

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
ACADEMIA DE PROPRIEDADE INTELECTUAL, INOVAÇÃO E
DESENVOLVIMENTO

VALÉRIA DELGADO BASTOS

INTELIGÊNCIA ARTIFICIAL: ECOSSISTEMAS E ESTRATÉGIAS EMPRESARIAIS DE
PROPRIEDADE INTELECTUAL E APROPRIABILIDADE DA INOVAÇÃO

Rio de Janeiro

2024

Valéria Delgado Bastos

Inteligência Artificial: Ecossistemas e estratégias empresariais de propriedade intelectual e apropriabilidade da inovação

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Propriedade Intelectual e Inovação, do Instituto Nacional da Propriedade Industrial.

Orientador: Prof. Dr. Sergio Medeiros Paulino de Carvalho
Coorientadora: Prof^a Dra. Adelaide Maria de Souza Antunes

Rio de Janeiro

2024

Ficha catalográfica elaborada pela Biblioteca de Propriedade Intelectual e Inovação – INPI
Bibliotecário responsável Evanildo Vieira dos Santos – CRB7-4861

B327 Bastos, Valéria Delgado.
Inteligência artificial: ecossistemas e estratégias empresariais de propriedade intelectual e apropriabilidade da inovação. -- 2024.

312 p. ; gráfs.; figs.; tabs.

Tese (Doutorado em Propriedade Intelectual e Inovação) - Academia de Propriedade Intelectual Inovação e Desenvolvimento, Divisão de Programas de Pós-Graduação e Pesquisa, Instituto Nacional da Propriedade Industrial – INPI, Rio de Janeiro, 2024.

Orientador: Prof. Dr. Sergio Medeiros Paulino de Carvalho.
Coorientadora: Profª Dra. Adelaide Maria de Souza Antunes.

1. Inteligência artificial – Propriedade intelectual. 2. Inteligência artificial – Inovação. 3. Inteligência artificial – Regime de apropriabilidade. 4. Ecossistema de Inovação. 5. Inovação aberta. I. Instituto Nacional da Propriedade Industrial (Brasil).

CDU: 347.77:004.89

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial deste trabalho, desde que citada a fonte.

Assinatura

Data

Valéria Delgado Bastos

Inteligência Artificial: Ecossistemas e estratégias empresariais de propriedade intelectual e apropriabilidade da inovação

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Propriedade Intelectual e Inovação, do Instituto Nacional da Propriedade Industrial.

Aprovada em 27 de novembro de 2024

Banca Examinadora:

Prof. Dr. Sergio Medeiros Paulino de Carvalho (Orientador)

Instituto Nacional da Propriedade Industrial

Prof^a. Dra. Adelaide Maria de Souza Antunes (Coorientadora)

Instituto Nacional da Propriedade Industrial

Prof. Dr. Antônio Marcio Buainain

Universidade de Campinas

Prof^a. Dra. Lia Hasenclever

Universidade Candido Mendes

Prof. Dr. Dr. Dirceu Yoshikazu Teruya

Instituto Nacional da Propriedade Industrial

Prof^a. Dra. Rita de Cássia Pinheiro Machado

Instituto Nacional da Propriedade Industrial

Prof^a. Dra. Cristina D'Urso Mendes Santos

Instituto Nacional da Propriedade Industrial

Rio de Janeiro

2024

DEDICATÓRIA

Aqueles que se foram antes de poder compartilhar deste momento

AGRADECIMENTOS

No começo desta tese, em meio às incertezas e angústias do isolamento imposto pela pandemia da Covid-19, não poderia imaginar a profundidade e a urgência que o tema assumiria nos anos seguintes. Essa trajetória, solitária em seu esforço, é coletiva no propósito de compreender uma tecnologia que promete ser o nosso futuro. O resultado da pesquisa, mais do que uma obra fechada, é a soma da interação com livros, artigos, dados... e pessoas.

Agradeço

Aos amigos que incentivaram este desafio: Luciana Badin, Professora da PUC-RJ; Luiz Augusto Veloso, Embaixador da Driblab; Mario Cordeiro de Carvalho Jr, Professor da UERJ; André Nassif, Professor da UFF; Marcos Migon e Paulo Cesar de Araújo Barcellos, do BNDES; Luiz Martins de Mello e Jacob Frenkel, pela sempre instigante troca de conhecimento; Jorge Khalil Misky, Inês Patrício, Neide Sodre e Eliane Ebel pela empatia nas horas difíceis; e minha terapeuta, Flávia Passos Soares, pelo suporte nos momentos em que o mundo parecia ruir.

Aos colegas do doutorado, pelo apoio e por dividir angústias acadêmicas e intelectuais em meio às incertezas sobre o curso e sobre o futuro, durante o distanciamento social.

Aos professores do curso pelo pote de conhecimento no final do arco-íris.

À equipe da Academia pelo suporte necessário em todos os momentos.

Aos profissionais que atuam na área da inteligência artificial que tive a oportunidade de conhecer, pessoal ou virtualmente, e aprender sobre um tema tão hermético.

Às comunidades I2AI e AI Brasil, que me mostraram o mundo real da IA e suas muitas oportunidades.

Aos membros da banca pelos comentários que enriqueceram a versão final deste trabalho.

À minha coorientadora, professora Adelaide Maria de Souza Antunes, por todo suporte acadêmico, leitura atenta e metódica de todas as versões da tese e apoio nos momentos difíceis.

Ao meu orientador, professor Sergio Medeiros Paulino de Carvalho, pelo apoio incondicional e por toda troca de conhecimento sobre os mecanismos para apropriabilidade da inovação.

Sem vocês esta tese certamente não teria sido concluída.

Any new technology, if it's used by evil people, bad things can happen. But that's more a question of the politics of the technology

Geoffrey Hinton

RESUMO

BASTOS, Valéria Delgado. **Inteligência Artificial: Ecossistemas e estratégias empresariais de propriedade intelectual e apropriabilidade da inovação.** Tese (Doutorado em Propriedade Intelectual e Inovação) - Instituto Nacional da Propriedade Industrial, Rio de Janeiro, 2024.

O tema da tese é a análise do regime de apropriabilidade da inovação da inteligência artificial, especificamente no subcampo de aprendizado profundo em arquiteturas de redes neurais artificiais que se consolidou na década passada e avança em ritmo acelerado na década atual. O objetivo é compreender o sentido e o alcance dos mecanismos tradicionais de propriedade intelectual e, a partir de estratégias empresariais diferenciadas, identificar a integração com ativos/tecnologias complementares essenciais para a captura de valor da inovação da IA. O método empregado para a análise de um fenômeno em desenvolvimento no contexto real foi a investigação documental exploratória de *corpus* heterogêneos sobre apropriabilidade da inovação, ecossistemas, plataformas digitais e inovação aberta, complementada por evidências empíricas, tendo como unidade de análise a firma. Um padrão diferenciado de inovação, consubstanciado em ecossistemas de plataformas digitais, caracterizados por relações complexas de interdependências e complementaridades, e que exige amplos conjuntos de dados e crescente capacidade infraestrutural e computacional, em um contexto de fortes assimetrias. Constatou-se que o grupo reduzido de empresas chamadas BigTechs sustentam o ritmo e a direção do progresso recente em IA, diretamente por meio de sua própria inovação e, indiretamente, pelo fornecimento de recursos de infraestrutura de hardware para empresas emergentes que se somam no desenvolvimento de modelos de IA de fronteira, ainda que ao custo de seu aprisionamento financeiro e computacional um subconjunto das BigTechs, o chamado núcleo infraestrutural. Constatou-se que o regime de apropriabilidade da inovação da IA é amplo e complexo, combinando múltiplos direitos de propriedade intelectual, que convivem harmonicamente com segredos comerciais e regime de código aberto, juntamente com o controle sobre ativos/tecnologias complementares essenciais para a inovação da IA. O atual regime legal de propriedade intelectual é fraco, incerto e subjetivo, com interpretações da lei baseadas em atos administrativos, que não constituem regulamentação substantiva, sem força e efeito de lei, ocasionando grande incerteza jurídica. O cenário de incerteza é agravado pelas muitas promessas e inquietações relacionadas às tecnologias da IA, que estão acarretando formas variadas de intervenção governamental ao redor do mundo, inclusive motivadas por disputa geopolítica, mas capazes de promover mudanças na arquitetura da indústria, no padrão de competição e, consequentemente, o regime de apropriabilidade da inovação. A tese pretende contribuir para o debate sobre criação e captura de valor da inovação da IA a partir da análise dos modelos de redes neurais profundas, complementando a literatura sobre ecossistemas e plataformas digitais e recuperando o papel da firma como unidade de análise.

Palavras-chave: Inteligência Artificial. Regime de Apropriabilidade. Ecossistema de Inovação. Estratégias empresariais. Propriedade Intelectual. Código aberto. Inovação Aberta.

ABSTRACT

BASTOS, Valéria Delgado. **Artificial Intelligence: Ecosystems and business strategies for intellectual property and appropriability of innovation.** Tese (Doutorado em Propriedade Intelectual e Inovação) - Instituto Nacional da Propriedade Industrial, Rio de Janeiro, 2024.

The theme of the thesis is the analysis of the appropriability regime of artificial intelligence innovation, specifically in the subfield of deep learning in artificial neural network architectures that has been consolidated in the last decade and advances at a fast pace in the current decade. The objective is to understand the meaning and scope of traditional intellectual property mechanisms and based on differentiated business strategies, to identify the integration with complementary assets/technologies essential for capturing the value of AI innovation. The method used for the analysis of a phenomenon in development in the real context was the exploratory documentary investigation of heterogeneous *corpus* on the appropriability of innovation, ecosystems, digital platforms and open innovation, complemented by empirical evidence, having the firm as the unit of analysis. A differentiated pattern of innovation, embodied in ecosystems of digital platforms, characterized by complex relationships of interdependence and complementarities, and which requires broad data sets and growing infrastructural and computational capacity, in a context of strong asymmetries. It was found that the small group of companies called BigTechs sustain the pace and direction of recent progress in AI, directly through their own innovation and, indirectly, by providing hardware infrastructure resources to emerging companies that join in the development of frontier AI models, even at the cost of their financial and computational lock-in a subset of BigTechs, the so-called infrastructural core. It was found that the regime of appropriability of AI innovation is broad and complex, combining multiple intellectual property rights, which coexist harmoniously with trade secrets and open-source regime, together with control over complementary assets/technologies essential for AI innovation. The current legal regime of intellectual property is weak, uncertain and subjective, with interpretations of the law based on administrative acts, which do not constitute substantive regulation, without the force and effect of law, causing great legal uncertainty. The scenario of uncertainty is aggravated by the many promises and concerns related to AI technologies, which are leading to various forms of government intervention around the world, including those motivated by geopolitical disputes, but capable of promoting changes in the architecture of the industry, in the pattern of competition and, consequently, the regime of appropriability of innovation. The thesis aims to contribute to the debate on the creation and value capture of AI innovation from the analysis of deep neural network models, complementing the literature on ecosystems and digital platforms, recovering the role of the firm as a unit of analysis.

Keywords: Artificial Intelligence. Appropriability Regime. Innovation Ecosystem. Business Strategies. Intellectual Property. Open Source. Open Innovation.

LISTA DE SIGLAS E ABREVIações

3GPP	3th Generation Partnership Project (parceria que desenvolveu padrão tecnologia telecomunicação wireless 3G)
AIaaS	<i>Artificial Intelligence as a Service</i> (serviço de software provido em computação na nuvem)
AT&T	<i>American Telephone and Telegraph Inc.</i>
Big Data	Área do conhecimento que estuda como tratar, analisar e obter informações a partir de conjuntos de dados muito grandes
CAS	<i>Chinese Academy of Science</i>
CII	<i>Computer Implemented Inventions</i>
CNIPA	<i>China National Intellectual Property Administration</i>
ConvNets	<i>Convolutional Neural Networks</i>
CSP	<i>Cloud Service Provider</i>
DaaS	<i>Data as a Service</i> (Dados fornecidos como um serviço em computação na nuvem)
DABUS	<i>Device for the Autonomous Bootstrapping of Unified Sentience</i>
DARPA	<i>U.S. Defense Advanced Research Projects Agency</i>
Deep Neural Networks	Sistemas de aprendizado profundo em arquiteturas de redes neurais artificiais
DMA	<i>European Union Digital Market Act</i>
DPI	Direitos de Propriedade Intelectual
DSA	<i>European Union Digital Services Act</i>
DSM	<i>European Union Digital Single Market</i>
ENPIs	Escritórios Nacionais de Propriedade Intelectual
EPC	<i>European Patent Convention</i>
EPO	<i>European Patent Office</i>
GAN	<i>Generative Adversarial Networks</i>
GDPR	<i>European Union General Data Protection Regulation</i>
GitHub	Git é um programa de código aberto para acompanhamento de alterações em arquivos de texto, escrito pelo autor do sistema operacional Linux e é a principal tecnologia na qual o GitHub, a interface social e do usuário, se baseia.
GPT	<i>Generative Pre-Trained Transformer</i>
GPU	<i>Graphic Processing Units</i>
GVC	<i>Global Value Chains</i>
IA	Inteligência Artificial
IaaS	<i>Infrastructure as a Service</i> (serviço de infraestrutura em plataformas de computação na nuvem)
IBM	<i>International Business Machines</i>
IEC	<i>International Electrotechnical Commission</i>
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
ETSI	<i>European Telecommunications Standards Institute</i>
F&A	Fusões e aquisições
FLOPS	<i>Floating-point operations per second</i> (operações de ponto flutuante por segundo), unidade de medida computacional
FRAND	<i>Fair, Reasonable, and Non-Discriminatory</i>
IoT	Internet of Things
IPC	<i>International Patent Classification</i>
ISO	<i>International Standard Organization</i>
ITU	<i>International Telecommunication Union</i>
KIPO	<i>Korean Intellectual Property Office</i>
KUBERNETS	Sistema de orquestração de contêineres <i>open-source</i> que automatiza implantação, dimensionamento e gestão de aplicações em contêineres
LaMDA	<i>Language Model for Dialog Applications</i>
LLM	<i>Large Language Model</i>

LTE	<i>Long Term Evolution</i> (tecnologia móvel de transmissão de dados criada com base no GSM e WCDMA).
LTS	<i>Large Technical Systems</i>
MaaS	<i>Model as a Service</i>
MPEP	<i>Manual of Patent Examining Procedure</i> (USPTO)
NASA	<i>U.S. National Aeronautics and Space Administration</i>
NBER	<i>National Bureau of Economic Research</i>
NIH	<i>U.S. National Institutes of Health</i>
NIST	<i>U.S. National Institute of Standards and Technology</i>
NLP	<i>Natural Language Processing</i>
NSCAI	<i>U.S. National Security Commission on Artificial Intelligence</i>
NSF	<i>U.S. National Science Foundation</i>
OASIS	<i>Organization for the Advancement of Structured Information Standards</i>
ODD	<i>European Union Open Data Directive</i>
OECD	<i>Organisation for Economic Cooperation and Development</i>
OECD.AI	<i>OECD AI Policy Observatory</i>
OMC	Organização Mundial do Comércio
OMPI	Organização Mundial da Propriedade Intelectual
OPT-175B	<i>Open Pretrained Transformer</i> (modelo de IA sobre imagens da Meta)
OSS	<i>Open Source Software</i>
PaaS	<i>Platform as a Service</i> (serviço de plataforma provido em computação na nuvem)
PCT	<i>Patent Cooperation Treaty</i>
PFI	<i>Profiting from Innovation</i>
PI	Propriedade Intelectual
PLT	<i>Patent Law Treaty</i>
PSI	<i>European Union Public Sector Information Directive</i>
SaaS	<i>Software as a Service</i> (serviço de software provido em computação na nuvem)
SD-SEP	<i>Self-declared standard essential patents</i>
SEP	<i>Standard Essential Patents</i>
SSO	<i>Standard-Setting Organizations</i>
SSRN	<i>Social Science Research Network</i>
TCP/IP	<i>Transmission Control Protocol/Internet Protocol</i> (conjunto de protocolos de comunicação entre computadores em rede)
TH	Tecnologias Habilitadoras
TIC	Tecnologias de Informação e Comunicação
TPG	Tecnologias de Propósito Geral
TRIPS	<i>Trade-Related Aspects of Intellectual Property</i>
TTC	<i>European Union-United States Trade and Technology Council</i>
UKIPO	<i>United Kingdom Intellectual Property Office</i>
UNCTAD	<i>United Nations Conference on Trade and Development</i>
USPTO	<i>United States Patent and Trademark Office</i>
VLOPS	<i>Very Large Online Platforms</i>
VLOSEs	<i>Very Large Online Search Engines</i>
W3C	<i>World Wide Web Consortium</i>
WAN	<i>Wide Area Network</i>
WIPO	<i>World Intellectual Property Organization</i>
WWW	<i>World Wide Web</i>
XAI	<i>Explainable Artificial Intelligence</i>

LISTA DE FIGURAS, GRÁFICOS E TABELAS

Figura A	Representação gráfica da estratégia de revisão da literatura	30
Figura 1.1	Regime de Apropriabilidade das Inovações - Profiting from Innovation (PFI)	53
Figura 1.2	Ativos complementares necessários para comercializar uma inovação	58
Figura 2.1	Linha do tempo – As fases da inoação da IA	60
Figura 2.2	Principais abordagens de pesquisa da IA	63
Gráfico 2.1	As três eras da Computação de treinamento de sistemas notáveis de aprendizado de máquina ao longo do tempo	66
Tabela 2.1	Vinte principais detentores de patentes de IA nos EUA e de pedidos de patentes de IA no mundo	67
Quadro 2.1	Definições legais da Inteligência Artificial	69
Figura 2.3	Representação da IA, aprendizado de máquina e o subcampo de aprendizado profundo	70
Figura 2.4	Representação esquemática do treinamento de sistema de aprendizado profundo	72
Figura 2.5	Ecossistema de Inovação da IA	75
Tabela 2.2	Despesas com P&D e intensidade de P&D (pelo lucro bruto ou por receitas), empresas selecionadas (BigTechs, principais concorrentes e grandes firmas farmacêuticas)	81
Gráfico 2.2	Número de firmas adquiridas pelas BigTechs americanas, 1998-2020	82
Gráfico 2.3	Investimento corporativo global anual em IA, por atividade de investimento, 2013 e 2021, em porcentagem	83
Gráfico 2.4	Série temporal de publicações científicas de IA por instituições, da Scopus, 2010-2020	84
Tabela 2.3	Publicações científicas e patentes, BigTechs selecionadas, 2014-2019	85
Tabela 2.4	Top 10 organizações de coautores em publicações científicas - BigTechs e universidades, 2014-2019	85
Gráfico 2.5	Colaboração entre países em publicações de IA, 2010-2021	86
Gráfico 2.6	Inteligência Artificial, uso como OSS	88
Gráfico 2.7	Participação de mercado mundial dos principais provedores de serviços de infraestrutura em nuvem, Q1 2024	91
Figura 2.6	Representação gráfica das fontes de dados, inovação e capacidades dinâmicas da Alphabet/Google	93
Gráfico 2.8	Trinta principais proprietários de patentes de IA concedidos pelo USPTO, percentagem, 1976-2018	94
Figura 2.7	Cem maiores plataformas digitais por capitalização de mercado, agosto 2023	95
Gráfico 2.9	Número de famílias de patentes de IA e publicações científicas, países selecionados, 1960-2017	95
Tabela 3.1	Sumário de redes neurais artificiais de aprendizado profundo	102
Gráfico 3.1	Ascensão e Crescimento dos principais LLMs e bots associados, por número de parâmetros, principais desenvolvedores (posição até 20 de março de 2024)	106
Figura 3.1	Tipos de modelos da IA generativa e tipos de dados de entrada/saída	108
Gráfico 3.2	Modelos de IA generativa, por domínio (percentual do número total), acumulado 2014-2023	108

Gráfico 3.3	Evolução do número de modelos destacados da IA generativa, por domínio, 2014-2023	108
Figura 3.2	Representação da melhoria do desempenho e ampliação de tarefas com o aumento a escala do modelo	109
Tabela 3.2	Principais LLM e modelos multimodais, por número de parâmetros e desenvolvedor	110
Gráfico 3.4	Número de modelos de AI notáveis, por afiliação e ano de publicação, 1950-2022	111
Figura 3.3	Número de meses em que os modelos de referência GPT-3 (LLM da OpenAI), DALL-E (Modelo de imagem da OpenAI) e AlphaFold1 (Modelo de Proteínas da DeepMind) permaneceram fechados até serem implementados, clonados ou melhorados pela comunidade de código aberto	117
Gráfico 3.5	Crescimento no número de parâmetros dos modelos da IA no tempo, 1950-2023	119
Gráfico 3.6	Quantidade de dados usados para treinar sistemas de IA notáveis, por modelo e domínio	119
Gráfico 3.7	Crescimento na capacidade computacional, em petaFLOP, ao longo do tempo	119
Tabela 3.3	Custo de treinamento estimado de modelos de IA selecionados	121
Gráfico 3.8	Custo estimado de treinamento e uso de computação, modelos de IA selecionados	122
Gráfico 3.9	Desagregação dos custos de treinamento e desenvolvimento, modelos selecionados de IA	122
Gráfico 3.10	Consumo anual de eletricidade por empresas de data centers, em terawatt hora, 2018–2022	123
Figura 3.5	Nível de acessibilidade e estratégias de liberação de modelos de fundação	126
Figura 3.6	Pilha de tecnologia IA generativa	127
Gráfico 3.11	Uso de tipos de Chips relatados em artigos publicados, ente 2018 e 2023 (e estimativa 2024)	135
Gráfico 3.12	Diferentes tipos de Chips mencionados nos artigos sobre IA publicados	135
Gráfico 3.13	Tipos de chips empregados no treinamento dos modelos lançados em 2023	136
Quadro 4.1	Principais Desenvolvimentos Regulatórios da IA	147
Gráfico 4.1	BigTechs reguladas pelo Digital Market Act - Serviços de plataforma fornecidos pelos “gatekeepers” designados pelo DMA da União Europeia	156
Quadro 4.2	Dependência Líquida de Importação dos EUA em 2023	166
Figura 5.1	Regime Jurídico dos Direitos de Propriedade Intelectual da IA	177
Figura 5.2	Representação gráfica das <i>Computer Implemented Inventions</i> , incluindo invenções relacionadas a IA	182
Quadro 5.1	Situação dos pedidos de patentes em nome do Sistema DABUS em vários países	193
Gráfico 6.1.1	Principais empresas desenvolvedoras de modelos de IA, em termos do número de modelos lançados nos subperíodos 2010-2016 (a) e 2017-2023 (b)	220
Gráfico 6.1.2	Principais empresas desenvolvedoras de modelos de IA que representaram melhoria no estado da arte (SOTA improvement), em termos de percentual no total de modelos lançados nos subperíodos (a) 2010-2016 e (b) 2017-2023	221
Gráfico 6.1.3	Principais empresas desenvolvedoras de modelos de IA altamente citados, em termos de percentual no total de modelos lançados nos subperíodos (a) 2010-2016 e (b) 2017-2023	221

Gráfico 6.1.4	Medida de qualidade dos modelos desenvolvidos pelas principais organizações (isoladamente e em parcerias) pela percentagem de modelos que representaram melhoria no estado da arte (SOTA) e de modelos mais citados, em termos de percentagem do total, em 2010-2024 (até março)	222
Gráfico 6.1.5	Principais desenvolvedores dos modelos de IA intensivos em computação (percentual do número total desses modelos)	225
Gráfico 6.1.6	Principais modelos de IA generativa em termos de computação de treinamento (estimada), em bilhões de petaFLOP (10^{15} operações de ponto flutuante estimadas na literatura de IA)	225
Figura 6.1.1	Alianças sobrepostas das BigTechs com empresas emergentes da IA generativa	226
Figura 6.1.2	Relação entre BigTechs e Nvidia e desenvolvedores de modelos de fundação	228
Figura 6.1.3	Relação entre BigTechs/Nvidia e desenvolvedores de modelos de fundação, provedores de dados e provedores de ferramentas para “devs”	228
Tabela 6.1.1	Detalhes técnicos de modelos intensivos em computação, por desenvolvedor e ano de lançamento	229
Gráfico 6.1.7	Número de chips empregados no treinamento dos modelos de IA mais intensivos em computação	230
Gráfico 6.2.1	Número de pedidos de patentes de IA e proporção no total (famílias de patentes), 1980-2017	231
Gráfico 6.2.2	Famílias de patentes e publicações acadêmicas de IA generativa, por ano, 2014–2023	232
Gráfico 6.2.3	Vinte principais organizações detentores de famílias de patentes de IA generativa, 2014-2023	233
Gráfico 6.2.4	Principais desenvolvedores de modelos de IA generativa lançados em 2014-2023	234
Gráfico 6.2.5	Número de patentes de IA generativa e número de modelos lançados pelos vinte principais detentores de patentes, 2014-2023	234
Gráfico 6.2.6	Famílias de patentes globais em cinco principais tipos de modelos de IA generativa, acumulado 2014–2023 (a) e evolução 2014-2023 (b)	236
Gráfico 6.2.7	Lançamento de modelos de IA por domínio, acumulado 2014-2023 (a) e evolução 2014-2023 (b)	237
Gráfico 6.2.8	Dez principais depositantes de pedidos de patentes relacionadas a redes Transformers, USPTO, 2016-2022	237
Gráfico 6.2.9	Patentes de IA generativa por tipo de arquitetura e pelos países líderes, 2014-2022	238
Quadro 7.1	Quadro comparativo da apropriabilidade da inovação da IA preditiva e IA generativa	243
Gráfico 7.1	Famílias de patentes 5G (concedidas e pedidos) no mundo, principais empresas	249
Gráfico 7.2	Famílias de Patentes 5G ponderadas pela taxa de essencialidade, principais empresas	249
Gráfico 7.3	Receita publicidade, principais plataformas digitais, US\$ bilhões 2021	253
Tabela 7.1	Vinte principais bases de dados de acesso aberto da IA generativa	265
Figura 7.1	Representação visual da cadeia de valor dos modelos de fundação e possíveis direitos de propriedade	270
Gráfico 7.4	Evolução das maiores empresas em capitalização de mercado, US\$ bilhão, 2014 a ago. 2024	271

Gráfico 7.5	Capitalização de mercado, vinte maiores empresas de tecnologia, US\$ bilhão, 17 ago. 2024	272
Tabela 7.2	Valor de mercado dos principais unicórnios de IA, chips e bases de dados, em US\$ bilhão (nov. 2023)	273
Gráfico 7.6	Receita anual, vinte maiores empresas de tecnologia em capitalização de mercado, US\$ bilhão, 2023	274
Tabela 7.3	Gasto em P&D (em US\$ bilhões e como percentual da receita, 2022) e Salários (médio em US\$, 2023) e Salário CEO (US\$ milhão, 2023), empresas selecionadas	274

SUMÁRIO

INTRODUÇÃO	18
OBJETIVOS DA TESE	25
JUSTIFICATIVA, RELEVÂNCIA E ORIGINALIDADE	26
METODOLOGIA	27
ESTRUTURA DA TESE	32
CAPÍTULO 1 – REVISITANDO A PROPRIEDADE INTELECTUAL NO CONTEXTO MAIS AMPLO DA APROPRIABILIDADE	35
1.1. <i>Mainstream</i> econômico e a visão heterodoxa da propriedade intelectual	37
1.2. Propriedade intelectual nas tecnologias de informação e comunicação	42
1.3. Inovação aberta, ecossistemas, plataformas digitais e apropriabilidade	44
1.4. As tecnologias de interoperabilidade	47
1.5. A crescente importância dos dados na economia digital	52
1.6. Regime de apropriabilidade: além da propriedade intelectual	53
CAPÍTULO 2 – ECOSSISTEMAS DE INOVAÇÃO E PLATAFORMAS DIGITAIS: A CONSOLIDAÇÃO DA INTELIGÊNCIA ARTIFICIAL PREDITIVA	60
2.1. Antecedentes: A longa trajetória da IA	60
2.2. A IA preditiva: Redes neurais profundas, tipos de aprendizado e a não- explicabilidade	69
2.3. O ecossistema de inovação da IA preditiva	75
2.3.1. Sistemas de IA	77
2.3.2. Big Data	80
2.3.3. Investimentos em P&D e aquisição de empresas	81
2.3.4. Interação com a comunidade científica e acadêmica	84
2.3.5. Supercomputação e computação na nuvem	90
2.4. A liderança na inovação da IA: ecossistemas e modelos de negócios	93
CAPÍTULO 3 – A INTELIGÊNCIA ARTIFICIAL GENERATIVA: OS AVANÇOS EM REDES NEURAIS E A IMPORTÂNCIA CRESCENTE DOS ATIVOS TANGÍVEIS	97
3.1. O passo adiante com a IA generativa	99
3.2. Os primeiros modelos e aplicativos generativos	101
3.3. Muito além de LLMs e aplicativos conversacionais	108
3.4. Protagonistas e coadjuvantes da IA generativa	113
3.5. As incertezas na escala dos modelos: custos e grau de abertura	118
3.5.1. A importância estratégica da capacidade computacional	119
3.5.2. Custos crescentes de treinamento dos modelos	121
3.5.3. Divulgação dos modelos em regime <i>open-source</i>	125
3.6. Os ativos tangíveis da IA: onde o vencedor leva tudo	128
3.6.1. O complicado quebra-cabeça geopolítico da cadeia de valor dos semicondutores	128
3.6.2. Processamento paralelo e chips de IA	130
3.6.3. Os principais chips usados pelos modelos de IA	135
3.6.4. A infraestrutura física por trás da IA: Data centers, satélites, cabos submarinos	137
3.6.5. Computação na nuvem: a principal vantagem comparativa das BigTechs	140

CAPÍTULO 4 – INTERVENÇÃO GOVERNAMENTAL E EFEITOS POTENCIAIS SOBRE A ARQUITETURA DA INDÚSTRIA, O PADRÃO DE COMPETIÇÃO E O REGIME DE APROPRIABILIDADE DA IA	146
4.1. Avanços recentes na regulação sobre modelos de fronteira	147
4.1.1. A abordagem europeia	147
4.1.2. A regulação pragmática chinesa	150
4.1.3. A regulação americana marcada pela disputa tecnológica geopolítica	151
4.1.4. As novas iniciativas regulatórias em nível global	153
4.1.5. O posicionamento empresarial divergente sobre a regulação da IA	154
4.2. O escrutínio antitruste se transfere para a computação na nuvem	155
4.3. Políticas industriais e comerciais de chips e IA generativa	162
4.4. Tratados internacionais de comércio como instrumento de proteção das tecnologias da IA	168
CAPÍTULO 5 – DIREITOS DE PROPRIEDADE INTELECTUAL DA INTELIGÊNCIA ARTIFICIAL	171
5.1. O legado da proteção dupla do software	173
5.2. Propriedade intelectual dos insumos e produtos da IA	177
5.2.1. Insumos da IA	178
5.2.1.1. Modelos de IA	178
5.2.1.2. Bases de dados	179
5.2.1.3. A proteção dos chips de computador	181
5.2.2. Produtos da IA: invenções relacionadas à IA	182
5.3. Patentes de invenção relacionadas à IA na prática dos escritórios nacionais de propriedade intelectual	184
5.3.1. O caso DABUS	193
5.4. O que muda com a IA generativa?	197
5.4.1. Os novos desafios para invenções e criações da IA	199
5.4.2. O <i>open source</i> da IA como um quase-DPI	204
5.4.3. Direito autoral, <i>fair use</i> e os gargalos de bases de dados	209
5.4.4. Impactos sobre o funcionamento do sistema de propriedade intelectual	213
5.5. Considerações finais	214
CAPÍTULO 6 – ALGUMAS EVIDÊNCIAS DA IA	218
6.1. Evolução dos modelos de IA e seus principais protagonistas	218
6.2. Evolução das patentes de IA e seus principais protagonistas	231
CAPÍTULO 7 – DISCUSSÃO SOBRE O REGIME DE APROPRIABILIDADE DA IA ...	241
7.1. Apropriabilidade da inovação da IA preditiva	244
7.2. Apropriabilidade da inovação da IA generativa	260
7.3. Os verdadeiros vencedores da corrida da IA	271
CONCLUSÕES	278
Referências	286

INTRODUÇÃO

Inovações tecnológicas são a base da competição e do crescimento da firma no longo prazo, alcançando e preservando mercados e buscando auferir lucros extraordinários. As vantagens competitivas da firma dependem cada vez mais da sua capacidade para inovar e de produzir e controlar ativos intangíveis¹ como a propriedade intelectual (PI).

Direitos de propriedade intelectual (DPI) são instrumentos de proteção jurídica de criações do intelecto humano, conferindo a inventores (e autores) um direito exclusivo sobre a invenção (ou obra) por determinado período. Mecanismos formais são fornecidos por leis de propriedade intelectual e apontados por uma parcela da literatura como a principal ferramenta de proteção e, consequentemente, de incentivo à inovação.

Paradoxalmente, entretanto, o longo e consolidado paradigma da PI é ocasionalmente afrontado pela teoria e pelos fatos, reabrindo o debate sobre sua real capacidade de incentivar a inovação. Esta instituição é frequentemente contestada, sendo considerada insuficiente ou ineficaz para a apropriação do valor da inovação por uma parcela da literatura, teórica e empírica, principalmente porque o conhecimento relevante usualmente não pode ser inteiramente codificado e porque a captura de valor depende de outros mecanismos e das muitas formas de concorrência no mercado. Há uma variedade de mecanismos distintos e informais, muitas vezes ditados pela natureza da tecnologia, algumas vezes complementares, para apropriabilidade da inovação (Dosi e Nelson, 2010). Em contextos específicos, como nos setores de alta tecnologia e software, outros instrumentos, inclusive informais, são vistos como mecanismos mais efetivos de proteção da inovação.

A história é repleta de casos de empresas inovadoras e as primeiras requerentes de patentes que fracassaram e não foram capazes de transformar a inovação em lucros e vantagens econômicas duradouras. O artigo seminal de Teece (1986) foi o ponto de partida da literatura conhecida como *Profiting from Innovation* (PFI), que conjuga enfoques teóricos antes independentes para explicar por que nem sempre a firma pioneira na introdução de uma inovação no mercado a partir de uma invenção protegida por PI é aquela que se apropria do seu valor no longo prazo. A captura duradoura do valor, conforme esta abordagem, que combina inovação e gestão estratégica, depende do regime de apropriabilidade que contempla não somente (ou principalmente) proteção legal por PI. Outras variáveis, como converter a inovação no *design* tecnológico dominante e formular estratégias para obter ativos ou tecnologias

¹ Sobre o termo, ver Investopedia. Disponível em <https://www.investopedia.com/terms/i/intangibleasset.asp>. Acesso em 5 fev. 2023

complementares necessárias para o sucesso da inovação, são fundamentais e dependem de capacidades dinâmicas da firma em contextos distintos da concorrência, ao longo do tempo, principalmente quando o regime legal de PI é fraco.

A relativização do papel da PI é central nesta abordagem pela hipótese da usual fraqueza dos regimes legais e quase um corolário de como a inovação, e não uma mera invenção, alcança e se mantém competitiva no mercado. O conceito de complementaridade nas inovações é central, seja por meio de ativos, seja por meio de tecnologias, combinado à noção de competição (dinâmica, central em qualquer abordagem da linha Schumpeteriana), seja diretamente por inovadores de produtos similares, seja indiretamente por firmas previamente fornecedoras de ativos complementares à inovação que acabam se diversificando ou integrando verticalmente.

Os temas da apropriabilidade, capacidades dinâmicas e competição, pilares em blocos separados da produção acadêmica de David Teece, estão sintetizados no PFI e, quando considerados em seu conjunto, se aproximam da tradição de economistas focados no crescimento da firma no longo prazo, como Penrose (1959) e Chandler (1962, 1977), ou estudos sobre inovação, como Nelson (1991; 1959), Nelson e Winter (1982) e Rosenberg (1990), apenas para citar alguns trabalhos que tomaram a firma como unidade de análise. Esses autores compartilharam ideias precursoras, mesmo que com abordagens distintas, tendo em comum o ceticismo sobre a lógica por trás dos DPIs plantadas por Penrose (1951) e Machlup (1958).

A emergência de novas tecnologias disruptivas é um fenômeno sempre desafiador. Os estágios iniciais são marcados por possibilidades incertas e carecendo do instrumental analítico para seu entendimento, cuja essência é ocasionar mudanças, tema central de Schumpeter e da economia evolucionária (Nelson et al., 2018), em longos e não previsíveis processos endógenos envolvendo organizações, instituições e tecnologias em dinâmicas interativas complexas, suscetíveis aos impulsos e pressões do contexto (Mendonça et al., 2022).

A emergência dessas tecnologias díspares de padrões predefinidos frequentemente reabre o debate sobre a eficácia dos mecanismos de PI, eventualmente levando a mudanças nesses instrumentos (Carvalho, 2003), frequentemente resultante da influência de estratégias empresariais (Drahoš, 1995; Pisano, 2006a; Pisano e Teece, 2007) e assegurando uma adaptabilidade notável do sistema de PI.

Direitos de PI oferecem proteção no plano nacional, frequentemente exigindo dos titulares a busca de proteção também em outros países, particularmente em mercados globalizados. Com vistas a minimizar a fragmentação legal e reduzir custos de transação, a legislação foi progressivamente harmonizada em escala global por tratados internacionais (Charnovitz, 1998), em especial o Acordo TRIPS (acrônimo da expressão *Trade-Related*

Aspects of Intellectual Property), de 1994.

O surgimento das novas tecnologias digitais e, em especial as modernas técnicas de inteligência artificial (IA), promete ser um desses novos capítulos da PI e da apropriabilidade da inovação. A complexidade e dinamismo dessa tecnologia inspiradora, que se tornaram aparentes desde a década passada e que avança espetacularmente na década atual, parece novamente desafiar os padrões de como as firmas inovam e de como lucram com a inovação.

As tecnologias da IA, até agora inseridas no cenário das plataformas digitais, onde obtêm os insumos essenciais para a inovação e nas quais são empregadas como insumo cativo em sistemas automatizados de classificação, recomendação, persuasão e previsão, em produtos e serviços inéditos, emergiram, contudo, de ecossistemas mais amplos, que “combinam” complementaridades (cooperação) e competição, como mostrado em trabalhos recentes sob a noção de “coopetição” (Teece, 2018a; Teece, 2016; Petit e Teece, 2020).

A digitalização da economia e a disponibilização de expressivos conjuntos de dados abriram possibilidades antes impensadas para criação de valor a partir dos dados, que junto com complexas técnicas algorítmicas e crescentes recursos computacionais respaldam as inovações da IA, cuja apropriabilidade dependeu sobretudo do exitoso modelo de negócios de plataformas digitais.

De fato, novas tecnologias usualmente demandam novas estruturas, como importantes inovações organizacionais que deixaram no passado a firma multidivisional “chandleriana” verticalmente integrada, com as cadeias globais de valor (GVC, sigla de *global value chains*) na década de 1990 e mais recentemente nas plataformas digitais (Andreoni, Chang e Labrunie, 2021).

Adicionalmente, a IA tem sido frequentemente descrita como a principal inovação disruptiva² da quarta revolução industrial, ou Indústria 4.0, noção que abarca processos de fabricação mais inteligentes, incluindo a dependência de sistemas ciberfísicos, que interligam componentes cibernéticos e físicos, como veículos autônomos, fábricas inteligentes, conhecidas como indústria integrada³ (Khaitan e McCalley, 2014; Zhou et al., 2015).

O enorme potencial transformador da IA decorre de suas características de impactar aplicações setoriais e interfaces amplas com outras tecnologias (OECD, 2019), principalmente aquelas do campo digital, como Internet das Coisas (IoT, sigla de *Internet of Things*), impressão em três dimensões (3D printing), computação na nuvem (cloud computing), entre outras

² Sobre o conceito de inovação disruptiva ver Christensen (1997)

³ O termo “Indústria 4.0” é uma invenção de políticos de pesquisa alemães para referir-se explicitamente à dimensão digital das futuras estruturas industriais (Schroeder, 2016).

(Schwab, 2016; Sejnowski, 2018), cuja implementação articulada resultará em enorme poder transformador da economia digital (sobre o conceito, Bukht e Heeks, 2017).

O potencial de disseminação e transformação de vários setores econômicos levou alguns autores a qualificar a IA como uma tecnologia de propósito geral (TPG), que se distingue de tecnologias discretas pela transversalidade (Trajtenberg, 2018; Klinger et al., 2018; Lyu, 2020). As TPG são caracterizadas por uso amplo e generalizado (ubiquidade), melhoria técnica contínua e geração de inovações setoriais complementares (Bresnahan e Trajtenberg, 1995), como no caso da máquina a vapor, da eletricidade e das TICs, protagonistas das revoluções industriais. Em termos macroeconômicos, a emergência desse tipo de tecnologia, a partir de uma inovação radical original, da origem a *clusters* de inovações que ocasionam mudanças no paradigma tecno-econômico e abrem “explosões de oportunidades” (Perez, 1983), ou provocam mudanças de paradigma e criação de novas trajetórias tecnológicas (Dosi, 1982), gerando ondas de crescimento econômico e produtividade, além de mudanças estruturais profundas na economia.

Autores como Teece, 2018a; Teece, 2018b; Gambardella et al., 2021, preocupados com a apropriabilidade de valor no nível microeconômico, preferem trabalhar com o conceito menos ambicioso de tecnologia habilitadora (TH), em que a IA teria menor aplicabilidade setorial, mas mesmo assim central para muitas outras tecnologias⁴.

De fato, TPG envolvem maiores desafios relacionados ao conhecimento incompleto de inúmeras aplicações futuras potenciais (Rosenberg, 1998), com incertezas pelo lado da oferta e da demanda, que demandam uma infinidade de capacidades dinâmicas que extrapolam redes neurais profundas e ciência de dados (Pisano, 2017). De todo modo, tanto TPG quanto TH envolvem condições inerentemente distintas de apropriabilidade, uma vez que ambas geram fortes transbordamentos (*spillovers*) e, conseqüentemente, incentivos insuficientes para o investimento privado (Teece, 2018a; Teece, 2018b; Gambardella et al., 2021).

No caso da IA, as incertezas pareceriam mais complexas, por três razões. Primeiro, porque sua própria base técnica ainda não parece inteiramente desenvolvida. Segundo, pela necessidade de estruturação de ecossistemas de inovação mais amplos, com modelos de governança e apropriabilidade ainda indefinidos. Terceiro, por exigirem inovações

⁴ Sobre tecnologias habilitadoras ver David (1992). Há autores que, pela natureza infraestrutural da IA, consideram o conceito de *Large Technical Systems (LTS)* - sistemas ou redes de enormes proporções e complexidade do final do século XX que inspiraram novas tecnologias, novas organizações e novos estilos de gestão, como ARPANET, rede de informação interativa baseada em computador que lançou as bases da Internet - mais adequado para descrevê-la (Vannuccini e Prytkova, 2021). Cockburn, Henderson, Stern (2018), por outro lado, consideram a IA um novo “método de invenção” ou invenção dos métodos de invenção (IMI), capaz de remodelar a natureza do processo de inovação e a organização da P&D, conceito que Nardin (2021) expande para método geral de invenção.

complementares de *players* setoriais para uma tecnologia que gera desconfiança sobre sua transparência e confiabilidade, acrescida de desafios institucionais com potencial de promover mudanças na arquitetura da indústria e no padrão de competição, incluindo regulação da IA, controle sobre dados e escrutínio antitruste sobre BigTechs⁵, aos quais se soma um arsenal de medidas de política industrial e comercial em meio à disputa geopolítica/geoestratégica pela hegemonia da IA.

O florescimento da implantação da IA (Lee, 2018; 2020; Lee e Qiufan, 2021) em outros setores da economia ainda não se realizou, mas deverá ocorrer em contextos competitivos diferenciados com desafios suplementares para criação e apropriação de valor.

Contudo, mesmo sem uma ampla disseminação, as inovações tecnológicas da IA já impactaram a maneira como vivemos, trabalhamos, produzimos e interagimos, com potencial de uma profunda metamorfose da economia e da sociedade. Muitas vezes expressas em aplicativos de softwares ou complexos dispositivos de hardware, as inovações da IA avançam de forma ainda mais acelerada e disruptiva na década atual, gerando perspectivas de uso potencial em uma infinidade de setores econômicos.

Não obstante, muitos aspectos técnicos da IA permanecem obscuros e ambíguos, sequer existindo uma definição universalmente aceita pela comunidade científica. Com efeito, a IA é um mero conceito guarda-chuva, genérico, composto por diferentes técnicas⁶ e aplicações funcionais distintas⁷ (WIPO, 2019; OECD, 2019). Na prática, são tecnologias digitais emergentes usadas para enfrentar problemas complexos, muitos dos quais impraticáveis de serem resolvidos usando métodos tradicionais.

A maioria dos atuais sistemas exitosos são referentes ao aprendizado profundo, pertencente ao paradigma de pesquisa conexcionista que se tornou predominante após saltos no desempenho dos modelos em decorrência de revolucionária inovações algorítmicas, da

⁵ A maioria dessas empresas têm suas operações apoiadas em modelos de negócios de plataformas digitais, como Alphabet/Google, Meta (proprietária das plataformas Facebook, WhatsApp e Instagram), Amazon, Apple. A Microsoft não é comparável às demais operadoras de plataformas, razão de não ter sido incluída entre as 17 VLOPs (Very Large Online Platforms) – plataformas com mais de 45 milhões de usuários ativos mensais –, que devem atender às obrigações definidas pelo *Digital Services Act* (DAS) da União Europeia, pelo seu padrão de competição diferenciado. Contudo, o mecanismo de busca online Bing, da Microsoft, que passou a contar com ferramentas da IA generativa do ChatGPT, com perspectivas de ampliar seu número de usuários, sendo incluído, junto com o Google Search, como VLOSEs (Very Large Online Search Engines). Comissão Europeia https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413. Acesso em 20 set. 2024

⁶ Atualmente, as principais técnicas de IA são o aprendizado de máquina (ML, sigla de *machine learning*) e, principalmente, o subcampo aprendizado profundo (DL, sigla de *deep learning*), embora contemple também outras técnicas, como programação lógica, engenharia de ontologia, lógica difusa e raciocínio probabilístico (WIPO (2019); OECD, 2019).

⁷ Aplicações funcionais contemplam visão computacional, processamento de linguagem natural, reconhecimento de fala, entre outras.

disponibilização de gigantescos conjuntos de dados com a digitalização e de descomunal ampliação em capacidade computacional.

Sistemas de IA e algoritmos “autodidatas” – tomando emprestada a expressão de Varian (2018) – conseguem, a partir dessas correlações identificar preferências e padrões de comportamento, realizar tarefas específicas que envolvam tomada de decisão e previsão, como recomendar produtos (ao prever as preferências do cliente), dirigir veículos (ao prever a direção correta para virar o volante) ou compor músicas (ao prever sequências de notas) como apontado por alguns autores (Brynjolfsson et al., 2017; Agrawal, Gans e Goldfarb, 2017). Com os novos modelos da IA generativa, podem também escrever textos semântica e sintaticamente perfeitos (ao prever palavras) e criar imagens inéditas (ao prever *pixels*).

Contudo, não chegam a compreender no sentido da inteligência humana o significado de palavras, nem de imagens, carregam vieses contidos nos dados, há baixa transparência no seu funcionamento interno e de como alcança seus resultados. Isso tem gerado incertezas e inquietações sobre os riscos e reais benefícios da IA, que se refletem em disseminada regulação de seus usos, em direitos de PI e até nas políticas antitruste e defesa da concorrência relacionadas às plataformas digitais, ainda o principal locus de geração e utilização das técnicas da IA, com potencial de influenciar a estrutura da indústria, o padrão de competição, o regime de apropriabilidade e, conseqüentemente, os vencedores e perdedores da nova revolução tecnológica.

A presente pesquisa procurou avançar no entendimento dessas tecnologias e da inovação da IA, a partir da evolução do conhecimento e da ação de firmas precursoras que criaram ecossistemas peculiares que incorporam diferentes dimensões de ativos (tecnologias) complementares, viabilizadas pelo modelo de negócios de plataforma digital, com vistas a examinar a apropriabilidade da inovação. São também analisados fatores institucionais que podem, em algum grau, influenciar o avanço da IA ao modificar a arquitetura da indústria, como ocorreu com as Tecnologias de Informação e Comunicação (TICs) ao viabilizar um padrão de competição atípico, podendo produzir efeitos similares no futuro.

O desenvolvimento das tecnologias de IA é apresentado nesta tese sob uma perspectiva histórica, com a identificação de fases diferenciadas em termos de base técnica e fatores institucionais. A trajetória foi longa até chegar ao tipo de sistemas de IA e a consolidação de uma determinada abordagem de pesquisa, cujas raízes estão na expansão da Internet e na estruturação do ecossistema da IA em um contexto de mercado privado. A consolidação do aprendizado profundo foi devida a progressos notáveis em redes neurais artificiais multicamadas, avanços em capacidade computacional e na ampliação do volume de dados e

estruturação de mecanismos automatizados de coleta, organização, armazenamento, análise e tratamento dos dados, sob governança das firmas operadoras de plataformas, possibilitando o desempenho e a acuracidade crescentes dos modelos de IA.

Desde a década de 2010, os sistemas de aprendizado profundo em arquiteturas de redes neurais artificiais – também denominados redes neurais profundas – deram origem a inovações da popularmente chamada IA preditiva. Essa classe particular de sistema, que usa algoritmos, ou modelos, para imitar neurônios do cérebro e identificar padrões complexos, foi empregada nos produtos/serviços disponibilizados nas plataformas digitais e tendo como principais *players* um grupo reduzido de empresas conhecidas como BigTechs⁸, americanas e chinesas, como Apple, Microsoft, Alphabet/Google, Amazon e Meta Platforms/Facebook, Tencent, Baidu, Alibaba, Huawei entre outras.

Desde fins da década passada, uma nova rodada de avanços técnicos aceleraram a emergência de modelos distintos, da chamada IA generativa, cada vez mais poderosos, desenvolvidos em velocidade mais rápida do que poderiam ser comercializados, originalmente restritos à comunidade de desenvolvedores (doravante, “devs.”), até chegarem de forma transformadora ao mercado no início da década atual. Modelos com capacidade de criar conteúdo diverso foram agressivamente disponibilizados por empresas emergentes e incumbentes do setor, em um ambiente de rivalidade aberta pela definição do *design* tecnológico dominante e mudanças na estrutura da indústria por meio da integração vertical com tecnologias complementares essenciais para a inovação em IA, desafiando paradigmas e o próprio regime de apropriabilidade, o modelo de negócios e o próprio regime legal de PI.

A presente pesquisa analisa as características de criação de valor e apropriabilidade da inovação da IA, focado na firma e no arranjo de ecossistemas de inovação, complementando uma parcela relevante da literatura sobre ecossistemas de plataformas digitais. A pesquisa pretende contribuir para o debate a partir da análise das características técnicas dos modelos de redes neurais profundas, revisitando o processo de criação e captura de valor na fase de desenvolvimento da IA preditiva e os novos desafios colocados pela emergência da IA

⁸ O termo BigTech não é um conceito acadêmico, mas se popularizou e é amplamente utilizado com vistas a resumir as características de empresas líderes dos mercados digitais, que operam modelos de negócios de plataformas online e se tornaram as mais valiosas do mundo, a partir de um crescimento quase ininterrupto há pelo menos duas décadas. Não é claro quando a expressão surgiu pela primeira vez, mas faz parte de um conjunto de siglas e acrônimos originados na mídia e empresas de consultorias, como Tech Giants, Superstars (Autor et al., 2020), GAFA, GAFAM, Big Five (van Dijck, 2021), entre outras, que há cerca de uma década definem essas empresas associando-as aos grandes oligopólios que dominam setores tradicionais, como petróleo, farmacêutica e outros., e buscam retratar aspectos distintos, mas que têm em comum a ênfase na escala dessas empresas de tecnologia (Oremus, 2017). O conceito é hoje empregado de forma a abarcar as estrelas americanas e gigantes chinesas.

generativa, sem considerar outros que virão quando de sua efetiva implementação em setores *downstream*.

A despeito de recorrer a fontes diversas e *corpus* heterogêneos da literatura, a linha condutora da pesquisa é a apropriabilidade do valor da inovação, tomando como ponto de partida a estrutura do PFI, analisando de forma mais detalhada o regime legal de PI, tratado como “dado” na abordagem original, além de buscar maior poder explicativo nas características peculiares de ecossistemas e plataformas digitais tratadas em outras peças da literatura.

A abordagem do PFI, seja na versão original (Teece, 1986), seja nas revisitações destinadas a incorporar as características novas da economia digital (Teece 2006, 2012, 2016, 2018a), ainda parece um referencial teórico-conceitual útil para a compreensão do regime de apropriabilidade da inovação das novas tecnologias da IA, adaptado às mudanças impostas pela inovação aberta no contexto de ecossistemas e plataformas digitais. Tal regime, entretanto, poderá ser afetado pelos muitos desafios da intervenção pública com que se depara a IA, relacionados à regulação, à política antitruste e às políticas industriais e comerciais em curso.

Quanto ao regime legal de PI no contexto das tecnologias de IA, parece passar por fricções nas principais economias avançadas, principalmente sobre direitos de patentes de invenção, direito autoral e a ampliação do papel do código aberto⁹ e do segredo comercial. Com o avanço da inovação colaborativa, os DPIs, instrumentos típicos da inovação fechada, passaram a demandar das firmas a criação de mecanismos para a governança da abertura e estratégias diferenciadas de gestão.

OBJETIVOS DA TESE

O objetivo da pesquisa proposta é identificar o regime de apropriabilidade da inovação da IA, mais especificamente no subcampo das técnicas de aprendizado profundo em arquitetura de redes neurais artificiais, a partir dos elementos componentes da estrutura padrão e revisões do PFI, conciliada com outros *corpora* da literatura que analisaram ecossistemas, plataformas digitais e inovação aberta. A pergunta central a ser respondida é sobre a natureza do regime de apropriabilidade das inovações do tipo particular de IA das redes neurais profundas.

⁹ O código aberto (open-source) é um modelo de inovação colaborativo a partir do licenciamento livre que surgiu no setor de software, baseado na liberdade de acessar, executar, estudar, redistribuir e modificar o software usado e distribuir o derivado, respeitando os termos da licença de código aberto. Existem modalidades variadas de licenças, embora no software de código aberto (OSS, sigla de Open Source Software) tenha sido consagrada a definição proposta pela *Open-source Initiative* (OSI) – disponível em <https://opensource.org/osd>, acessada em: 10 ago. 2023 -, segundo a qual as licenças devem cumprir os dez critérios da OSI, com destaque para resistência a restrições de acesso em decorrência de PI. Na IA não existe ainda essa espécie de “convenção” e as modalidades de licenças são ainda mais variadas, embora já exista uma proposta recente da OSI para padronização das licenças de código aberto da IA, como será visto.

Título: Apropriabilidade da Inovação na Inteligência Artificial

Questão da Pesquisa: Qual o regime de apropriabilidade das inovações da Inteligência Artificial em redes neurais profundas?

Objetivos:

Geral: Identificar os elementos componentes do regime de apropriabilidade da inovação relacionada às fases da IA preditiva e da IA generativa

Específicos:

- 1) Identificar os padrões de inovação em IA, inclusive como resultado da pesquisa colaborativa de firmas entre si, instituições de ensino e pesquisa, e instituições governamentais;
- 2) Identificar, descrever e discutir as formas de proteção intelectual empregadas para proteção das tecnologias da IA e as indefinições ainda existentes que podem enfraquecer o regime de PI das invenções e bases de dados de treinamento dos modelos da IA, além de mecanismos informais empregados;
- 3) Explorar a natureza das firmas líderes na inovação da IA, principais depositantes de pedidos de patentes de invenção da IA nos respectivos escritórios de propriedade intelectual (ENPIs) ou que disponibilizam modelos no mercado em código aberto;
- 4) Identificar o padrão de competição no mercado de IA e o papel dos atuais fornecedores de ativos complementares.

JUSTIFICATIVA, RELEVÂNCIA E ORIGINALIDADE

Direitos de propriedade intelectual são uma das mais antigas e polêmicas formas de intervenção pública com vistas a estimular a inovação. Mesmo assim, continuam utilizados pelas firmas, ainda que inseridos em estratégias amplas que contemplam outras variáveis para o sucesso da inovação e o crescimento a longo prazo. Com o surgimento de tecnologias emergentes, a literatura sobre apropriabilidade da inovação vem sendo revisitada (e adaptada) ao contexto da economia digital, da mesma forma que outras peças da literatura sobre inovação – como inovação aberta, ecossistemas, plataformas digitais e antigas abordagens de tecnologias de propósito geral ou habilitadoras - que começam a se entrelaçar.

A originalidade da pesquisa é tentar combinar essas diferentes abordagens com vistas a lidar com o fenômeno complexo e em rápida evolução do padrão de inovação e do seu regime de apropriabilidade nas distintas fases de desenvolvimento da IA e identificando idiossincrasias das empresas inovadoras, complementando análises consagradas de plataformas vencedoras (:

A pesquisa procurou analisar como o regime de PI recepcionou as invenções relacionadas a IA e bases de dados usadas no treinamento dos modelos, particularmente com

vistas a avaliar se corresponde à hipótese do PFI de um regime legal fraco, com riscos de contestação futura de direitos, o que parece contribuir para o avanço de outras formas de proteção da inovação.

O tema da PI relacionada a invenções da IA e bancos de dados subjacentes tem produzido um número importante de estudos, predominante do campo do direito e desde 2019 recebe atenção especial da Organização Mundial da Propriedade Intelectual (OMPI) (WIPO, sigla da expressão em inglês *World Intellectual Property Organization*), que produziu estudos (WIPO, 2019; WIPO, 2024a), elaborar documento onde elenca os aspectos mais desafiadores para o sistema de PI (WIPO, 2020a), consulta pública (WIPO, 2020b) e um total de dez seminários ou sessões anuais ou bianuais, entre 2019 e 2024, disponíveis no Website do fórum *WIPO Conversation on Intellectual Property and Frontier Technologies*¹⁰. Discussões semelhantes também mobilizaram a agenda dos principais escritórios nacionais de PI a partir da virada da década, como o IP5, fórum dos cinco maiores escritórios nacionais de propriedade intelectual, além de outros escritórios importantes, como o Alemanha, Reino Unido, Canadá, entre outros.

O debate, que parecia se tornar mais insípido, é retomado de forma intensa com o surgimento da IA generativa, embora muitas vezes ocorrendo a portas fechadas, e sem uma posição clara, recomendações conclusivas e harmonização internacional de direitos de PI sobre inovações que transcendem fronteiras nacionais de uma tecnologia com impactos indubitavelmente globais como parece ser a IA.

A principal contribuição da pesquisa talvez venha a emergir da investigação da análise da ampla e dispersa literatura. Essa literatura, que começava a se debruçar sobre a implementação da IA em outros setores da economia e potenciais modelos de negócios - inclusive a partir de estudos de caso de plataformas digitais voltadas para setores da economia -, acabou sendo sobrecarregada pela nova onda avassaladora de modelos da IA generativa. O que obrigou a reformulação da tese, com vistas a melhor compreensão dessas novas tecnologias, usos potenciais e possíveis impactos sobre a apropriabilidade.

METODOLOGIA

A pesquisa é de natureza qualitativa e exploratória, que busca analisar um fenômeno ainda em processo de desenvolvimento no contexto real, quando os limites entre o fenômeno e o contexto não estão bem definidos (Yin, 2015; Flick, 2009). O estudo lidou com fenômenos

¹⁰ Disponível em: https://www.wipo.int/about-ip/en/frontier_technologies/frontier_conversation.html. Acesso em: 12 out. 2024

novos e complexos, ainda em formação, analisados por diferentes campos de pesquisa, onde o procedimento técnico principal é a investigação documental de *corpus* heterogêneos de fontes diversas (Godoy, 1995; Sandelowski et al., 2012), considerada adequada para investigar fenômenos novos e mal compreendidos que possuem elementos múltiplos, complexos e que evoluem ao longo do tempo.

A pesquisa documental foi associada por meio de uma estrutura coerente que toma como inspiração a abordagem consagrada da apropriabilidade da inovação, combinada a ecossistemas e inovação aberta, com vistas a construir ligações e identificar pontos críticos para a inovação e o regime de apropriabilidade de uma tecnologia ainda incipiente. A principal contribuição é aprofundar a análise das variáveis do regime de apropriabilidade, explorando a dimensão dos ativos complementares representados pela infraestrutura de computação que só recentemente começou a ser tratada pela literatura especializada desde o surgimento dos grandes modelos generativos. Simultaneamente, buscou-se analisar a força ou fraqueza do regime legal de PI relacionado às invenções da inteligência artificial. Por fim, o surgimento de *designs* tecnológicos dominantes, fenômeno complexo da investigação acadêmica em diversos campos, foi abordado sob a ótica da apropriação de valor. A conexão de diferentes construções teóricas e conceituais com vistas a explicar o mesmo fenômeno resultou em uma síntese conceitual, na forma de estudo de caso múltiplo, rastreando as operações das empresas líderes, baseado em análise de *corpus* documentais.

Muito tem sido escrito sobre as chamadas BigTechs, termo consagrado na mídia e nas grandes firmas de consultorias, tratadas como um conjunto uníssono de empresas que operam modelos de negócios de plataformas digitais e, de uma forma crescente, se destacam no campo da IA. No entanto, como regimes de apropriabilidade não ocorrem no vazio, a unidade de análise da tese é a firma, mesmo que algumas vezes a pesquisa avance sobre conceitos meso- e macroeconômicos que tenham implicações sobre a arquitetura da indústria, o padrão de competição e o regime de PI, com impactos sobre a firma e seu regime de apropriabilidade, sem a pretensão de esgotar essas dimensões mais amplas.

Em alguma medida, a tese procura contribuir para a compreensão das vantagens competitivas das firmas inovadoras líderes em IA, complementando análises de plataformas vencedoras (Yoffie, Gawer e Cusumano, 2019) e do caráter diferenciado da PI na IA. Adicionalmente, indícios de estratégias empresariais diversas exigiram que a investigação se apoiasse em extensivo monitoramento das operações das notórias firmas BigTechs e de empresas emergentes de destaque. Isso foi improvisado por meio da análise de dados heterogêneos de fontes documentais diversas. Contemplou desde artigos de pesquisa visando a

obtenção de informações técnicas sobre modelos de IA, coletada na literatura especializada, sobretudo na base de distribuição gratuita e acesso aberto de pré-prints de artigos ainda não revisados/publicados disponibilizados no website arXiv¹¹, da Universidade de Cornell, consagrada como principal veículo de divulgação das invenções algorítmicas e modelos de IA; na base de dados também livre e gratuita do instituto de pesquisa multidisciplinar Epoch¹²; em estudos de consultorias especializadas e organizações de pesquisa que se dedicam à investigação da trajetória da IA, como o State of AI Report¹³ e Semianalysis¹⁴, além das publicações anuais do AI Index Report do importante Stanford Institute for Human-Centered Artificial Intelligence (HAI)¹⁵, entre outras instituições que serão mencionadas ao longo da tese; bem como artigos de mídia e dados comerciais públicos das empresas. Documentos de política industrial e comercial, tratados internacionais de comércio, iniciativas regulatórias e da política antitruste complementaram a análise, tendo em vista seu potencial de impactar a arquitetura da indústria, o padrão de inovação e competição e, conseqüentemente, o regime de apropriabilidade da inovação.

A despeito da dificuldade de teorizar sobre um fenômeno recente, cuja evolução ainda está em curso e com grandes incertezas inerentes, a pesquisa procurou avançar na discussão dos temas fundamentais para a criação e captura de valor da inovação em IA. A contribuição é baseada na discussão de aspectos técnicos dos modelos de redes neurais profundas, um melhor entendimento dessa promissora tecnologia cuja complexidade exigiu a estruturação de amplos ecossistemas de plataformas digitais, ainda que seu objeto não seja o estudo desses modelos de negócios *per se*.

A identificação dos aspectos conceituais e teóricos da tese exigiu a elaboração de uma metodologia própria, que interrelaciona a literatura sobre IA, PI e apropriabilidade da inovação, complementada por achados relacionados à defesa da concorrência, regulação e políticas públicas. A sistematização de diferentes campos da literatura teórica foi complementada por documentos base, como o Acordo TRIPS e Convenções internacionais, legislações nacionais de patentes (Estados Unidos, União Europeia e China¹⁶), documentos de diretrizes de exame de patentes dos principais Escritórios Nacionais de PI (ENPIs) – *United States Patent and*

¹¹ Disponível em: <https://arxiv.org/>. Acesso em: 20 mar. 2022

¹² Disponível em: <https://epochai.org/>. Acesso em: 6 jul. 2024

¹³ Disponível em: <https://press.airstreet.com/p/state-of-ai-report-5-years>. Acesso em: 12 mar. 2021

¹⁴ Disponível em: <https://www.semianalysis.com/>. Acesso em: 6 jul. 2024

¹⁵ Disponível em: <https://aiindex.stanford.edu/report/>. Acesso desde 2019.

¹⁶ EUA e China são as nações hegemônicas na inovação da IA e líderes em propriedade intelectual relacionada ao tema, enquanto a União Europeia vem tendo a liderança regulatória da IA e temas correlatos, com perspectivas de produzir impactos globais por meio do *Brussels effect*, expressão cunhada por Bradford (2012, 2020).

Trademark Office (USPTO), *European Patent Office* (EPO) e *China National Intellectual Property Administration* (CNIPA) -, bem como artigos de mídia, relatórios e documentos públicos das empresas, incluindo blogs, relatórios, documentos e apresentações em conferências da Organização Mundial da Propriedade Intelectual (WIPO, sigla da expressão em inglês *World Intellectual Property Organization*) e dos principais ENPIs, além de informações quantitativas de fontes secundárias.

A revisão da literatura foi em grande medida responsável pela revisão do tema e do escopo inicial da pesquisa. Por um lado, obrigou a ampliação do seu escopo original, antes restrita aos instrumentos legais de propriedade intelectual, passando a incorporar as variáveis da estrutura do PFI e conceitos nascentes de ecossistemas, plataformas digitais, inovação aberta e tecnologias de propósito geral que se entrelaçam nas discussões sobre a apropriabilidade da inovação da IA.

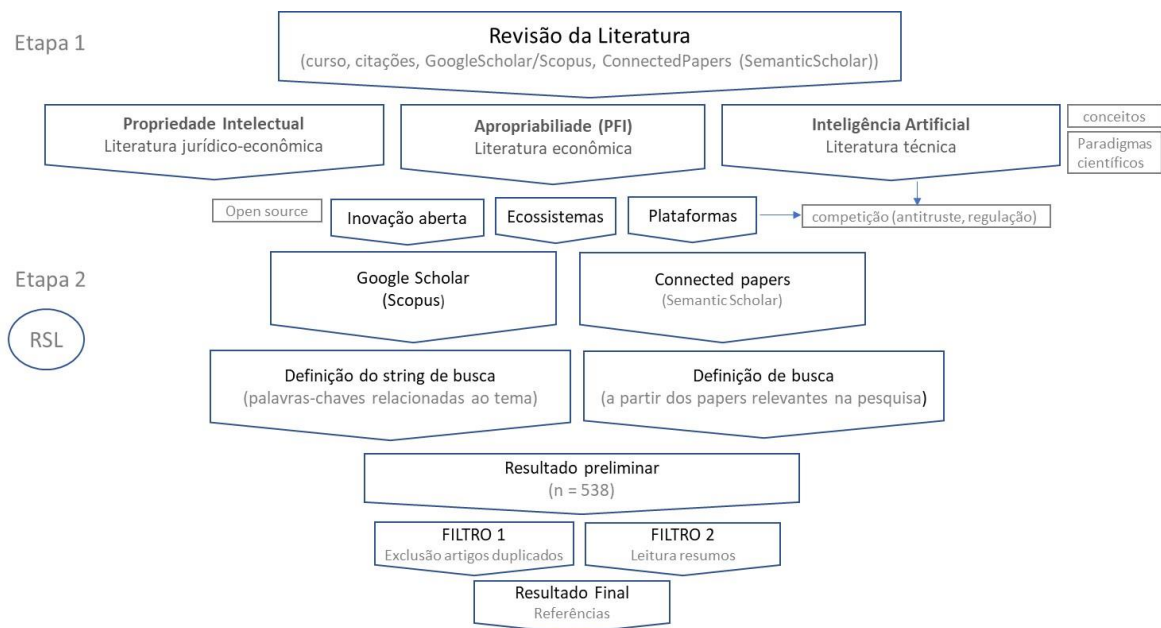
A revisão da literatura foi feita em duas etapas, conforme a Figura A. A pesquisa bibliográfica tomou por base as referências das disciplinas do curso de doutorado de Pós-Graduação *Stricto sensu* em Propriedade Intelectual e Inovação do INPI (PPGPI) da Academia de Propriedade Intelectual, Inovação e Desenvolvimento (ACAD) do INPI. A partir daí, foi realizada a busca “livre” das citações contidas nessa literatura e daquela relacionada aos temas-chaves da pesquisa. No tema PI, a revisão contemplou, a literatura jurídico-econômica convencional, além de buscas livres no website da SSRN/Elsevier, relatórios, leis, documentos e diretrizes de exame de patentes disponibilizados nos websites dos respectivos escritórios nacionais de PI (ENPIs) e em conferências da OMPI, EPO, USPTO e CNIPA. No tema da IA, a revisão contemplou publicações disponibilizadas no projeto de pesquisa em IA do National Bureau of Economic Research (NBER), da Organisation for Economic Co-operation and Development (OECD), arXiv, além de artigos e relatórios de organizações especializadas, como o mencionado AI Index Report do HAI/Stanford e instituições de pesquisa, artigos apresentados em conferências, teses e dissertações, cotejados com achados até mesmo de consultorias como McKinsey & Co., Boston Consulting Group/Henderson Institute e Goldman Sachs.

O tema da apropriabilidade da inovação exigiu uma pesquisa mais estruturada, que corresponde a segunda etapa da revisão bibliográfica, com vistas a identificação de publicações mais recentes que pudessem considerar as inovações da IA. Neste caso foram implementados mecanismos de revisão sistemática da literatura a partir de duas fontes de informações. Foi pesquisado o website Google Scholar¹⁷, que contém “metadados” da literatura acadêmica da

¹⁷ Disponível em <https://scholar.google.com/>. Acesso em 29 set. 2024

base Scopus, com filtro por algoritmo de classificação que permite a busca orientada (livros, publicações, revisados ou não, artigos de conferências, pré-impressões, teses e dissertações, resumos, relatórios técnicos) a partir de palavras-chaves, definido o período de 1998, data da fundação do Google, até hoje). A definição do “string de busca” contemplou as expressões, no idioma inglês, em várias combinações e truncagens, de “*artificial intelligence*”, “*deep learning*”, “*artificial neural networks*” (e os nomes dos principais tipos dessas redes, como “*convolutional neural networks*”, “*recurrent neural networks*”, “*generative adversarial networks*”, “*Transformers*”, entre outras), “*digital platforms*”, “*ecosystems*”, “*general-purpose technologies*”, “*enabling technologies*”. Por outro lado, foi pesquisado no website Connected papers,¹⁸ ferramenta visual que relaciona papers por citação ou por similaridade a partir de um algoritmo de cocitação e acoplamento, a partir da base Semantic Scholar. O uso da ferramenta foi feito para os artigos considerados centrais na abordagem empregada nesta tese. A utilização dessas bases, apesar de suas limitações técnicas, menos filtros e algoritmos simples, do que bases como a Web of Science/Clarivate, tem a vantagem de permitir pesquisa livre de pagamento (ou *freemium* no caso do Connected papers).

Figura A – Representação gráfica da estratégia de revisão da literatura



Fonte: Elaboração própria

À medida que a revisão da literatura avançava, alcançavam-se resultados importantes, mas outros talvez muito mais importantes pareciam fora de alcance. Por conseguinte, o objeto da pesquisa se modificou quase que de forma autônoma contemplando além da pretendida avaliação do passado e do presente, o estímulo para também vislumbrar o futuro. Isso levou à

¹⁸ Disponível em <https://www.connectedpapers.com/>. Acesso em 29 set. 2024

escolha de uma abordagem empírica mista, a partir do levantamento exaustivo de modelos de redes neurais profundas. Esses dados primários objetivaram mapear o lançamento dos modelos, a partir do extenso levantamento do instituto multidisciplinar da Epoch como *proxy* das tendências da inovação relacionada à IA, em especial a partir de 2010 e, principalmente, 2017, como forma de identificar a IA generativa, até os dias atuais.

O levantamento compreenderia também pedidos de patentes depositados ao redor do mundo, para o mesmo período. No entanto, a divulgação de recente relatório da OMPI (WIPO, 2024a) evidenciou a precariedade de empregar esse tipo de informação e análises patentométricas para prospecção tecnológica no caso de tecnologias que avançam em ritmo tão acelerado como a IA e conflitam com os prazos de sigilo e concessão de patentes. A velocidade da inovação da IA, sem qualquer precedente histórico, principalmente no subcampo da IA generativa nos últimos dois anos, decorrente e explicada pela expansão desenfreada do regime *open-source*, mostrou as limitações de utilizar informações de patentes, que mostram resultados defasados de invenções e empresas líderes, principalmente para tecnologias para as quais sequer foi ainda definido um *design* tecnológico dominante.

ESTRUTURA DA TESE

A tese inclui, além deste capítulo introdutório, seus objetivos, justificativa e metodologia, sete outros capítulos, além das conclusões.

O Capítulo 1 apresenta uma breve revisão da literatura de propriedade intelectual, avançando para abordagens que discutem a questão da PI em indústrias como TICs, passando pelas questões relevantes para a inovação na economia digital, como conjuntos de dados, tecnologias de interoperabilidade e a noção de inovação aberta e os conceitos de ecossistemas de plataformas digitais. Finalmente, é apresentada a abordagem do regime de apropriabilidade do PFI, fio condutor da análise das inovações da IA desta tese.

O Capítulo 2 analisa como a estruturação do robusto ecossistema de inovação, avanços técnicos do aprendizado profundo em arquiteturas de redes neurais artificiais, conjugados à principal inovação organizacional das plataformas *online* operadas pelas empresas BigTechs, levaram às exitosas inovações da popularmente chamada IA preditiva. É destacada a importância da inovação aberta e da extensa rede de pesquisa colaborativa estruturada por essas empresas. O capítulo inicia com a descrição da longa trajetória da IA, que enfrentou fases distintas até a consolidação das redes neurais profundas como paradigma de pesquisa, a partir das invenções algorítmicas, estruturação de amplos conjuntos de dados e avanços em capacidade computacional, junto a mudanças institucionais, que levaram a transformação do

ecossistema de inovação de natureza pública para outro privado, resultando em inovações efetivas no mercado desde a década passada.

O Capítulo 3 discute as mudanças iniciadas há poucos anos, com o avanço de uma nova classe de modelos de redes neurais profundas cada vez mais poderosos, da denominada IA generativa. Esses modelos proliferaram desenfreadamente nos últimos dois anos, agregando ao cenário quase exclusivo das empresas líderes da fase anterior, um conjunto amplo de empresas emergentes, sugerindo uma nova mudança de paradigma, principalmente dos chamados ativos tangíveis da IA.

O Capítulo 4 apresenta as principais formas de intervenção pública que as tecnologias da IA estão enfrentando ao redor do mundo, com perspectivas de influenciar a arquitetura da indústria e, conseqüentemente, impactar o padrão de apropriabilidade, de concorrência e da própria inovação.

O Capítulo 5 discute o regime legal de PI das invenções relacionadas a inteligência artificial e dos seus principais insumos, como grandes conjuntos de dados que são essenciais para o treinamento dos modelos de IA. Tal regime legal é extenso e complexo, de natureza essencialmente híbrida, que combina múltiplos direitos legais de PI (direito autoral, direitos *sui generis* e patentes, entre outros instrumentos como marcas, inclusive submarinas) sobre a invenção propriamente dita e sobre as bases (ou conjuntos) de dados, que convive harmoniosamente com regime de código aberto e segredos comerciais.

O Capítulo 6 apresenta evidências empíricas da produção da IA e dos principais inovadores, elaboradas a partir da ampla e atualizada base de dados do instituto de pesquisa multidisciplinar Epoch, que acompanha quase em tempo real o lançamento de novos modelos da IA, mostrando um avanço recente sem precedente histórico de invenções nesse campo de conhecimento. Essas evidências são confrontadas com dados recentes de patentes da IA generativa divulgados pela OMPI, com vistas a mostrar suas limitações enquanto indicadores de tendências tecnológicas em um campo ainda em desenvolvimento e cujo ritmo acelerado parece incompatível com os prazos e características dos instrumentos formais de PI.

A comparação desses dois tipos de indicadores levou à própria revisão da pesquisa da tese ao evidenciar que a análise patentométrica da IA em setores selecionados não produziram informações relevantes de prospecção tecnológica. Os dados levantados mostraram que um estudo prospectivo da identificação do direcionamento (setorial) parece prematuro, diante das mudanças representadas novos modelos generativos e outros ainda mais incipientes. As principais redes neurais artificiais da IA preditiva, ConvNets (sigla de *Convolutional Neural Networks*, cuja aplicação consagrada é em reconhecimento de imagem) e RNNs (sigla de

Recurrent Neural Networks, usada em processamento de linguagem natural - PLN), parecem ter sido suplantadas, primeiro pelas GANs (sigla de Generative Adversarial Networks) e, depois, por modelos de difusão, nas aplicações funcionais de imagem e, principalmente pelas revolucionárias arquiteturas Transformers em PLN. Com isso, as principais patentes de invenção de aprendizado profundo em setores como saúde, empregavam ConvNets, que não corresponderiam mais a fronteira do conhecimento – motivo que talvez explique sua disseminação em equipamentos de imagem, como tomógrafos computadorizados e ressonância magnética, onde alcançam acuracidade superior a 95%, explicando o importante número de pedidos de patentes por fabricantes de equipamentos que já incorporam essas técnicas de IA. Com isso, a análise patentométrica acabaria por sinalizar tendências que não necessariamente subsistirão ao enorme potencial dos novos modelos generativos.

No Capítulo 7 são analisadas as variáveis do arcabouço teórico-conceitual do PFI para os casos específicos da inovação da IA preditiva e da nova fase da IA generativa. O capítulo evidencia mudanças importantes no regime de apropriabilidade, que parece dar início a uma nova disputa pelo *design* tecnológico dominante e da reestruturação das estratégias empresariais relacionadas aos ativos/tecnologias complementares, em particular as infraestruturas de capacidade computacional que se tornam cada vez mais essenciais não só para a captura de valor, como para a própria criação de valor da inovação em IA.

As conclusões da tese são sintéticas, apontando para a necessidade de estudos complementares, inclusive considerando as novas tendências relacionadas a agentes inteligentes e as prometidas AGIs (sigla de *Artificial General Intelligence*). Algumas reflexões dizem respeito aos desafios de apropriabilidade para empresas emergentes, sobretudo diante da brutal escalabilidade dos modelos e da crescente importância dos ativos complementares relacionados a chips, datacenters e hardware em geral, aos muitos desafios regulatórios com efeitos imprevisíveis sobre o ecossistema e regime de apropriabilidade da IA, particularmente em meio a crescentes disputas geopolíticas, além de umas poucas recomendações para economias com menor grau de desenvolvimento e sugestão de temas de pesquisas futuras.

CAPÍTULO 1 – REVISITANDO A PROPRIEDADE INTELECTUAL NO CONTEXTO MAIS AMPLO DA APROPRIABILIDADE

Inovações tecnológicas são fontes de desenvolvimento econômico e elementos centrais da competição entre as firmas com vistas a alcançar e preservar mercados, auferir lucros extraordinários e crescer no longo prazo. Em termos macroeconômicos, as inovações estão na base da expansão do sistema capitalista e podem provocar rupturas e mudanças profundas em toda a estrutura econômica. Em termos microeconômicos, a forma crucial da competição é baseada na capacidade de inovar, não em preços:

"[...] in capitalist reality as distinguished from its textbook picture, it is not that kind of competition which counts but the competition from the new commodity, the new technology, the new source of supply, the new type of organization (the largest-scale unit of control for instance) - competition which commands a decisive cost or quality advantage and which strikes not at the margins of the profits and the outputs of the existing firms but at their foundations and their very lives." (Schumpeter 1942:84). E ainda "Every piece of business strategy acquires its true significance only against the background of that process and within the situation created by it. It must be seen in its role in the perennial gale of creative destruction; it cannot be understood irrespective of it or, in fact, on the hypothesis that there is a perennial lull" (Schumpeter, 1942:83-4).

No entanto, as vantagens competitivas das firmas dependem da sua capacidade de inovar que, para tal, cada vez mais competem por capacidades dinâmicas que se reflete em produzir e controlar ativos intangíveis (Buainain et al., 2004), aumentando a importância da propriedade intelectual como instituição para dar proteção e facilitar a valorização econômica dos ativos intangíveis.

Direitos de propriedade intelectual (DPI) são considerados importantes instrumentos de intervenção governamental respaldados pela teoria econômica convencional e pelas teorias do direito (Barbosa, 2010; Posner, 2001). O sistema legal de PI corresponde ao arcabouço legal visando proteger criações do intelecto humano, conferindo a autores e inventores um direito exclusivo sobre sua obra ou invenção. É dividido em dois grandes grupos, a propriedade industrial – que contempla marcas, patentes de invenção e modelos de utilidade, desenho industrial e indicações geográficas - e direitos autorais, que protegem obras artísticas, literárias e científicas, incluindo programas de computador e bases de dados subjacentes, além de direitos *sui generis*, destinados a casos não são adequadamente contemplados pelos demais instrumentos.

A dimensão da PI de transitar pelos campos do direito e da economia, é como as duas faces de uma moeda. Uma dessas faces, do do direito, reconhece, por meio de leis específicas, direitos de propriedade às criações do intelecto humano, enquanto a outra face, do campo da economia, possibilita que invenções e DPIs subjacentes se traduzam, em termos microeconômicos, em lucros extraordinários, expansão de mercados e ampliação do valor da

firma inovadora expresso em capitalização de mercado.

O aparente monopólio jurídico muitas vezes atribuído a PI não se traduz, contudo, necessariamente, em um monopólio econômico. O próprio Schumpeter é taxativo:

“New methods of production or new commodities, especially the latter, do not per se confer monopoly [...]. The product of the new method has to compete with the products of the old ones and the new commodity has to be introduced, i.e., its demand schedule has to be built up. As a rule, neither patents nor monopolistic practices avail against that.” (...). Isso ocorre somente em casos “[...] of spectacular superiority of the new device [...]” (Schumpeter 1942:102).

Nesse sentido, as firmas não devem sofrer do que Laursen e Salter (2005) chamaram de miopia de protecionismo, confiando demais na apropriação em detrimento de outras atividades. Como lembram Dosi e Stiglitz (2013:11), a reflexão crítica sobre os mecanismos de apropriabilidade parece se perder em uma parte da literatura sobre inovação, presumindo que a única forma eficiente é por meio de DPIs, mesmo na ausência de evidências que sustentem uma relação positiva entre PI e taxa de inovação, principalmente em regimes rígidos de PI mal projetados.

Evidências empíricas repetidamente demonstraram que o regime legal de PI, particularmente patentes, não são, muitas vezes, verdadeiramente cruciais para a inovação (Levin et al., 1987; Cohen et al., 2000; Cimoli et al., 2011), exceto em alguns poucos setores econômicos. Mesmo nesses, patentes em particular podem, em alguns casos, ser contornadas (“*inventing around*”), a custos modestos, sugerindo que poucos setores parecem depender efetivamente das leis de patentes para capturar os retornos da inovação (Mansfield, 1984).

A natureza em boa parte intuitiva, tácita e não codificada do conhecimento, cria frequentemente condições de apropriabilidade da inovação na ausência de proteção por PI, por meio de mecanismos alternativos como segredos comerciais e industriais¹⁹, *know-how*, *lead time*, vantagens da curva de aprendizado e, usualmente, a mera complexidade dos produtos, que dificultam a imitação mesmo na ausência de proteção legal (Dosi et al., 2006b).

Tanto na economia, como no direito, existem justificativas distintas para os DPI. Nos Estados Unidos a lei tem sido amplamente baseada na abordagem utilitarista, onde, em tese, a legislação da concorrência contrabalançaria dos detentores de direitos que poderiam bloquear inovações subsequentes ou impor custos excessivos, de forma a proteger a sociedade (Lianos e Dryfuss, 2013).

¹⁹ Em alguns países há legislação específica relativa a *trade secrets*, consideradas entre os DPIs, como o *US Defense of Trade Secrets Act – DTSA/2016* – disponível em <https://www.uspto.gov/ip-policy/trade-secret-policy>. Acesso em 05 fev. 2021 - e a Diretiva Europeia 2016/943. – Disponível em <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0943&from=PT>. Acesso em 04 fev. 2021

Apesar de tão contestado ou, pelo menos, ter sua importância relativizada por uma parcela grande da literatura teórica e empírica, o instituto da PI sobrevive por quase seis séculos e é cada vez mais espalhado pelo mundo, amparado pelo *mainstream* econômico.

Entre os instrumentos de PI, talvez o mais controverido sejam as patentes. Diversos autores exploraram evidências empíricas sobre a racionalidade do sistema de patentes que ora aparece como incentivo às atividades inventivas, ora como contrapartida à revelação do conhecimento. A racionalidade desse sistema reflete-se nos requisitos formais e nas condições exigidas para o patenteamento, em particular a descrição detalhada da invenção no ato do depósito, que deve ser suficiente para ser reproduzível por uma pessoa qualificada (suficiência descritiva). Esse detalhamento da invenção – além, obviamente, de seus custos e requisitos formais – é muitas vezes apontado pela literatura como a razão para as empresas preferirem mantê-la em sigilo ou protegida por outros instrumentos.

Como qualquer forma de propriedade, é afetada pela natureza dos bens e suas características intrínsecas, não existindo sem o bem, seja ele físico (tangível) ou imaterial (intangível). Para o *mainstream* econômico a PI é um tipo específico de atribuída a um tipo particular de bem, o bem público. No entanto, a opção pelo conceito de bem, no sentido de mercadorias que são entes usados para fins de transação comercial, ao contrário de tratar a inovação com conceitos mais realistas e amplos como um ativo, que são recursos cujo valor econômico está no seu uso para geração de receitas e valorização futura, acabou dificultando a compreensão de outras funções da PI que parecem bem mais efetivas e atuais, como moeda de barganha, uso defensivo e valorização da empresa.

1.1. *Mainstream* econômico e a visão heterodoxa da propriedade intelectual

A teoria econômica convencional classifica invenções, do mesmo modo que conhecimento, informação e dados, como um tipo de bem, o chamado bem público, que não pode ser produzido eficientemente pelo setor privado devido a falhas de mercado (Samuelson, 1954), justificando, assim, a intervenção governamental (seja através do financiamento, da tributação, da produção direta ou da concessão de DPI). Os bens públicos devem ser amplamente disponíveis (o que na terminologia econômica, significa que não são excludentes) e não envolvem competição porque sua oferta não é insuficiente e não há escassez (na terminologia econômica, não são rivais)²⁰. Bens públicos se diferenciam dos demais por gerar

²⁰ Exemplos de bens públicos são o policiamento, a defesa e outros onde ninguém deve ser excluído do consumo mesmo sem contribuir para o seu custo de produção, que é suportado por poucas pessoas, dando origem ao *free-rider problem*. Eles diferem dos bens privados, onde os mecanismos de mercado operando de forma eficiente por meio dos preços negam o acesso a alguns indivíduos, e seu consumo por alguns indivíduos drena o consumo de outros, sendo assim rivais. Existem outras categorias de bens entre esses dois extremos, os bens quase públicos

benefícios em bem-estar, com retornos sociais superiores ao retorno privado (pois geram *spillovers*, ou externalidades) (Mansfield et al., 1977) e em virtude dos seus atributos de não excludência e não-rivalidade, firmas privadas não estão dispostas ou são incapazes de fornecê-los em quantidades socialmente ótimas.

O conhecimento é não rival e não excludente, tornando-se um bem público (Arrow, 1962; Stiglitz, 1999), e como tal o conhecimento tem sido muitas vezes associado à informação (Archibugi e Filippetti, 2015). Por essa abordagem, invenções são consideradas não excludentes porque não é possível excluir terceiros do uso do bem mesmo sem pagar seu custo de produção, além de não rivais porque o consumo de um indivíduo não reduz a quantidade disponível para os demais. Embora a produção de conhecimento possa ser cara, o custo marginal de compartilhar informações com um indivíduo adicional é zero (Lévêque e Ménière, 2004).

Os fundamentos econômicos dos DPI baseiam-se nesse argumento padrão e o Estado intervém concedendo direitos de propriedade sobre bens intelectuais, especialmente patentes, criando assim uma escassez artificial (Dosi et al., 2006a) que garante o direito de exclusão de terceiros e, por meio da cobrança de *royalties* pelo uso dos direitos, o problema de não rivalidade.

A racionalidade da PI no caso das patentes foi tratada pelo *mainstream* sob a perspectiva da eficiência alocativa. De acordo com a teoria econômica convencional, as inovações envolveriam falhas de mercado decorrentes de externalidades e dificuldades de apropriação privada dos resultados, que estariam na base do surgimento de sistemas de PI, possibilitando a remuneração de investimentos em pesquisa e desenvolvimento (P&D), lucros de monopólio e exclusão de terceiros. Por um lado, sem direitos de PI não seria possível excluir terceiros que não contibuíram com o custo de produção de usar o bem; por outro, seu consumo por um indivíduo não reduz a quantidade disponível para terceiros. Essas duas “falhas” características de um bem público, não excludente e não rival, são enfrentadas pela PI, sobretudo patentes, resolvendo o problema da não exclusão, ao assegurar o direito de excluir terceiros e afastar o risco de imitação por *free-riders* e, através da cobrança de *royalties* pelo uso da tecnologia, o problema da não rivalidade.

(ou bens públicos incompletos ou impuros) que atendem apenas parcialmente às características mencionadas. São os bens comuns, não excludentes, mas rivais, como recursos naturais ou minerais esgotáveis; e bens do clube, um tipo particular de bem excludente, mas não rival, onde indivíduos (não pertencentes ao “clube”) podem ter seu consumo negado, como parques, assinaturas de TV etc. Sobre bens de clube, ver Ostrom, Hess (2008).

Os DPI, sobretudo patentes, ao conferir um direito exclusivo por determinado período, contornaria essas duas falhas de mercado típicas de bens públicos, eliminando o problema do “carona” ao obrigar usuários a pagar *royalties* ao titular, garantindo que inventores se apropriem do valor de suas criações. Há um esforço permanente de equilibrar o incentivo a inovação e a difusão de seus resultados, numa aparente contradição traduzida em termos econômicos no *trade-off* entre eficiência estática e dinâmica. É, portanto, um instrumento imperfeito, marcado por ineficiência estática no curto prazo - pelo custo inerente a PI imposto à sociedade durante o período de exclusividade (peso morto) -, compensada por ganhos de eficiência dinâmica no longo prazo - ao estimular inovações que beneficiam a sociedade (Pinto, 2009).

A importância das patentes é controvertida na literatura. A economia dominante reconhece a intervenção pública quando o mercado falha. Devido às características de bem público do conhecimento e da informação subjacentes à inovação (não excludentes e não rivais), terceiros podem se beneficiar, com perda de bem-estar social, uma vez que sem a proteção da PI o conhecimento fluiria para a sociedade, reduzindo a apropriação privada da inovação e desencorajando a P&D. Patentes resolvem o duplo problema de não exclusão e não rivalidade, pela concessão de direitos exclusivos e estimulando o investimento em P&D que de outra forma não ocorreria.

A abordagem teórica da eficiência alocativa e das falhas de mercado foi amplamente contestada pela literatura econômica heterodoxa, da escola evolucionária, a partir de uma visão radicalmente distinta do processo de inovação, desde Arrow (1962), Nelson (1959) e outros, com implicações sobre a apropriação privada do conhecimento e a direção e ritmo da inovação, oportunidades específicas de determinado paradigma tecnológico (Dosi, 1982).

A abordagem heterodoxa considera que existem formas altamente diferenciadas de conhecimento que influenciam seu entendimento como bem privado ou público (Nelson, 1959), rejeitando a comparação entre informação e conhecimento porque, dependendo das características específicas de como é transferido, o consumidor pode experimentar altos custos de transação, minando a suposição de que todo conhecimento pode ser considerado um bem público.

O conhecimento tem características diferentes de bens, envolvendo indivisibilidade, cumulatividade, desenvolvimento específico e, principalmente, incerteza fundamental. Com isso, raramente há apropriabilidade completa pelo primeiro inovador, mesmo que conte com direitos de propriedade sobre sua invenção, porque a maior parte do conhecimento repousa em bases tácitas e não codificadas, onde mecanismos informais de proteção são usualmente mais eficazes. Dosi et al. (2006a:5) assim resumem o problema da visão convencional: “... *the*

(misleading) identification of knowledge with information [...] makes one forget that processes through which new knowledge is generated are strongly dependent on the specificities of each technological paradigm”.

Estudos empíricos mostraram que as condições legais de apropriabilidade têm efeito limitado sobre a inovação, com patentes menos eficazes em comparação com as mencionadas vantagens da curva de aprendizado, *lead time*, sigilo e capacidades complementares, exceto em poucos setores e tecnologias discretas de produtos. Além disso, Cohen et al. (2000) constataram que a decisão de não patentear invenções decorre da necessidade de demonstrar novidade (em 32% dos casos), divulgar informações (24%) e da facilidade de *inventing around* (25%).

No entanto, patentes, como lembram Macedo e Barbosa (2000), podem ser vistas como um contrato entre o inventor e a sociedade, onde o Estado concede um monopólio (jurídico), representado pelo direito exclusivo sobre a invenção por determinado período, em troca da sua divulgação, que de outro modo poderia ser mantida em sigilo. Representa a recompensa pelo conhecimento que passará para a sociedade após expirada a validade da patente, quando se torna de domínio público. A exclusividade no uso de uma invenção patenteada é a maneira da sociedade reconhecer esse direito de propriedade, impedindo terceiros de vender produtos que incorporem a mesma ideia.

Essa abordagem corresponde a um dos argumentos que sustentam a racionalidade das patentes analisados por Machlup e Penrose (1950), como um contrato social em que o Estado concede direitos exclusivos sobre a invenção em troca de sua divulgação (*disclosure*), de modo possibilitar a disseminação da informação, a ampliação do estoque de conhecimento da sociedade e o progresso econômico, mas com um claro *trade-off* entre direitos exclusivos e divulgação.

Esta função ou racionalidade do sistema de patentes remete ao chamado dilema de Arrow (1962) e o desafio de alcançar simultaneamente incentivos ótimos tanto para produção quanto para disseminação do conhecimento, dois objetivos aparentemente conflitantes do sistema de patentes (Pénin e Neicu, 2018). Esse dilema não está present em outra dimensão dos direitos de PI, o direito autoral, forma de proteção escolhida para os programas de computador, que não possui um sistema de registro obrigatório, com a proteção concedida desde o ato da criação da obra e automaticamente concedida ao autor. Com isso, dispensa um elemento importante que é sua divulgação, com implicações óbvias para a ampliação do conhecimento pela sociedade, enfrentado por meio de movimento do software de código aberto.

A PI sempre foi um tema envolto em grande controvérsia, seja sobre essa racionalidade, seja sobre sua eficácia no estímulo à inovação. Invenções chegam ao mercado, a partir de

decisões tomadas por firmas, na forma de inovações tecnológicas capazes de gerar retornos econômicos. Embora nos primórdios dos direitos de propriedade intelectual a figura do inventor pudesse ser confundida com a do titular dos direitos, com o tempo se tornaram *players* cada vez mais distintos (Coriat e Weinstein, 2009:6-7).

Mudanças profundas, a partir da década de 1980, primeiro nos EUA e depois ao redor do mundo com o Acordo TRIPS, levaram a um desenvolvimento sem precedentes da PI, com o conhecimento transformado em ativo estratégico das firmas, admitindo patentes sobre tópicos antes explicitamente excluídos (software e genes e matéria viva) (Coriat e Orsi, 2002; Andersen, 2004). Mais tarde, ocorreria um salto na mercantilização do conhecimento por meio de condições permissivas para a participação do capital financeiro (*venture capital*) na produção de conhecimento (financiando até mesmo firmas inovadoras deficitárias, com DPI como ativos intangíveis²¹) (Coriat e Weinstein, 2009:17). Hoje, proprietários dos DPI são principalmente firmas, capitalistas de risco, investidores financeiros e até mesmo universidades, que “representam” os inventores por meio de relações contratuais (principalmente contratos de trabalho) e “intermediam” DPI.

Alguns autores identificaram nessas estratégias empresariais mecanismos de apropriação privada da ciência e do conhecimento, frequentemente atacada pela abordagem heterodoxa, uma vez que “*science base largely is the product of publicly funded research and the knowledge produced by that research is largely open and available for potential innovations to use*” (Nelson, 2003:1).

Pénin e Neicu (2018:58) também criticam a abordagem das patentes como forma de exclusão de imitadores, baseada em uma visão irrealista do processo de inovação - resultado da ação individual e isolada para produção de um bem público, onde conhecimento equivale a informação (de fácil reprodução) e que negligencia sua dimensão tácita e o processo parcialmente coletivo de geração que requer colaborações e interações. Os autores enfatizam a ligação entre patentes e inovação aberta (*open innovation*), cada vez mais apontada como a realidade da maioria dos setores.

Em síntese, apesar de DPI, especialmente patentes, serem temas controvertidos, as firmas têm se apoiado fortemente nesse regime jurídico, compondo portfólios amplos e diferenciados desses ativos intangíveis, muitas vezes sobrepostos, para proteger invenções e os

²¹ Além de patentes, marcas registradas, direitos autorais e outras propriedades intelectuais são exemplos comuns de ativos intangíveis que as firmas podem possuir ou controlar para gerar receitas ao longo do tempo. Hoje também incluem conjuntos de dados, provavelmente o ativo mais valioso da economia digital, ainda que de difícil identificação no balanço das firmas. Birch et al. (2021) e trabalhos mais recentes (WEF 2021; Coyle et al. 2020; Coyle, Manley 2022) têm buscado sua mensuração.

conjuntos de dados que permitem a eficiência e precisão dos sistemas de IA. Nesse sentido, se convertem em mecanismos para orquestrar novos ecossistemas de inovação aberta com potencial de expansão na fase de implementação da IA e valorização de mercado das firmas que os detêm. Gradualmente, amplas carteiras de patentes tornam-se um dos ativos intangíveis mais importantes na avaliação do mercado das firmas. Bessy (2019:5) encara esse papel distinto como uma nova convenção do uso das patentes, que se soma às várias convenções prévias e depende apenas de um mínimo de *players* aderir coletiva e espontaneamente ao novo uso das patentes.

1.2. Propriedade Intelectual nas tecnologias de informação e comunicação

A indústria de TICs se caracteriza por produtos complexos de alta tecnologia, de natureza transversal, em que a difusão é tão fundamental quanto a inovação em si, constituindo um caso peculiar com respeito aos DPI. Dosi et al. (2006a) constataram que mesmo invenções não patenteadas ou patenteadas sob regimes fracos obtiveram retornos econômicos significativos nessa indústria. Isso ocorre pelo típico curto ciclo de vida dos produtos e ritmo acelerado da inovação, que reduzem a importância das patentes e tornam sigilo e *lead time* mecanismos mais efetivos. Quando invenções são patenteadas, frequentemente visam bloquear o uso de componentes de rivais e adquirir força de barganha em negociações de licenciamento cruzado, diferentemente de produtos discretos, usadas para bloquear substitutos (Dosi, 2006a:12).

Este foi o caso no início da indústria de semicondutores (compensados por outra forma de intervenção pública, as compras públicas) e da telefonia móvel, em que vigoraram regimes de PI fracos até o final da década de 1980. O intenso patenteamento e grande expansão dos portfólios pelas firmas nas décadas seguintes não se destinaram ao retorno ou proteção da inovação, mas seu uso defensivo, como “moedas de barganha” nas trocas de tecnologias entre firmas (Hall e Ziedonis, 2001). Parte substancial dessas patentes não é utilizada internamente nem para transações de mercado, mas usadas como patenteamento estratégico, nas chamadas *blocking and sleeping patents* (Torrise et al., 2016). Esses usos menos ortodoxos tornaram os requisitos legais para patenteamento (não obviedade, utilidade, novidade) menos relevantes e diminuíram a qualidade das patentes.

Tais estudos incorporam características das firmas e do ambiente externo no qual interagem para moldar estratégias de patenteamento. Esse tipo de estratégia é mais agressiva em mercados altamente fragmentados e propriedade dispersa das tecnologias, de forma mais pronunciada nas firmas com grandes investimentos em ativos tecnológicos específicos e sob

forte regime de apropriabilidade legal (Ziedonis, 2004).

O caso do software de código aberto (OSS) é apontado como resultado de estratégias de negócios que, em algumas situações, moldaram o regime jurídico da PI (Pisano, 2006a). Outras vezes foram fruto da regulamentação governamental e da política antitruste, que transformaram a estrutura originalmente integrada da indústria de TICs em uma estrutura modularizada.

Originalmente protegido por direitos autorais, o software passou por importante mudança no regime de apropriabilidade com o surgimento do OSS, tendo a IBM como grande patrocinadora como parte de sua reorientação estratégica desde o surgimento dos computadores pessoais e o impacto sobre seu mercado de *mainframes*, moldando a arquitetura da indústria e o próprio regime jurídico de PI (Pisano, 2006a; 2006b). As mudanças na estrutura da indústria precisam também ser entendidas como resultado da ação antitruste contra a empresa, que se estendeu do final da década de 1960 ao início dos 1980 (Rosenberg, 2010:44), promovendo a separação entre software e hardware e transformando uma estrutura antes integrada para um formato modularizado, o que permitiu a expansão de firmas especializadas em módulos específicos, como Microsoft e outras que emergiram após seu revés antitruste (Lohr, 2019).

De fato, a política antitruste orientada para assegurar o ingresso de novos entrantes (Rosenberg, 2010:48) foi decisiva na reorientação de firmas como AT&T, IBM e Microsoft, que mudaram estratégias de licenciamento, desinvestimento e desverticalização, rejuvenescendo periodicamente a estrutura da indústria de TICs ao abrir o mercado para novos *players*. Teece (2018a:1378) observa essas mudanças institucionais que transformaram a indústria eletrônica e telecomunicações originalmente altamente concentrada em uma estrutura com empresas especializadas. Cita os casos da AT&T (e o importante Bell Labs, hoje Nokia) e de empresas como Motorola, Ericsson e DEC que se especializaram em segmentos *downstream*. A modularização (ou modelo de negócios modular, nas palavras do autor) provocaria um ritmo acelerado de inovação, mas simultaneamente comprimiria os lucros e as condições de apropriabilidade em cada módulo.

As poucas exceções seriam o microprocessador e o sistema operacional, resistentes à imitação e à concorrência por expertise tecnológica, PI e, em casos como Microsoft, pelos efeitos de rede. Outra exceção ocorre com a inovação sistêmica ou arquitetônica (que envolve padrões), em sistemas complexos, a qual oferece os maiores retornos. Uma consequência da modularização são *pools* de patentes e fragmentação de DPI, como ocorre atualmente com as tecnologias digitais (Den Uijl, Bekkers e de Vries, 2013). Por conseguinte, a modularização na fase de consolidação da indústria de TICs parece resultar em uma tendência ao crescimento da importância dos DPI com o objetivo de gerenciamento estratégico (sobretudo de patentes), não

para proteger a própria tecnologia, mas sim para se defender de riscos de infrações de terceiros.

A modularização do sistema de produção de telefonia móvel limita o progresso futuro das tecnologias habilitadoras da indústria – módulos muito competitivos, onde fornecedores de tais tecnologias (incluindo proprietários de patentes essenciais padrão) não alcançam o retorno devido ao desenvolvimento de padrões. Esse é um ponto fundamental de divergência de Teece, que coloca tanto IA no mesmo patamar como tecnologias de interoperabilidade, considerando ambas como tecnologias habilitadoras. Aqui, ao contrário, a IA é considerada de forma distinta, como TPG ou habilitadora que prepara/instrumentaliza outras tecnologias e inovações de outros setores, enquanto tecnologias como *wireless*, protegidas por patentes essenciais de padrões, não interferem nas tecnologias setoriais, mas possibilitam a interoperabilidade e funcionamento conjunto das distintas tecnologias, estabelecendo interfaces entre elas. A função da IA não é fazer com que diferentes tecnologias falem e se conectem, mas (principalmente na fase de implementação downstream) criar algo novo por meio do desenvolvimento de novas tecnologias setoriais de distintas bases técnicas.

1.3. Inovação aberta, ecossistemas, plataformas digitais e apropriabilidade

A inovação aberta é um tema que tem recebido atenção considerável da literatura acadêmica recente de inovação e estratégia em função do caráter cada vez mais colaborativo requerido para a geração de muitas inovações, que dependem de conhecimentos e capacitações de outros atores, não restritas à firma responsável pela inovação focal. Sobre o conceito de inovação aberta, Chesbrough (2003:43) argumenta que “...*valuable ideas can come from inside or outside the company and can go to market from inside or outside the company as well*”.

Ao contrário do senso comum, a noção de inovação aberta no nível da firma não exclui a propriedade intelectual e uma literatura crescente tem analisado sua relação com a questão da apropriabilidade (Chesbrough 2003; Helfat e Quinn, 2006; Chesbrough e Appleyard, 2007; Bogers, 2011; Laursen e Salter 2014; Chesbrough, 2012; Holgersson et al., 2018; Chesbrough et al., 2018; Holgersson e Granstrand, 2022), complementando abordagens setoriais de estratégias de negócios, principalmente nas indústrias de alta tecnologia, como o software, tecnologias de telecomunicações sem fio e, recentemente, tecnologias da IA.

A literatura distingue entre estratégias formais e informais de inovação aberta e busca compreender sua influência na captura de valor em diferentes modelos de negócios. Engloba *stakeholders* heterogêneos - universidades, instituições de pesquisa e firmas, inclusive interdependentes ou rivais - que implementam diferentes escolhas de apropriabilidade que protegem a “parte” do conhecimento de cada organização, compartilhada em colaborações

formais (Laursen e Salter, 2014).

Diversos autores defendem que as patentes promoveriam (ao invés de dificultar) a inovação aberta ao mitigar riscos, delimitar o escopo da interação e a parcela do conhecimento de cada *player*. Para isso, a qualidade das informações das patentes torna-se ainda mais importante, uma vez que aquelas de baixa qualidade - usualmente associadas ao patenteamento crescente em smartphones e software (Graham e Vishnubhakat, 2013) - gerariam incerteza sobre o que está sendo protegido, dificultando a inovação aberta.

O conceito de inovação aberta, erroneamente associado a informações de domínio público, é cada vez mais apoiada por interesses baseados no lucro (Lianos, 2013:11), como no caso mencionado do suporte da IBM ao OSS como forma a ter uma base livre disponível para “rodar” seus programas proprietários. É comum usuários desenvolverem novos produtos por meio de troca livre dentro de suas comunidades de desenvolvedores de software, mas que podem envolver DPI quando passam ao estágio comercial.

Ao envolver múltiplos *players* com interesses heterogêneos, a inovação aberta requer estratégias de PI e apropriação, principalmente em contextos dinâmicos e sistêmicos de inovação (Holgersson, Granstrand e Bogers, 2018). Como sintetizam Laursen e Salter (2014), enquanto a criação de inovações requer cada vez mais abertura, a comercialização de inovações requer proteção.

Com isso, embora o tema da inovação aberta e colaborativa seja muitas vezes erroneamente associado a temas como o OSS, envolvem questões distintas (ainda que possam se confundir), que alguns estudos que tratam os temas conjuntamente acabam por reforçar (Kilamo et al., 2012). A inovação aberta não implica em qualquer contradição com DPI, embora a legislação não regule como esses direitos podem ser coordenados ou administrados -, de modo que a inovação aberta acaba se baseando em modelos de negócios nos quais a governança ocorre por meio de contratos de troca de informações e cláusulas de confidencialidade (Lee et al., 2013; Trott e Hartmann 2009).

No campo da IA o modelo de inovação aberta começa a ser uma tendência identificada pela literatura e promete ser crucial no estágio de sua implementação em setores a jusante, como tecnologia de uso geral/habilitadora, impondo desafios maiores de apropriabilidade (Teece, 2018a; Teece, 2018b; Gambardella et al., 2021; Conti et al., 2019; Jacobides et al., 2021; Hurmelinna-Laukkanen e Yang, 2022).

Um conjunto amplo de autores da tradição evolucionária trabalha há décadas na noção de criação “coletiva” de valor por meio de conceitos como sistemas de inovação - nacionais,

setoriais, locais e até mesmo sistemas corporativos²² -, que envolvem elementos e conexões, onde a inovação resulta da interação da firma com outras organizações (sobretudo, acadêmicas e de pesquisa, mas também firmas entre si), a partir de instituições (normas, formais e informais) que influenciam a *interação* ao definir as regras do jogo. Contudo, mesmo essas abordagens mais realistas do processo de inovação foram incapazes de prover o instrumental para análise das novas formas de organização e governança da firma no mundo globalizado, que só surgiu com a literatura de GVCs (sigla de *global value chain*) e plataformas digitais.

Um conceito central desta pesquisa é o de ecossistemas de inovação²³, que engloba uma literatura ampla, multifacetada e dispersa, desenvolvida nas últimas décadas, que procura enfatizar a interação, a partir da noção de complementaridades, mas com interrelações e interdependências em constante evolução.

É um conceito que remonta a Moore (1996) e a noção de que as empresas são parte de um sistema de interações evolutivas, onde a captura de valor depende de como interagem e coevoluem com outros *players* (fornecedores, clientes e parceiros) e da sua adaptação às mudanças, colaboração estratégica e liderança na evolução do ecossistema. O tema ganhou interesse recente nos campos da competição e estratégia, associado aos tópicos de plataformas digitais ou inovação aberta que amiúde surgem entrelaçados na literatura, ainda que não sejam conceitos similares, mas interrelacionados diante das tecnologias complexas emergentes da economia digital.

Correspondem a formas distintas de organizar atividades econômicas ligadas por tipos específicos de complementaridades e interdependências entre organizações independentes que participam trazendo habilidades e conhecimentos específicos, combinando colaboração e competição (Jacobides, Cennamo e Gawer, 2018; Adner, 2017; Teece, 2018a, este último a partir da noção de capacidades dinâmicas, que correspondem às habilidades da empresa para adaptar, integrar e reconfigurar recursos internos e externos em ambientes de rápida mudança, como nos ecossistemas de inovação). O conhecimento flui entre os diferentes *stakeholders* do ecossistema por meio de interações formais (como parcerias e participações acionárias) e informais (como redes).

Os estudos enfatizam diferentes aspectos do ecossistema, dependendo da unidade de análise escolhida, tal como a empresa e seu ambiente no caso dos ecossistemas de negócios; ou

²² O conceito de sistemas corporativos de inovação (Granstrand, 2000), que teria ficado um pouco esquecido na literatura, foi recentemente recepcionado por Rikap e Lundvall (2021) na análise da IA. Ainda que não prescindida da ideia da criação ecossistêmica de valor, é relevante quando se procura não perder de vista a unidade de análise da empresa.

²³ Para a revisão da literatura e origens de ecossistemas de inovação ver Granstrand e Holgersson (2020).

inovações específicas e seus vários participantes; ou organização dos distintos *stakeholders* em torno dos ecossistemas de plataforma (Jacobides, Cennamo e Gawer, 2018).

Ecossistemas de inovação referem-se a redes amplas e sistêmicas de interação entre diferentes *players* que colaboram no desenvolvimento de novas tecnologias e padrões e competem no mercado de forma a implementar a inovação e capturar valor. A inovação emerge como um processo dinâmico que surge da interação e interdependência entre diferentes *players*, tecnologias e contextos, em lugar dos limites isolados da empresa.

Plataformas digitais podem, em muitos casos, criar ou sustentar ecossistemas de inovação, mas não são sinônimos. As plataformas, que podem ter diferentes formatos (transações, infraestrutura e inovação ou híbridas) (Gawer, 2014; Cusumano et al., 2019; Parker et al., 2016; Plantin et al., 2016), enquanto um modelo de negócios, podem servir como infraestrutura ou conexão dentro de um ecossistema maior. Jacobides, Cennamo e Gawer (2018) identificaram condições necessárias para o surgimento dos ecossistemas, como a modularidade e diferentes *tipos* de complementaridades, que levam empresas a influenciar e moldar a formação de ecossistemas para lidar com a dinâmica distinta de criação e captura de valor. A ideia de "ecossistema de plataforma" é o ponto de ligação central, que considera como os *stakeholders* se organizam em torno de uma plataforma. Do ponto de vista formal, os ecossistemas não são gerenciados hierarquicamente, mas pelo gerenciamento de padrões e interfaces, direitos de propriedade intelectual e outras formas contratuais, incluindo a importância das interfaces tecnológicas para controle, *lock-in* e outros.

1.4. As tecnologias de interoperabilidade

Uma questão importante relacionada as TICs e eletrônica é demandarem cada vez mais padrões técnicos de interoperabilidade²⁴ e um tipo diferenciado de tecnologias que refletem esses padrões. Computadores, smartphones e tablets cada vez mais interagem entre si e se conectam à Internet por meio de tecnologias padronizadas. Padrões de interoperabilidade tornaram-se essenciais com a proliferação de CGVs, possibilitando que produtos e componentes fabricados por distintos fornecedores ao redor do mundo operem conjuntamente. Nos mercados de tecnologia digital, padrões tornaram-se essenciais e amplamente globais,

²⁴ O conceito de interoperabilidade, que parece ter um apelo intuitivo, refere-se, essencialmente, à capacidade de os sistemas controlados por diferentes operadores serem compatíveis e partilhar informação em conjunto. Várias indústrias exigem padrões similares, tais como redes ferroviárias transfronteiriças, baseadas na interoperabilidade dos trens devido aos padrões de bitola ferroviária comum. Sistemas telefônicos que operam com base em um padrão devem interoperar com outros padrões para permitir chamadas telefônicas entre redes distintas em regiões diferentes. Arquivos interoperáveis permitem que documentos feitos em um programa de computador sejam abertos e editados em outros sem a perda de dados.

como protocolos e tecnologias de rede sem fio (Wi-Fi e Bluetooth), telecomunicações sem fio (as 5 gerações existentes até o momento e a 6G em fase de desenvolvimento).

De acordo com a International Standard Organization (ISO)²⁵, padrão é um documento, estabelecido por consenso e aprovado por órgão reconhecido, que fornece, para uso comum e repetido, regras, diretrizes ou características para as atividades. Grandes empresas globais participam ativamente de várias SSOs com atuação internacional (Contreras, 2017).

Padrões são importantes porque ao estabelecer protocolos consistentes podem ser universalmente compreendidos e adotados, ajudando na compatibilidade de bens e serviços e estimulando a implementação de tecnologias em setores a jusante. São voluntários, mas sua aplicação oferece uma série de vantagens, como melhoria da compatibilidade, qualidade e níveis mínimos de desempenho dos produtos e serviços no mercado. Em particular, permitem que sistemas mecânicos, elétricos, de informação, telecomunicações e outros interoperem (Wong-Ervin e Wright, 2016).

São aprovados/definidos por organizações especiais (SSOs, sigla da expressão em inglês *standard-setting organizations*)²⁶ a partir de discussões técnicas com as partes interessadas, incorporando as melhores tecnologias disponíveis à época, cobertas por PI, as chamadas patentes essenciais de padrões (SEP, acrônimo de *standard essential patents*), que protegem invenções imprescindíveis para produtos e serviços interoperarem. De outro modo, cada fabricante teria que desenvolver por si mesmo as informações técnicas que viabilizassem a operação conjunta com outros módulos²⁷. Os benefícios da interoperabilidade são ampliados em mercados com efeitos de rede, onde o valor do produto/serviço para o consumidor individual depende do número total de consumidores que adotam produtos compatíveis.

O tema da PI relacionada a padrões avançou desde a década de 1980, ganhando relevância nas TICs, em função da inovação modularizada e a necessidade de interoperabilidade

²⁵International Standard Organization (ISO). Disponível em <https://www.iso.org/home.html>. Acesso em 10 out. 2022

²⁶ O papel das SSOs é coordenar a definição de padrões e, em alguns casos, estabelecer regras de divulgação dos termos de licenciamento, sem neles intervir. Há várias SSOs ativas, algumas grandes organizações com foco mundial - como ISO e International Telecommunication Union (ITU) – ou associações profissionais, como o Institute of Electrical and Electronics Engineers (IEEE), que desenvolveu padrões da Ethernet, Wi-Fi, Firewire e Bluetooth. A Internet Engineering Task Force (IETF), com suporte da Internet Society, surgiu nos primeiros esforços de padronização da Internet com os protocolos TCP/IP. O consórcio Organization for the Advancement of Structured Information Standards (OASIS), concentrada no desenvolvimento, convergência e adoção de padrões da Web. O World Wide Web Consortium (W3C), fundado pelo inventor da Web, desenvolve padrões de conexão com a cultura de licença gratuita para infraestrutura básica global. O European Telecommunications Standards Institute (ETSI), tem atuação regional e constituiu o consórcio de desenvolvimento colaborativo de tecnologias 3G, como GSM, WCDMA e LTE.

²⁷ Patentes são o instrumento mais comum para implementação de um padrão, mas não o único, podendo também contemplar direitos autorais, modelos de utilidade, banco de dados, know-how e até *trade secrets*.

e efeitos de rede (Bekkers e Updegrove, 2012). Regimes de PI buscam o equilíbrio entre, de um lado, a proteção da propriedade individual sobre ativos intangíveis e, de outro, o interesse público e o bem-estar da sociedade. Padrões técnicos, contudo, requerem um olhar distinto sobre a PI, mais orientado para seu uso prático. Isso porque enquanto patentes tradicionais são baseadas na exclusão, patentes de padrões são instrumentos de difusão das tecnologias, requisito chave em TICs e outras indústrias modulares, pois o sucesso do padrão e o valor das patentes subjacentes dependem da ampla disseminação. Os detentores de SEP obviamente desfrutam de grande vantagem econômica, mas não podem excluir terceiros (inclusive rivais) (Baron e Spulber, 2018). Com isso, o tema perpassa questões que abarcam DPI e seu entrelaçamento com o direito contratual, transferência de tecnologia e defesa da concorrência, sendo um campo complexo e marcado por uma infinidade de casos de litígios.

Quando o padrão é definido, parcela substancial da inovação relevante já foi patenteada e é improvável que requisitos técnicos possam ser atendidos sem as SEP. Na ausência de DPI, a adoção de um padrão uniforme criaria incentivos a “caronas”, eliminando o incentivo para desenvolver/aprimorar padrões (Sidak, 2013).

Algumas tecnologias são escolhidas em detrimento de outras para constar do padrão, quando todos os investimentos *downstream* passam a ser feitos segundo o padrão definido. A mudança/migração posterior para uma tecnologia alternativa é difícil, senão impossível, para terceiros (como implementadores de aplicativos, por exemplo), pelos altos custos de transação (Williamson, 1979), efeitos de rede e a própria funcionalidade técnica (Lemley e Simcoe, 2019).

Isso assegura um potencial poder de mercado aos titulares das patentes relacionadas ao padrão, seja pela imprescindibilidade de utilizar as tecnologias, seja pelo risco de aprisionamento (*lock-in*), com as SEP devendo seguir condições especiais. Esse poder de mercado potencial de seus titulares leva os SSOs a estabelecerem condições especiais de licenciamento, de modo a evitar desvios (Lemley, 2002). Os titulares dessas patentes se comprometem a licenciá-las (voluntariamente) em condições especiais chamadas FRAND (*fair, reasonable and non-discriminatory*), ou RAND, como é chamado nos EUA, em alguns casos sem custos, as chamadas condições *royalty-free* (RF) ou RAND-Z (RAND com *zero royalty*), como no caso dos padrões abertos da Web que foram fundamentais para a acessibilidade.

Um padrão envolve, na prática, um conjunto enorme de patentes que reivindicam invenções que devem ser usadas para seu atendimento (Shapiro, 2001). As dimensões importantes relacionam-se ao conceito de essencialidade, ao escopo de divulgação e ao compromisso de licenciamento e suas condições. A maioria das SSOs não faz qualquer revisão do conteúdo, baseando-se na declaração de essencialidade do titular e negociada com os

membros da SSO, cuja confirmação depende dos participantes e do mercado. A consolidação do padrão passa pela confiança de que a tecnologia é robusta, testada, comercialmente viável e capaz de oferecer suporte para aplicativos. Algumas SEP declaradas essenciais no momento de definição do padrão (que inclui apenas um subconjunto estimado entre 20% e 30% das autodeclaradas essenciais) podem deixar de ser essenciais em data posterior, com seus titulares passando a ser livres para decidir se estão dispostos a licenciar e em que termos.

Além disso, herdamos padrões anteriores. Muitos novos padrões de telecomunicações usam tecnologias de Internet (TCP/IP) ou protocolos de codificação de imagem, áudio e vídeo existentes, protegidos por SEP existentes. Em tais casos, a implementação compatível do padrão em questão requer que os padrões anteriores também devam ser empregados, levantando ao frequente questionamento se esses direitos prévios devem ser considerados essenciais no contexto do novo padrão (Bekkers e Updegrave, 2012).

No entanto, à medida que aumenta a complexidade dos sistemas em rede, a exigência de interoperabilidade torna-se tema controvertido, com seus benefícios questionados, da mesma forma que a não portabilidade de dados, por constituírem barreiras à inovação e riscos à competição, com tensão de interesses e direitos legais sobrepostos (Kraemer, 2020; Graef, 2021).

Na prática, as SEP podem ser mal utilizadas, barrando concorrentes, promovendo discriminação de preços e condições de licenciamento, possibilitando a acumulação de *royalties* e atrasos no licenciamento. Com isso, são fonte de intensa judicialização e escrutínio antitruste, com disputas legais dispendiosas, demoradas e resultados imprevisíveis (Erixon e Bauer, 2017). A frequência em que ocorrem supera o licenciamento usual de patentes, com probabilidade de judicialização cinco vezes maior que patentes não-SEP (Contreras, 2016), envolvendo aspectos relacionados ao valor dos *royalties* e violação da lei antitruste (Lemley e Shapiro, 2007; 2013).

Os compromissos FRAND são projetados com o duplo objetivo de garantir que a tecnologia incorporada no padrão seja acessível a todos e recompensar financeiramente os titulares das SEP (Competition Policy Brief, 2014). A lógica é que o benefício de mercado das tecnologias patenteadas que constituem padrões técnicos seja contrabalançado por garantias que impeçam seus titulares de abusos da posição dominante no mercado pela adoção generalizada do padrão e impeçam práticas anticompetitivas e comportamentos oportunistas. Na prática, contudo, o mecanismo usualmente pode não ter a eficácia pretendida, pois o entendimento dos termos FRAND/RAND não é claro, não sendo precisamente definido pelas SSOs, mas negociados entre as partes em contratos de licenciamento (India Discussion Paper, 2016). Se por um lado isso poderia assegurar flexibilidade nas negociações bilaterais de

licenciamento, por outro, tem levado a disputas da interpretação dos termos de licenciamento e litígios envolvendo os termos FRAND. Assim, essa interpretação acaba surgindo por meio da evolução da jurisprudência nas decisões dos tribunais e órgãos reguladores da concorrência. Como sintetizado por Swanson e Baumol (2005), o compromisso FRAND se traduz em eficácia limitada e um contrato incompleto entre licenciadores e licenciados devido à ambiguidade do seu significado, com casos frequentes de *hold-up* (*royalties* acima da taxa justa, devido ao *lock-in*, ou aprisionamento tecnológico) e *hold-out* (*royalties* abaixo da taxa justa) identificados pela literatura (Lemley e Shapiro, 2007; Contreras, 2016; Langus et al., 2013).

A complexidade do licenciamento é ampliada porque muitos padrões englobam várias SEP, como WiFi e 3G que envolveram centenas ou milhares de patentes²⁸ e risco de "empilhamento de *royalties*" (*royalty stacking*) se cada titular exigir uma parcela desproporcional da receita do produto, ao invés de manter o conceito royalty FRAND baseado no valor conjunto das patentes relevantes e da contribuição incremental de cada titular. São comuns *pools* de SEP (várias patentes de muitos titulares empacotadas e licenciadas por uma terceira entidade ou por um dos titulares), principalmente nos padrões de tecnologias digitais, permitindo seu agrupamento em um contrato de licença padrão global, com repartição da taxa de licenciamento entre os membros do *pool*. Outra situação frequente é o licenciamento cruzado, com compartilhamento mútuo de patentes entre titulares, concedendo a cada um o direito de usar a patente do outro. Os *pools* de patentes e as licenças cruzadas envolvem muitos problemas de transparência dos royalties.

Nas últimas décadas, houve um aumento exponencial das SEP, em especial de telecomunicações sem fio, a maioria de propriedade de um grupo seleto de firmas, como Qualcomm, InterDigital, LG Electronics, Nokia, Samsung, Ericsson e Motorola, evidenciando grande concentração em firmas de economias desenvolvidas, incorporando, recentemente, firmas chinesas como Huawei e ZTE, bem posicionadas no ranking das patentes autodeclaradas essenciais dessa tecnologia²⁹.

Enquanto algumas empresas possuem SEP como resultado da intensa P&D e inovação, outras formaram carteiras importantes a partir da compra de patentes. Em dispositivos móveis tiveram destaque a compra do portfólio da Nortel pelo consórcio Apple, Microsoft e RIM, em

²⁸ Produtos e serviços relacionados são redes emaranhadas de SEPs. Até 250.000 patentes podem ser requeridas para uma especificação técnica ou elemento de design em um único smartphone.

²⁹ Teece (2021) discute detalhadamente o uso dessas patentes como *proxy* da liderança das empresas nas inovações 5G, sobretudo chinesas, com base em críticas usuais do uso de patentes como indicador de capacitação tecnológica.

2010, por US\$ 4,5 bilhões³⁰ e, em 2011, da Motorola pelo Google, por US\$ 12,5 bilhões, incluindo suas muitas SEP, evidenciando sua importância para as plataformas digitais³¹.

1.5. A crescente importância dos dados na economia digital

Algumas observações relevantes para a discussão sobre DPI de IA relacionam-se aos dados. Embora pareça sem sentido falar sobre propriedade de dados, individuais e brutos, há controle, acesso e direitos sobre dados agregados. Atualmente, as grandes firmas de plataformas são as principais responsáveis pela sua coleta, análise e armazenamento e quem se apropria do seu valor, enquanto os reais geradores dos dados (principalmente os pessoais) não têm qualquer controle sobre seu uso, pois uma vez extraídos, não capturam qualquer parcela do valor econômico ou o capturam por meio de acesso a serviços “gratuitos” das plataformas digitais graças a expressivas receitas de propaganda online. Assim, são as firmas orquestradoras dessas plataformas que têm controle e os direitos sobre grandes conjuntos de dados.

Em termos agregados, dados fazem jus a DPI, embora sua dimensão como recurso econômico seja complexa. Savona (2019:5) lembra que “*The economic nature of goods in terms of excludability and rivalry is determined both by law, which regulates the contractual rights over their ownership and use, and by the scale of their supply*”.

A natureza econômica dos dados, segundo o *mainstream* econômico, está baseada na mesma noção de bens públicos, não excludentes e não rivais, que respalda os DPI, que podem ser usados simultaneamente pelas firmas (sem esgotá-los), inclusive sem respeito a privacidade dos usuários acumulados de maneira socialmente ineficiente (Jones e Tonetti, 2020:1).

Dados individuais gerados como informação (diferente de conhecimento) são não rivais, mas excludentes, podendo ser considerados bens de clube³². Contudo, uma vez agregados e tratados para gerar valor, sua natureza muda de bens do clube para bens públicos ou privados (Savona, 2019:8-9).

O termo BigData foi popularizado para denotar a ampla gama de dados cada vez mais gerados na sociedade (UNCTAD, 2019:28). A BigData tem alta possibilidade de exclusão, mas rivalidade muito baixa, pela sua elevada escala, o que torna seu custo marginal nulo. A combinação de altos custos fixos da infraestrutura digital para coleta e tratamento dos dados, com custo marginal zero dos dados individuais, torna as plataformas digitais monopólios quase naturais e cria assimetrias no valor dos dados para os membros do clube. A análise dos dados

³⁰ Incluiu cerca de 6.000 patentes e pedidos em *wireless*, *wireless* 4G, rede de dados, óptica, voz, Internet, provedor de serviços, semicondutores e outras relacionadas aos padrões 3G e 4G (FULCRUM, 2011).

³¹ Se o Google tivesse adquirido as patentes da Motorola antes do início do litígio relacionado a smartphones com a Apple, a ameaça de uma retaliação por deter patentes da Motorola poderia ter evitado os processos.

³² Ver Nota de rodapé 19.

por BigData, os converte em conhecimento, podendo ser considerado um bem público (não rival e não excludente) quando usado sem fins lucrativos e administrado por atores públicos, mas um bem privado, excludente e privado, se apropriado por atores privados com fins lucrativos.

Os dados, na forma de BigData, certamente não são bens públicos. Dados que alimentam algoritmos de redes neurais profundas são mantidos em sigilo, com cada firma coletando, agregando, tratando e armazenando suas próprias bases. A importância dos conjuntos de dados estaria mudando as fontes de vantagens competitivas da IA transferindo a busca da exclusividade na invenção para a exclusividade nos dados (Hartmann e Henkel, 2020). Grandes conjuntos de dados que treinam os atuais sistemas de IA recebem amplos e sobrepostos DPI, além de serem monitorados e tornados exclusivos por meio do uso de restrições técnicas impostas pelos detentores de bancos de dados digitais, controlando usuários e bloqueando o acesso de rivais (Alvarenga, 2019; Jones e Tonetti, 2020:4).

1.6. Regime de apropriabilidade: além da propriedade intelectual

Teece (1986) definiu uma abordagem inteiramente distinta para tratar a questão da apropriabilidade da inovação. Associou as literaturas de inovação e estratégia da firma, isolando, para fins explicativos, os processos de captura e criação de valor da inovação, com vistas a compreender por que inovadores pioneiros raramente são capazes de transformar inovação em lucros e vantagens econômicas de longo prazo (Pisano e Teece, 2007).

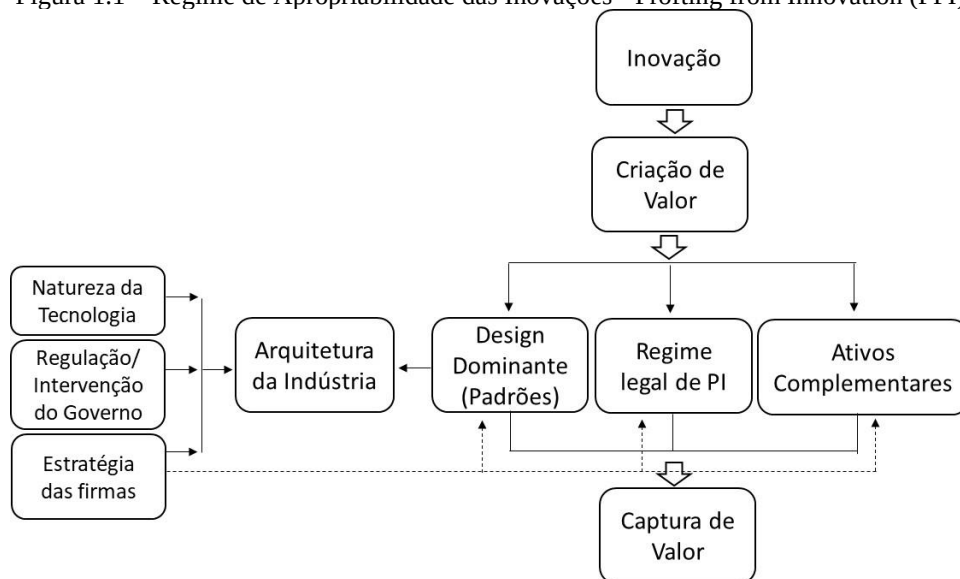
Pela abordagem do *Profiting from Innovation* (PFI), a captura do valor da inovação depende não somente do regime legal de PI no qual a firma opera, que pode ser forte ou fraco de acordo com a natureza da tecnologia e da eficácia do sistema jurídico para proteger a PI (Teece, 1986:287), mas da sua capacidade de definir um *design* tecnológico dominante e da gestão estratégica de ativos complementares exigidos para levar e manter a inovação no mercado. Em especial, em regimes fracos de patentes, as firmas precisam definir estratégias de negócios e construir capacidades dinâmicas de forma a obter retorno com a inovação. A Figura 1.1 resume e revisita a abordagem do regime de apropriabilidade do PFI, identificando as variáveis chaves para a captura de valor da inovação (*design* dominante, regime de PI e ativos complementares), além da arquitetura da indústria (como o caso da modularização da indústria de TICs, influenciada pelas estratégias empresariais e pela política pública). As linhas pontilhadas correspondem as variáveis endógenas, fruto de decisões da firma, com vistas a capturar o valor da inovação.

O avanço da ciência evolui de um estágio pré-paradigmático, onde os conceitos ainda

não estão consolidados, até chegar ao estágio paradigmático, quando os conceitos passam a ter aceitabilidade universal em função da maturidade de padrões de pesquisa científica, culminando na definição de um **design dominante**. Segundo Teece (1986:288), “*At some point in time, and after considerable trial and error in the marketplace, one design or a narrow class of designs begins to emerge as the more promising.*” Este *design* não é necessariamente o “melhor”, mas aquele que consegue se manter e suplantar outras tecnologias no mercado por efeitos de rede no início das escolhas de tecnologias (Teece, 2006:1134).

Teece (1986) dedicou pouca atenção a explicitar claramente noções relacionadas ao *design* tecnológico dominante, particularmente na indústria de TICs. O próprio autor reconhece em artigo posterior essa falha da literatura do PFI: “*Although PFI highlighted the role of dominant designs, it did not fully develop the role of standards, and standard setting bodies, on the innovation process. Nor did it explore the interaction between standards and intellectual property protection*” (Teece 2006:1134). De fato, um tratamento menos superficial do tema dos padrões, essenciais nas TICs e mais ainda nas TPGs/THs, só ocorre quando o tema foi tratado no âmbito da economia digital (Teece 2016; 2018a), ainda que quase exclusivamente limitado às tecnologias de telecomunicações sem fio e Internet – cada vez mais fundamentais para a operação conjunta da IA e outras tecnologias, inclusive digitais, no cenário das plataformas e seus ecossistemas -, mas mesmo assim sem qualquer referências ao *design* dominante das técnicas da IA propriamente dita.

Figura 1.1 – Regime de Apropriabilidade das Inovações - Profiting from Innovation (PFI)



Fonte: Elaboração própria baseado em Teece (1986, 2006, 2012, 2018a); Pisano (2006a; 2006b); Pisano e Teece (2007)

Um *design* tecnológico que se torna dominante constitui um dos elementos que definem as condições de apropriabilidade. O padrão é uma atividade arriscada, mas pode render retornos

significativos, de forma que a competição pelo padrão (ou *design* dominante) torna-se, de certa forma, a competição pelo mercado (Teece, 2006:1141). No entanto, as vantagens do *design* pioneiro não começam até que o mercado e/ou órgão de definição de padrões o convertam no *design* padrão.

Esse divisor de águas é importante para o padrão de competição e a distribuição de lucros entre o inovador e seguidores, pois quando a imitação é possível por meio de *designs* alternativos durante a fase pré-paradigmática os seguidores terão oportunidade de converter sua tecnologia no padrão da indústria e inclusive promovendo-a junto aos órgãos definidores.

Na fase pré-paradigmática, a competição recai justamente na disputa pelas firmas pela posição de tecnologia dominante frente a tecnologias alternativas, enquanto no estágio paradigmático, que se inicia quando uma das tecnologias se torna o *design* dominante e o padrão praticado pelo mercado, a competição passa a depender de outras variáveis, em especial estratégias empresariais sobre ativos complementares.

O regime de apropriabilidade, formulado originalmente para tecnologias discretas, principalmente de produto, típicas da economia industrial, decorre da natureza da tecnologia, que muitas vezes, por si só, impõe barreiras naturais à imitação, e da eficácia dos mecanismos legais de proteção à PI. Na abordagem do PFI, o sistema legal é analisado em dois casos extremos: regimes "rígidos" (tecnologias de fácil proteção) ou regimes "fracos" (tecnologias com características que tornam a proteção quase impossível). Nesses últimos, a captura de valor da inovação passa a depender fortemente de um conjunto de ativos, tecnologias e capacidades complementares.

Teece (1986) dedica pouca atenção a explicitar claramente essas noções, partindo da suposição de que para uma ampla gama de tecnologias e indústrias, o regime de proteção por patentes é fraco (Teece, 1986:290), para a partir daí discutir de forma mais aprofundada o conceito de ativos complementares.

No entanto, as implicações econômicas são muito distintas caso um **regime legal de PI seja forte ou fraco**. Nos regimes fracos há probabilidade significativa de patentes serem anuladas ou contornadas com relativa facilidade (Anton et al., 2006), levando as firmas a dependerem de ativos complementares, como proposto pelo PFI, ou de outros mecanismos de proteção, como sigilo, consequentemente reduzindo o conhecimento divulgado publicamente, em um ambiente no qual a firma carece de informações economicamente relevantes sobre as capacidades de seus concorrentes.

Teece e Sherry (2017:20) argumentam que uma proporção significativa de patentes concedidas pode ser invalidada ou reduzida, uma vez que direitos de patente, ao contrário da

propriedade sobre bens tangíveis, não podem ser retidos fisicamente, não são autoaplicáveis e não podem ser automaticamente resguardados após a patente ser publicada e a invenção divulgada. O titular precisará recorrer aos tribunais para impedir infração de direitos e buscar compensação pelo uso não autorizado, em processos caros, demorados e incertos. Estudos empíricos mostram que somente metade das patentes litigadas serão consideradas válidas. Allison e Lemley (1998) constataram que 46% das patentes contestadas nos EUA entre 1989 e 1996 foram anuladas (antes da criação do Circuito Federal, em 1982, era 65%). Segundo Mansfield et al. (1981), 60% de um conjunto de invenções (70% delas protegidas por patentes) foram imitadas em até 4 anos, sugerindo regimes não fortes para impedir a imitação.

Lemley e Shapiro (2005) argumentam que direitos de patentes são apenas probabilísticas³³. No entanto, constataram que apenas 1,5% das patentes são litigadas simplesmente porque a maior parte das patentes concedidas acaba tendo pouco ou nenhum significado comercial, não levadas ao mercado.

Cada vez mais é reconhecido que uma patente não confere ao seu titular o direito de excluir, mas sim o direito de tentar excluir por meio de ações nos tribunais, com disputas crescentes à medida que aumenta o número de patentes (Shapiro, 2003:391). Teece e Sherry (2017), refutam a expressão “monopólio”, que deve ser usada com cuidado, lembrando que enquanto nos livros texto de economia o monopólio é definido pela existência de um único fornecedor, na política antitruste diz respeito a um mercado relevante, que abrange bens idênticos e também intercambiáveis ou substituíveis pelo consumidor, devido a suas características, usos ou preços. No caso de tecnologias, algumas delas, que são patenteadas, terão como substitutas, próximas ou distantes, patenteadas ou não, que podem competir no mercado de tecnologia relevante. Em síntese, Teece e Sherry (2017) enfatizam que antes de ter poder de monopólio, o titular da patente pode ser vítima de infração, só lhe restando processar judicialmente.

Direitos de patente fracos aumentam a probabilidade de litígio. Contudo, Anton et al. (2006:3-4) observaram que algumas patentes originalmente percebidas como fracas, poderão ser futuramente vistas como fortes, em campos que envolvam uma nova classe de tecnologia, onde pode existir incerteza sobre a proteção e escopo da patente, como nos organismos vivos geneticamente modificados e nos métodos de negócios. A força ou fraqueza do regime de PI depende da natureza da tecnologia, mas também da adequabilidade do regime jurídico, da

³³ Autores como Gant (2021) contestou a teoria probabilística da patente, pois ainda que o litígio de patentes possa ser probabilístico, a propriedade em disputa não é.

qualidade das patentes e do exame em si, bem como da convergência dos requisitos estabelecidos na lei à realidade de tecnologias complexas.

A contribuição de Pisano (2006a) ao PFI mostrou a endogeneidade do regime legal de PI, moldado a partir de estratégias das firmas, que também influenciaram a arquitetura da indústria de TICs de uma arquitetura vertical para um modelo horizontal (Pisano 2006b). A especialização de firmas em módulos particulares requer interfaces bem definidas entre as tecnologias com alto grau de modularidade³⁴, o que permite que componentes interdependentes de um sistema sejam fabricados por diferentes produtores, com coordenação limitada – limitação esta suplantada com os ecossistemas de plataformas digitais e tecnologias de interoperabilidade.

Ao admitir a hipótese da fraqueza do regime de PI na abordagem do PFI, o êxito da inovação, na fase paradigmática, passa a depender principalmente de **ativos complementares**. Nas inovações da economia industrial, o PFI mostrou que para ser bem-sucedida no mercado uma inovação requer outras atividades e capacidades complementares. Ativos complementares foram essenciais para inúmeras inovações retratadas na literatura, controlados por fornecedores especializados, o que explica por que inovadores pioneiros nem sempre foram capazes de colher seus frutos duradouros. A noção de ativos complementares agregou a competição real ou potencial, indireta, no mercado como dimensão relevante para a apropriabilidade e o sucesso da inovação.

Ativos complementares podem ser de três categorias: genéricos, especializados ou co-especializados. Os genéricos são ativos de uso geral, que não demandam qualquer tipo de adaptação à inovação focal. Os ativos especializados são aqueles em que existe dependência unilateral entre a inovação e o ativo complementar, como, por exemplo, instalações fabris e equipamentos necessários para fabricação da inovação focal. Finalmente, os co-especializados são aqueles onde existe uma dependência bilateral entre a tecnologia focal e o ativo, sugerindo que teriam um perfil de tecnologias complementares.

O fornecimento de ativos complementares passa a ser, na fase paradigmática e em regimes fracos de PI, a dimensão crucial para a apropriabilidade e o lócus central da competição. Podem ser desenvolvidos pelo próprio inovador focal (por meio da construção de capacidades dinâmicas) ou obtidos de outras empresas (por meio de aquisição, acordos de co-produção,

³⁴ Baldwin e Clark (2000) analisaram o impacto da modularidade nos ecossistemas de inovação, mostrando como sistemas modulares permitem que diferentes empresas contribuam para o desenvolvimento de produtos complexos, capturando valor por meio do controle de módulos críticos de sistemas maiores, componentes que são centrais e dos quais dependem os *players* de outras partes do ecossistema.

acordos de distribuição cruzada, licenciamento de tecnologia ou mesmo parcerias, como em *joint-ventures*). Para dispor de ativos complementares exigidos pela inovação, a firma inovadora precisa adquirí-los de terceiros ou desenvolver capacidades dinâmicas para produzi-los internamente (Teece, Pisano e Shuen, 1997; Eisenhardt e Martin, 2000; Pisano, 2017). Assim, quanto mais fraco o regime de proteção legal, maior será o incentivo para a firma se integrar em ativos complementares.

A decisão da firma inovadora entre integrar atividades ou acessar ativos de terceiros é estratégica e dinâmica, dependendo de contextos específicos do mercado, que mudam no tempo e envolvem formas distintas de interação com outros *players* (atividades interfirmas, para usar a expressão de Teece, 1986:285), sejam eles inovadores seguidores, fornecedores ou distribuidores, criando interdependências e complementaridades, os quais podem se converter em futuros concorrentes. É nisso que reside o papel crucial dos ativos complementares, materialização da relação de complementariedade e/ou competição entre firmas, envolvendo decisões penrosianas típicas. Adquirir ativos complementares de terceiros não é algo simples e envolve questões cruciais para o sucesso da inovação focal no longo prazo pelo risco de surgimento de gargalos e competição potencial. Isso seria particularmente desafiador em regimes de apropriabilidade fracos, onde nem a complexidade intrínseca da tecnologia dificulta sua imitação, nem há mecanismos de PI suficientes para sua proteção, o que deixa o inovador focal em posição desvantajosa frente aos detentores de ativos complementares, incentivando-o à integração.

Esses ativos, no caso de inovações discretas - conceito não bem definido na versão original, mas explicitado em uma das revisões do PFI, como inovações destinadas a solução de um problema específico (Teece, 2018a:1371) - originalmente abarcavam serviços de marketing, capacidade de produção competitiva e suporte pós-venda, como canais de distribuição e assistência técnica, bem como tecnologias complementares (Figura 1.2). Este seria o caso quando os ativos complementares são parte da mesma inovação sistêmica, como o software especializado, sistemas operacionais e aplicativos exigidos pelo hardware de computador (Teece, 1986:288). De acordo com Teece (2018a:1379), a estrutura PFI baseava-se implicitamente em ideias da teoria da barganha, que geralmente mostrava como as rendas seriam repartidas³⁵, que parece bastante útil na análise das plataformas digitais.

³⁵ O princípio da barganha, central do direito contratual, se baseia, na ideia de que a “barganha” é um fenômeno que envolve duas partes, que podem cooperar para a criação de um excedente comumente desejável, cuja distribuição as partes estão em conflito.

Figura 1.2 - Ativos complementares necessários para comercializar uma inovação



Fonte: Adaptado da Fig.5 de Teece (1986:289)

A abordagem do PFI vem sendo revisitada com vistas a identificar as condições de apropriabilidade na economia digital (Teece, 2012; 2016; 2018a; 2018b; 2020; Petit, Teece, 2020), que envolvem muito menos tecnologias discretas e mais tecnologias sistêmicas, transversais³⁶, tecnologias de propósito geral ou habilitadoras. Esses trabalhos contribuíram para um melhor entendimento das especificidades da criação e captura de valor em ecossistemas de plataformas digitais, considerando a importância de distintos tipos de ativos (ou tecnologias) complementares (verticais e laterais) no contexto digital, como o papel dos dados (e seu controle), que não constituem meros mecanismos de captura de valor, mas elementos necessários para o funcionamento da própria tecnologia. No entanto, a criação de valor das tecnologias da IA se dá por meio de um amplo ecossistema e dinâmico, não pela ação da firma isolada, mas por uma crescente complementaridade entre diferentes *players* em um regime de apropriabilidade que contempla as variáveis padrão do PFI com uma natureza um pouco diferenciada.

³⁶ Sobre o conceito de tecnologia transversal, ver Technologies.org, disponível em: [https://technologies.org/ what-is-transversal-technologies/](https://technologies.org/what-is-transversal-technologies/). Acesso em 20 set. 2023

CAPÍTULO 2 – ECOSSISTEMAS DE INOVAÇÃO E PLATAFORMAS DIGITAIS: A CONSOLIDAÇÃO DA INTELIGÊNCIA ARTIFICIAL PREDITIVA

2.1. Antecedentes: a longa trajetória da IA

A IA é uma tecnologia de fronteira que pareceu ter chegado muito cedo a este mundo (Shrobe, 1996) e precisou esperar por inovações complementares - expansão em poder computacional e BigData -, para habilitar altos níveis de precisão dos modelos de IA e promover a consolidação de um determinado paradigma de pesquisa. Sua trajetória foi longa e apenas a partir da segunda década do século XXI ocorre a efetiva transformação de invenções notáveis em inovações exitosas no contexto de modelos de negócios de plataformas digitais.

As primeiras referências a uma IA surgiram na época de ouro da ficção científica com livros como *Runaround*, de Isaac Asimov (1942), e nos trabalhos científicos precursores de McCulloch e Pitts (1943). Uma das primeiras inovações parece ter sido o desenvolvimento de um dispositivo eletromecânico, por Alan Turing, pioneiro da área, que deu origem ao “The Bombe”, artefato usado pelo governo britânico na Segunda Grande Guerra para quebrar o código Enigma do exército alemão (Norvig e Russell, 2010; Nardin, 2021).

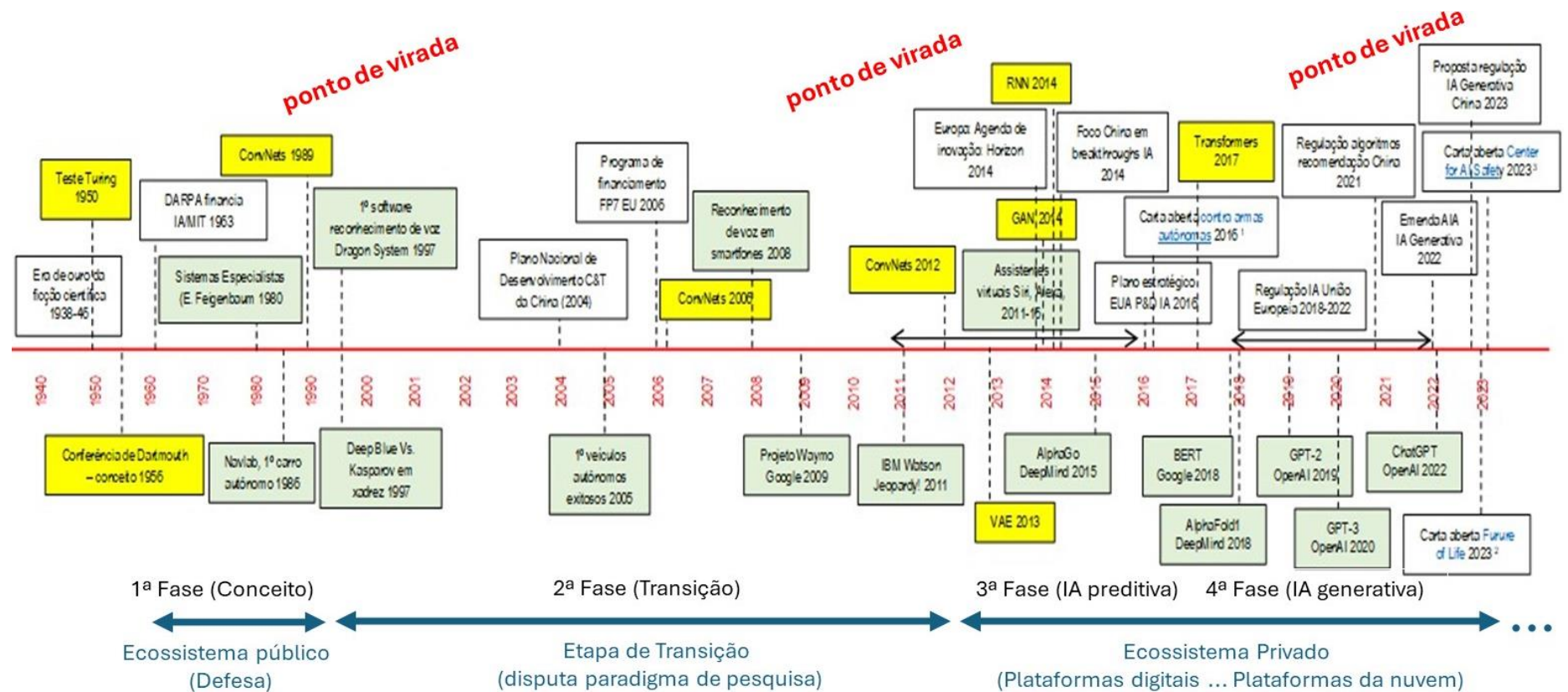
Entretanto, em termos formais, o surgimento da IA só ocorreu em meados dos anos 1950, quando o termo foi cunhado na Conferência de verão da Universidade Dartmouth (McCarthy et al., 1955), dando origem ao novo campo de pesquisa acadêmica (Russell e Norvig, 2010).

A compreensão da evolução e as perspectivas dessa poderosa tecnologia exigem uma interpretação de fatos relevantes a partir de uma organização lógica que não se limite à mera identificação de invernos e verões que normalmente norteiam as linhas do tempo da IA (Floridi, 2020). De uma perspectiva histórica, técnica e institucional, a evolução das tecnologias de IA nesta tese é proposta a partir da identificação de quatro fases distintas (Figura 2.1).

Uma primeira fase, desde o desenvolvimento do conceito e a estruturação de um robusto ecossistema governamental de pesquisa nos Estados Unidos voltado para a área de defesa, em virtude do caráter dual dessa tecnologia³⁷, se estendeu dos anos 1960 até a década de 1990. Desde cedo, projetos científicos de ponta foram desenvolvidos por uma ampla rede colaborativa de pesquisa militar, contemplando a interação entre universidades importantes, como Carnegie Mellon, Princeton, Instituto de Tecnologia de Massachusetts (MIT) e, mais tarde, Stanford, junto com empresas privadas como IBM, AT&T Bell Labs, Lucent Technologies, RAND Corporation e *Digital Equipment Corp* (DEC), além de *startups* como Nuance Corporation e

³⁷ A primeira demonstração exitosa da invenção dos perceptrons de Rosenblatt foi em um computador IBM704 do US Office of Naval Research, em 1958 (Lefkowitz, 2019)

Figura 2.1 – Linha do tempo - As fases da inovação da IA



Fonte: Elaboração própria

Dragon Systems, sob o comando da *Defense Advanced Research Projects Agency* – DARPA (na época, ARPA), que coordenava o ecossistema, definia os rumos da pesquisa e fornecia forte apoio financeiro. Também envolveu outras unidades do Departamento de Defesa (DoD) e organizações com outras missões, como a *National Aeronautics and Space Administration* (NASA) os *National Institutes of Health* (NIH) e a *National Science Foundation* (NSF) (National Research Council, 1999).

De fato, a pesquisa colaborativa foi o arranjo institucional chave para muitas inovações e desenvolvimento da indústria de alta tecnologia americana como retratado pela literatura³⁸, a partir de ecossistemas da NASA, DARPA e outras unidades do DoD, do qual participava um seleto grupo de empresas e universidades³⁹. O caráter dual da maioria das tecnologias da indústria de eletrônica e TICs não pode ser negligenciado na construção da competitividade da economia americana e afetou o ritmo e o padrão de inovação. O paradigma civil-militar acabou por influenciar, em diversos momentos, estratégias empresariais, entrelaçando-as à política industrial com conotação de segurança nacional e geopolítica. Com a IA não foi diferente, contemplando relações colaborativas intrincadas ao longo de décadas, em projetos em parceria da DARPA com pesquisadores acadêmicos, empresas privadas e *spin-offs* universitários.

No entanto, as primeiras inovações da IA só arriscaram timidamente chegar ao mercado nos anos 1980, com o desenvolvimento de computadores suficientemente poderosos para a operação de sistemas de IA, como os fluorescentes, mas tecnicamente limitados, sistemas especialistas (McCarthy, 1984), pertencentes à abordagem de pesquisa simbólica, e os primeiros sistemas conexionistas em redes neurais simples (*perceptrons*), simples classificadores lineares em uma única camada. Muitos desses projetos não foram meros experimentos, mas invenções que geraram patentes “embarcadas” em conjuntos maiores

³⁸ Os semicondutores, as TICs, o *Global Positioning System* (GPS) e a Internet são alguns importantes *breakthroughs* tecnológicos levados ao mercado pelo setor privado, a partir de sementes lançadas pelos projetos de defesa e financiamento público ao longo de décadas. O caso da Internet é um dos mais ilustrativos, onde um conjunto de iniciativas contribuíram lançando suas bases e gerando amplo debate sobre a origem pública ou privada da inovação (Moyer, 2012; Crovitz, 2012). Iniciativas pioneiras envolveram pesquisadores, redes governamentais experimentais (ARPANET, pela antiga ARPA) e infraestrutura pública (NSFNET, pela NSF), até o desenvolvimento da Ethernet pela Xerox PARC, na década de 1980, que transporta o *Internet Protocol* (IP), uma das tecnologias chaves da Internet. Essas organizações já foram apontadas pela literatura como parte importante no desenho e implementação da política industrial “invisível” associada à segurança nacional nos EUA, contribuindo com importantes inovações privadas em diversos setores (Wade, 2012; 2017; Mazzucato, 2013; 2021).

³⁹ Reconhecer o papel das agências governamentais no financiamento e orquestração de ecossistemas de inovação não minimiza, entretanto, a importância da pesquisa acadêmica, central para boa parte das inovações nos EUA, com uma enorme dívida do sucesso comercial americano em setores de alta tecnologia às atividades empreendedoras da universidade (Rosenberg, 2003:119).

compostos por sistemas de IA, software e hardware (National Research Council, 1999).

Essas inovações resultaram, evidentemente, de avanços técnicos nas duas principais abordagens de pesquisa no campo da IA, mas também resultaram de importantes mudanças institucionais do período. Primeiro, a reorientação da pesquisa em defesa de um enfoque prévio que cobria todo o ciclo de P&D e ênfase em pesquisa básica, passou a privilegiar a pesquisa aplicada depois das duras críticas de representar um financiamento à academia encoberto sob o véu da segurança nacional desde a aprovação da Emenda Mansfield, de 1969 (Walsh, 1979) e do intenso debate que se seguiu na década seguinte (Umpleby, 2003). No entanto, uma tecnologia ainda incipiente e um paradigma de pesquisa indefinido, além das limitações em capacidade computacional do período, não foi capaz de gerar os resultados esperados em termos de inovações no mercado, o que contribuiu para a progressiva desestruturação do ecossistema público de inovação.

Segundo, a regulação antitruste americana, cujo longo escrutínio sobre grandes empresas verticalmente integradas no campo das TICs, entre o final dos anos 1960 e início dos 1980, provocou a mudança da estrutura da indústria, originalmente integrada, para uma arquitetura modular, resultando em mercados fragmentados, tecnologias dispersas e especialização das empresas em módulos particulares (Rosenberg, 2010:44), desmembrando hardware, software e IA, o que possibilitou o surgimento e ascensão de novos *players*, como Microsoft, e, mais tarde, de uma nova geração de empresas do Silicon Valley, como Google, Meta, entre outras.

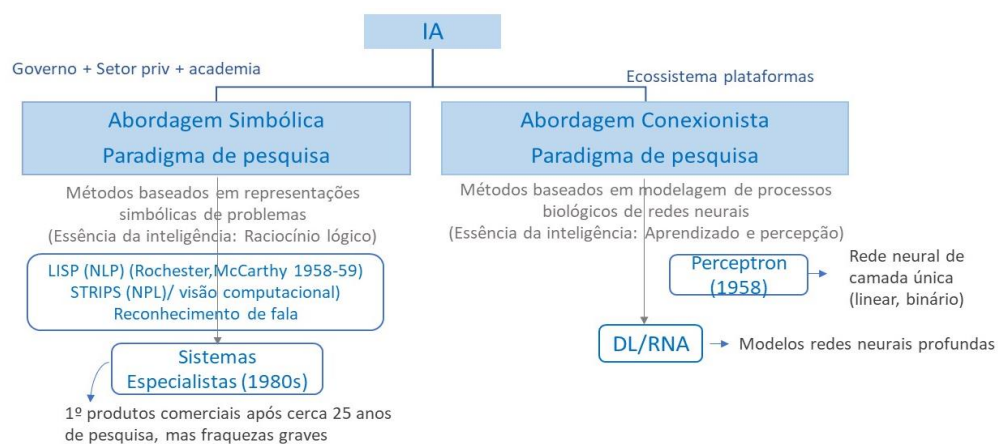
Terceiro, a importante grande guinada, a partir do final dos anos 1990, com a evolução da supercomputação, a expansão da Internet de natureza privada e a crescente disponibilização de dados digitais, junto aos avanços no campo do conhecimento de redes neurais artificiais, promoveu o primeiro grande ponto de inflexão da IA.

Com isso, tem início uma segunda fase da IA, ainda de transição, que se estendeu dos anos 1990 a 2010. O final da década de 1990 marcou um ponto de inflexão efetivo para a evolução das técnicas da IA, em meio ao avanço da Internet e da World Wide Web, em bases privadas, abrindo caminho para a estruturação privada do ecossistema de inovação da IA consolidado na fase subsequente, por empresas que surgiram com a economia digital. Mudanças no arcabouço legal e institucional alteraram o ambiente e a política pública dos EUA, delimitando uma nova fase de evolução da IA, sob governança privada, mas cujas inovações auspiciosas só se concretizaram na fase seguinte.

Esta segunda fase foi marcada por importante mudança ecossistêmica, em um contexto de coordenação privada e a disputa entre as principais abordagens de pesquisa da IA pelo papel

de paradigma dominante. Essas duas abordagens, a simbólica (ou logicista) e a conexionista (Norvig e Russel, 2010; Hinton 2021)⁴⁰, conviveram por décadas (Figura 2.2), disputando o lugar de paradigma de pesquisa, numa espécie de fase pré-paradigmática. A IA simbólica é inspirada no raciocínio lógico e entende a essência da inteligência como fundamentada no uso de regras para manipular expressões simbólicas. A abordagem conexionista, por seu turno, é biologicamente inspirada e a inteligência é a capacidade de aprender os pontos fortes das conexões em uma rede neural artificial.

Figura 2.2 – Principais abordagens de pesquisa da IA



Fonte: Elaboração própria a partir de Hinton (2021), Marcus (2022), National Research Council (1999)

A fase se inicia com a introdução e prova de conceito dos sistemas de IA em um arranjo típico da área, as competições da comunidade de desenvolvedores (devs.), cujo papel foi relevante para demonstrar o desempenho e a precisão dos sistemas de IA, em particular aqueles baseados em aprendizado profundo em arquiteturas de redes neurais artificiais, em sistemas de jogos complexos. O potencial dos sistemas de IA começa a ser percebido pela sua capacidade de superar campeões humanos, como o Deep Blue em xadrez - batizado como uma espécie de paráfrase ao apelido Big Blue da IBM -, remanescente dos sistemas especialistas; o IBM Watson, vencedor no programa televisivo Jeopardy! de perguntas e respostas, por meio de processamento de linguagem natural (NLP, sigla da expressão em inglês *Natural Language Processing*); e os modelos da DeepMind (adquirida pelo Google em 2014), com aprendizado por reforço nos jogos Atari 2600 e depois nos vitoriosos AlphaGo, AlphaZero e AlphaMu em

⁴⁰ Atualmente, há abordagens alternativas, como a IA neuro-simbólica, que combina características das mencionadas simbólica e conexionista (Chaudhuri et.al., 2021; Marcus, 2022; Garcez, Lamb e Gabbay, 2009), destinada a superar limitações para a evolução do aprendizado profundo, ou a IA Causal (von Kügelgen et al. 2022; Karimi et al., 2020), que busca a causalidade e não a correlação dos dados; ambas, contudo, ainda incipientes do ponto de vista de prover soluções de mercado.

Go⁴¹ e shogi, popularizando o conceito do aprendizado de máquina e evidenciando a eficiência e superioridade desses sistemas para o público. Mais que meras curiosidades, as competições de jogos popularizaram e consagraram o conceito do aprendizado de máquina e evidenciaram seu enorme potencial. Paralelamente, tem início a estruturação de um amplo e complexo ecossistema privado de inovação e a consolidação do exitoso modelo de negócios de plataformas digitais.

Uma terceira fase, que correspondeu à efetiva consagração dos modelos de aprendizado profundo em arquiteturas de redes neurais artificiais, possibilitou que inovações da popularmente chamada IA preditiva alcançassem exitosamente o mercado. A evolução da supercomputação e a crescente disponibilização de dados digitais pela Internet, junto aos avanços disruptivos no campo do conhecimento de redes neurais artificiais, entre meados dos anos 2000 e início dos 2010, representaram um importante ponto de inflexão e início dos *breakthroughs* tecnológicos da IA.

A década de 2010 marcou o início uma promissora jornada econômica e social da IA, com auspiciosos sistemas de redes neurais ganhando visibilidade, deixando o campo de artigos científicos, a predominância da pesquisa militar e competições complexas de jogos de tabuleiro para chegar ao mercado com inovações baseadas nesse um tipo particular de sistemas de redes neurais profundas, habilitando serviços inéditos como mecanismos de pesquisa *online*, assistentes virtuais, tradução automática, recomendações de *streaming*, comércio eletrônico e pagamentos digitais, disponibilizados primordialmente em plataformas digitais comandadas por um pequeno grupo de empresas, as chamadas BigTechs.

Nesse sentido, a IA é muitas vezes vista como um subproduto dos negócios em plataformas digitais, que criam valor a partir de massivos e variados dados e empregando extensa infraestrutura computacional, embora desenvolvida a partir de ecossistemas de inovação mais amplos.

Os avanços dos sistemas de aprendizado profundo foram extraordinários, consolidando a abordagem de pesquisa conexionista, a partir de importantes invenções algorítmicas - entre as quais as ConvNets do modelo AlexNet, em 2012 (Krizhevsky, Sutskever e Hinton, 2012) -, a partir da evolução em capacidade computacional e da quantidade de dados amplamente

⁴¹ Go é um complexo jogo de tabuleiro estratégico de soma zero criado na China há 2.500 anos. O AlphaGo é um sistema de IA com capacidade de reconhecer padrões e realizar previsões utilizáveis em outros campos. Foi desenvolvido pela DeepMind, posteriormente adquirida pelo Google. Em suas várias versões posteriores usam um algoritmo de busca para encontrar movimentos, com base no conhecimento previamente aprendido por uma rede neural artificial e treinamento extensivo. A versão AlphaZero, disponibilizada em 2017, é baseada em aprendizado por reforço, sem dados previamente rotulados por humanos, orientação ou conhecimento de domínio além das regras do jogo, apontada como o alvorecer dos modelos atuais.

disponibilizados na economia digital, atrelados ao fenômeno emergente dos modelos de negócios das plataformas online, principal inovação organizacional da economia digital.

Conforme sintetizaram LeCun et al. (2015:436):

“We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress.”

Nos últimos cinco anos, as fronteiras dos sistemas de aprendizado profundo evoluíram significativamente, representando um divisor de águas e início de uma nova fase da chamada IA generativa. Uma série de avanços nas arquiteturas dos modelos entre 2014 e 2017, no tipo de treinamento⁴² empregado, a ampliação do tamanho dos modelos, do volume e variedade dos dados e a expansão exponencial na capacidade computacional empregada, além de mudanças institucionais importantes relacionadas ao regime *open-source*, à expansão da computação na nuvem⁴³ como modelo emergente de negócios e à integralização de ativos de hardware pelas principais empresas de IA.

O ano de 2023, em particular, foi apontado como a efetiva mudança de paradigma, marcado por uma explosão de modelos e o surgimento de novos *players* que se juntaram às BigTechs cuja ascensão foi descomunal a partir da apropriação dos resultados da inovação da IA. Novos e poderosos modelos, mais intensivos computacionalmente, de diferentes domínios, apresentam capacidades inéditas de criação de novos conteúdos/dados (similares aos conjuntos de dados com que foram treinados), com destaque para modelos de texto, imagem, áudio, vídeo e código de computação, como os notórios grandes modelos de linguagem (LLMs, sigla de *large language models*), os modelos multimodais e os modelos de fundação.

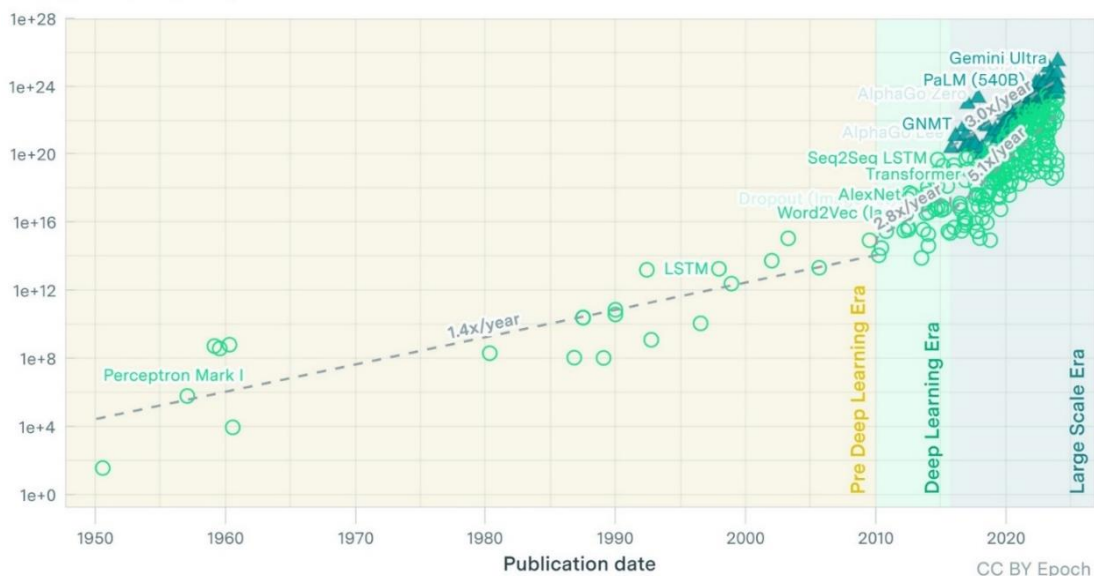
A constatação das fases da IA proposta nesta tese a partir de aspectos técnicos e institucionais guarda convergência com as três eras de progresso dos modelos de redes neurais profundas identificadas por Sevilla et al. (2022) com base no a partir da computação utilizada no seu treinamento (Gráfico 2.1).

⁴² O termo "treinamento" refere-se ao ajuste iterativo e ajuste fino de sua arquitetura em camadas, com o objetivo de melhorar a capacidade de um modelo de calcular previsões precisas. Computacionalmente, o objetivo é "definir de forma eficiente e otimizada os pesos e vieses" (Yui, 2019) de neurônios individuais, que são ligados entre si por diferentes camadas dentro da rede neural. Ferrari (2023) observa que o processo é tão intensivo em recursos que se tornou chave na competição entre as empresas líderes em torno de quem possui hardware mais potente para sustentar o treinamento.

⁴³ A computação na nuvem, expressão coloquial que corresponde à disponibilidade sob demanda de recursos computacionais, particularmente para armazenamento de dados e capacidade de processamento, para outras empresas, teve origem na 1ª fase da inovação da IA, com a DARPA encomendando o Projeto MAC ao MIT, visando desenvolver uma tecnologia que permitisse que um computador pudesse ser usado por duas ou mais pessoas, simultaneamente (Foote, 2021)

Segundo os autores, uma primeira era, pré-aprendizado profundo, entre 1950 e 2010, que corresponde às duas primeiras fases desta tese, foi marcada por crescimento lento da computação de treinamento, próxima da Lei de Moore (dobrando a cada 20 meses), que correspondente às duas primeiras fases definidas na tese a partir de avanços nas técnicas e mudanças institucionais relevantes. Uma segunda fase, a era do aprendizado profundo, tem início com modelos desenvolvidos a partir de 2010, em particular o modelo AlexNet, quando com aceleração da intensidade da computação de treinamento, passando a dobrar a cada 6 meses, coincidindo com a fase da IA preditiva. Por fim, a era da grande escala, iniciada em meados da década passada, quando as empresas passaram a desenvolver modelos com requisitos 10 a 100 vezes maiores em computação de treinamento, relativamente convergente com a fase da IA generativa. Este e outros trabalhos elaborados a partir das informações da base de dados da Epoch abriram novas perspectivas para esta tese ao chamar atenção para extraordinária escalada dos modelos da IA generativa em termos de capacidade computacional - ativos tangíveis que parecem se tornar o ativo complementar fundamental para a invasão e captura de valor da IA, apontando para novas perspectivas em termos do regime de apropriabilidade como será analisado.

Gráfico 2.1 – As três eras da Computação de treinamento de sistemas notáveis de aprendizado de máquina ao longo do tempo (02/07/1950 a 30/10/2023) por afiliação dos pesquisadores, medida em petaFLOPS (10^{15} operações de ponto flutuante)



Fonte: Sevilla et al. (2022) com dados da base (Epoch) (atualizados até 30/10/2023)

Nota: A computação é medida em petaFLOP, que é 10^{15} , que corresponde às operações de ponto flutuante (*Float pointing operation*, ou FLOP), um tipo de operação de computador, equivalente a uma adição, subtração, multiplicação ou divisão de dois números decimais, realizadas por segundo (FLOPS).

Essas duas últimas eras correspondem, respectivamente, às fases da IA preditiva e IA generativa identificadas nesta tese, com modelos que avançam exponencialmente em capacidade computacional, tornando este o principal ativo complementar da inovação em IA.

Ainda que os termos IA preditiva e IA generativa sejam genéricos, não científicos, se difundiram na mídia e serão empregados nesta tese com vistas a delimitar as características distintas da evolução das tecnologias da IA, acompanhadas por mudanças estruturais e redefinição de várias dimensões do regime de apropriabilidade e modelo de negócios. Serão esses dois tipos de IA que serão analisados de forma mais detalhada no restante deste capítulo e no capítulo seguinte.

As características distintas dos novos sistemas de IA generativa que proliferam aceleradamente no mercado, particularmente sua forte dependência de colossais recursos de *hardware* especializado, tanto por líderes incumbentes quanto empresas emergentes, parecem desafiar o padrão vigente de apropriabilidade e uma perspectiva inédita para o regime legal de PI, entrelaçada à reestruturação de seu modelo de negócios.

A emergência de novos sistemas baseados em modelos robustos e versáteis pareceu dar início a um novo capítulo na história dessa poderosa tecnologia, criando enormes expectativas em benefícios para a economia e sociedade e extensas possibilidades de inovação pelas perspectivas de sua implementação em setores *downstream*, ao mesmo tempo que gerou inquietações generalizadas e desafios regulatórios para mitigação de danos e riscos potenciais da IA, junto a preocupações sobre o excessivo poder de mercado e capacidade de direcionar o progresso técnico pelo conjunto reduzido de empresas BigTechs. Isso seu origem a várias formas de intervenção governamental, com potencial de influenciar as condições e o contexto no qual se dá a criação e captura de valor da inovação da IA, motivo pelo qual são retratadas no capítulo 5.

A implementação das tecnologias de IA em variados setores da economia e a realização de seu potencial de TPGs, parece ter tido que aguardar mais alguns anos. A IA 2.0 (Lee, 2023), não mais restrita a tarefas específicas da IA preditiva, em plataformas digitais, mas o uso de modelos com múltiplas e integradas funcionalidades da IA generativa, prometendo ampla disseminação econômica, foi atropelada pela necessidade de maturidade das novas técnicas à espera de sistemas generalisáveis e versáteis, ainda não alcançados, que alguns acreditam estará baseada em modelos de agentes de IA e da Artificial General Intelligence (AGI), possivelmente embarcados em smartphones, smartwatches e dispositivos que sequer foram desenvolvidos.

É interessante observar que patentes relacionadas a IA apresentaram uma trajetória quase tão longa quanto a evolução dessa tecnologia, conforme mostrado na Tabela 2.1. São discriminados os principais detentores de patentes de IA nos Estados Unidos, entre 1976 e 1997, na sua maioria de invenções de IA “embarcadas” em conjuntos contemplando hardware e software. A tabela também apresenta os principais titulares de pedidos de patentes relacionadas

a IA no mundo, para um período mais longo (1960-2017), em que, apesar do forte crescimento no número de pedidos e a emergência de novos *players* (em especial Microsoft, já na segunda posição do *ranking*, e Alphabet/Google, na décima posição), é interessante notar a resiliência de um conjunto não desprezível de firmas que se destacam no campo da IA, como a IBM, detentora de um portfólio expressivo de patentes em IA até a fase dos modelos preditivos, mas que não conseguirá alcançar o dinamismo dos novos *players*, em particular as BigTechs americanas e chinesas, na fase da IA generativa, como será visto⁴⁴. Oito empresas se mantiveram nas primeiras posições, identificadas pela mesma cor, sete delas empresas japonesas, incluindo a Matsushida, que passou a se chamar Panasonic. A maioria delas, entretanto, praticamente desaparecerá do cenário recente das patentes da IA, evidenciando as características distintas dos modelos de redes neurais profundas e a importância da estruturação de ecossistemas de inovação.

Tabela 2.1 – Vinte principais detentores de patentes nos EUA (1976-1997) e pedidos de patentes em IA no mundo (1960-2017)

20 principais titulares de patentes nos EUA (1976-1997) (NRC, 1999)					20 principais de pedidos de patentes no mundo (1960-2017) (WIPO, 2019)				
Institution	País	Criação	number of patents		Institution	Country	Criação	number of patent applications	
1 IBM Corp	US	1911	292		1 IBM Corp	US	1911	8.290	
2 Hitachi Ltd	JP	1910	192		2 Microsoft Corp.	US	1975	5.930	
3 Motorola, Inc.	US	1928	114		3 Toshiba Corp	JP	1875	5.223	
4 Mitsubishi Corp	JP	1870	94		4 Samsung Electronics Co	KR	1938	5.102	
5 Toshiba Corp	JP	1875	92		5 NEC Corp	JP	1899	4.406	
6 GE Co.	US	1892	91		6 Fujitsu Ltd	JP	1935	4.303	
7 NEC Corp	JP	1899	73		7 Hitachi Ltd	JP	1910	4.233	
8 Taligent Corp	US	1992	67		8 Panasonic Corp	JP	1918	4.228	
9 Toyota Motor Corp	JP	1937	60		9 Canon Inc	JP	1937	3.959	
10 US Philips Corp	US	n.a.	59		10 Alphabet/Google	US	1998	3.814	
11 Fujitsu Ltd	JP	1935	58		11 Siemens AG	DE	1847	3.539	
12 Lucent Technologies	US	1996	57		12 Sony	JP	1946	3.487	
13 Ford Motor Co	US	1903	53		13 Toyota Motor Corp	JP	1937	2.890	
14 Digital Equipment Corp	US	1957	53		14 NTT	JP	1985	2.772	
15 Westinghouse Electric	US	1886	48		15 State Grid (SGCC)	CN	2002	2.685	
16 Eastman-Kodak Co	US	1888	44		16 Mitsubishi Corp	JP	1870	2.683	
17 AT&T Inc	US	1883	44		17 Chinese Academy of Science	CN	1949	2.652	
18 Hughes Aircraft Co	US	1932	42		18 Ricoh Company	JP	1936	2.593	
19 Matsushida Electric Industrial Co	JP	1918	42		19 LG Corp	KR	1947	2.213	
20 Texas Instruments	US	1951	42		20 ETRI	KR	1976	1.936	

Fonte: National Research Council (1999) e WIPO (2019)

2.2. A IA preditiva: Redes neurais profundas, tipos de aprendizado e a não explicabilidade

As tecnologias da IA apenas há pouco mais de uma década alcançaram o mercado. Isso se reflete no fato de ainda não haver uma definição universalmente aceita pela comunidade

⁴⁴ Uma observação importante, que diz respeito a política de PI da China, com a única organização chinesa que aparece na tabela é a Chinese Academy of Science (CAS) na décima sétima posição no *ranking* de patentes acumuladas ao longo do longo período. A política parecia seguir uma orientação de atribuir a titularidade das invenções a CAS, mas parece ter mudado no período mais recente, com crescimento expressivo de pedidos de patentes de firmas como Tencent, Baidu, Alibaba, StateGrid, Ping An Insurance (WIPO, 2024; Statista, 2022).

científica. No Dicionário de Cambridge⁴⁵, a IA é descrita como o estudo da forma de produzir máquinas que possuam algumas das características da mente humana, como a capacidade de entender a linguagem, reconhecer imagens, resolver problemas e aprender. É também entendida como a "capacidade do sistema de interpretar corretamente os dados externos, de aprender com esses dados e de usar esse aprendizado para alcançar metas e tarefas específicas por meio de adaptação flexível" (Kaplan e Haenlein, 2019:15).

Com efeito, IA é um conceito geral e significado polêmico ao comportar a ideia de que máquinas podem mimetizar a inteligência humana. Estritamente falando, é um simples conceito guarda-chuva que abrange diferentes técnicas, aplicações funcionais e usos potenciais em diversos setores econômicos (WIPO, 2019; OECD, 2019).

Os propósitos deste trabalho, contudo, relacionados à apropriabilidade da IA, exigem definições formais mais precisas, requisito para garantir sua segurança jurídica. O Quadro 2.1 resume as definições usadas pelo *European Patent Office* (EPO) e *US Patent and Trademark Office* (USPTO), além da OECD, que tem definido os princípios da IA, e do regulamento europeu aprovado para a IA (*Artificial Intelligence Act*, 2024), evidenciando um entendimento distinto por essas organizações.

Quadro 2.1 – Definições legais da Inteligência Artificial

Organização	Definição IA	Referência
OECD	"An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations, or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy"	OECD.AI
EU AI Act (Article 3(1))	"'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"	EU AI Act (21/05/2024)
EPO	"AI corresponds to the ability of computer programs and machines to perform mental tasks commonly associated with humans, such as learning, reasoning and problem solving."	EPO (2022)
USPTO	"For patent applications and grants, we define AI as comprising one or more of eight component technologies" (planning control, knowledge processing, speech, AI hardware, evolutionary computation, natural language processing, machine learning, vision). These components span software, hardware, and applications, and a single patent document may contain multiple AI component technologies	Toole et al. (2020:2)

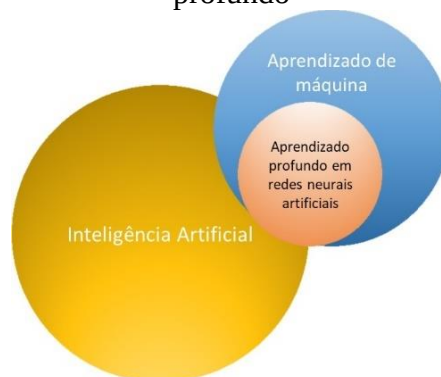
Fonte: Elaboração própria a partir das referências citadas

Os atuais sistemas auspiciosos de IA são constituídos por modelos de aprendizado profundo, subconjunto do aprendizado de máquina, em arquiteturas de redes neurais artificiais (Figura 2.3). São técnicas abrangidas pelo hiperônimo IA que não objetivam a inteligência humana, mas sim treinar máquinas para realizar tarefa específica e fornecer resultados precisos

⁴⁵ Disponível em: <https://dictionary.cambridge.org/us/dictionary/english-portuguese/artificial-intelligence>. Acesso em 15 dez. 2020

pela identificação de padrões. É, assim, um caso de IA “...que permite às máquinas extrair conhecimento dos dados e aprender com eles de maneira autônoma” (Google Cloud).⁴⁶

Figura 2.3 – Representação da IA, aprendizado de máquina e o subcampo de aprendizado profundo



Fonte: Elaboração própria

O aprendizado de máquina é talvez a técnica mais conhecida da IA por possibilitar as primeiras aplicações bem-sucedidas. Emprega algoritmos e modelos estatísticos/matemáticos para organizar dados e reconhecer padrões de forma a automatizar decisões com base no aprendizado. Embora os primeiros sistemas de aprendizado tenham sido desenvolvidos no início da década de 1960, seu desempenho somente melhorou após a disponibilização de grandes conjuntos de dados de treinamento (Nilsson, 2009).

O aprendizado profundo data da década de 1980 e, após relativo ostracismo na década seguinte, recuperou o interesse da pesquisa até se tornarem comercialmente importantes a partir de meados dos anos 2000⁴⁷. Engloba técnicas promissoras, sofisticadas, complexas, construídas sobre arquiteturas de redes neurais artificiais, capazes de trabalhar com BigData. Seu êxito não decorre simplesmente de escrever um conjunto de instruções para executar uma tarefa, mas permitir que o sistema determine o melhor modelo de previsão para um problema com base em grandes quantidades de dados e um número limitado de parâmetros definidos por humanos (Nardin, 2021:12). Expressivos conjuntos de dados de treinamento de qualidade se tornaram importantes e marcando uma mudança de paradigma em relação a abordagens anteriores (Russell e Norvig, 2010).

A literatura de redes neurais data da década de 1950, mas só se tornou viável com a disponibilidade de grandes conjuntos de dados da Web e poderosas unidades de processamento gráfico (WIPO 2019a). Seus avanços foram em grande parte devidos à longa pesquisa sobre redes neurais artificiais, com contribuições revolucionárias entre 2006 e 2012 (LeCun et al.,

⁴⁶ Google Cloud. Disponível em <https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning?hl=pt-br>. Acesso em 11 out. 2024

⁴⁷ Invenções relacionadas à aprendizado de máquina tiveram crescimento exponencial e representam grande parte dos pedidos de patentes e metade das publicações científicas no mundo em IA entre 1998/2016 (WIPO 2019a).

2015) para eficiência e precisão dos modelos⁴⁸. Novos algoritmos e arquiteturas de aprendizado desenvolvidos para redes neurais profundas aceleraram esse progresso de modelos treinados para classificar e agrupar em clusters dados de entrada, popularmente chamados de IA preditiva.

As redes neurais convolucionais (ConvNets, sigla de convolutional neural networks), por exemplo, as mais utilizadas da IA preditiva, quase desapareceram da agenda de pesquisa em visão computacional e aprendizado de máquina até que avanços nos anos 2000 a 2010 as tornaram base do desenvolvimento de produtos/serviços de grandes firmas de tecnologia (Google, Facebook/Meta, Microsoft, IBM, Yahoo!, Twitter e Adobe) e startups (LeCun et al., 2015).

Os modelos de redes neurais profundas usados em aplicações funcionais como processamento de linguagem natural, reconhecimento de imagem e vídeo, entre outras, funcionam inspirados no cérebro humano, construídos em várias camadas não lineares sobrepostas (*hidden layers*) de processamento de dados, cada uma representando características diferentes dos dados e, ao ajustar pesos das conexões entre esses “neurônios”, podem aprender e identificar características dos dados, como em imagens, bordas, cores, texturas, etc., distinguindo diferentes classes de imagens.

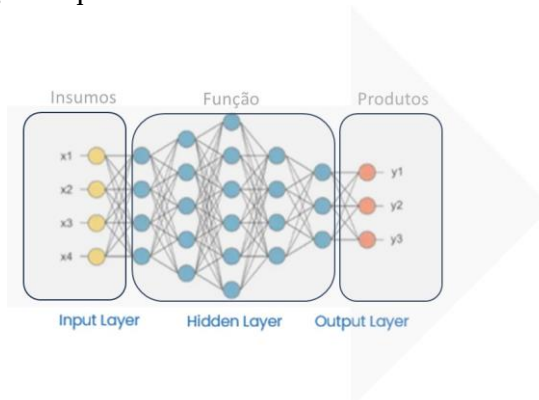
O processo de treinamento a partir dos dados é crucial para o desempenho dos modelos. A Figura 2.4 é uma representação genérica do treinamento de modelos de redes neurais profundas, estruturado em camadas, onde a primeira camada recebe um vetor que representa os dados de entrada e cada camada subsequente aplica alguma transformação da saída da camada anterior. Correspondem a funções complexas que relacionam entradas (insumos, ou uma variável independente) e saídas (ou produtos, uma variável dependente) compostas por diversas camadas de redes neurais.

A primeira camada recebe um vetor que representa os dados de entrada e cada camada subsequente aplica uma transformação na saída da anterior. Cada camada consiste em “nós” que recebem dados de entrada, multiplicam cada entrada por um coeficiente (peso), somam o resultado e passam esse resultado por uma função de ativação. A entrada é retroalimentada por várias camadas ocultas, com cada uma delas gerando resultados aos quais são aplicadas transformações que podem ser aprendidas automaticamente, gradualmente ajustando e refinando os pesos da unidade neural para minimizar sua função de perda. Por meio de

⁴⁸ Há vários tipos de redes neurais artificiais, com destaque para as redes neurais convolucionais, conceito introduzido por LeCun (1989) e contribuições importantes de Hinton e Salakhutdinov (2006) e Krizhevsky, Sutskever e Hinton (2012).

aplicações repetidas, essas transformações são ajustadas para mapear as entradas próximas das saídas (soluções computacionais) desejadas.

Figura 2.4 - Representação esquemática do treinamento de sistema de aprendizado profundo



Fonte: Elaborado a partir de imagens livres na Internet

A abordagem mais comum de treinamento era o aprendizado supervisionado (onde o modelo “aprende” a partir de observação de dados brutos previamente categorizados ou rotulados), até outras começarem a se destacar. São elas: o aprendizado não supervisionado (ou auto supervisionado, que não requer rotulagem prévia dos dados de treinamento), o semi-supervisionado (que requer rotulagem parcial), o aprendizado por reforço (que aprende de forma iterativa e interativa a partir do ambiente), o aprendizado de transferência (onde o aprendizado obtido em uma tarefa é transferido para outras, exigindo menos dados em novos aprendizados) e o aprendizado federado, no qual os algoritmos aprendem enquanto processam dados que não precisam ser hospedados juntos no mesmo local⁴⁹.

Os algoritmos e a arquitetura do modelo envolvem funções matemáticas, com pesos e parâmetros de treinamento definidos pelo programador, que mudam com o aprendizado com os dados, apresentando ganhos crescentes em acurácia e precisão, sem que os especialistas saibam claramente. Isso ocorre porque são baseados em métodos empíricos (heurísticos), não teóricos (Drexler, 2019). Os especialistas sabem que as técnicas de IA funcionam, mas não exatamente como funcionam (Geiger, 2020).

No estágio atual, o funcionamento interno dos modelos não é completamente compreensível mesmo para especialistas (Meske et al., 2022; Camburu, 2020) por uma “autonomia” relativa frente à programação original que os diferencia do software tradicional (WEF, 2018). Sistemas de redes neurais profundas têm um duplo atributo - não comparável a nenhuma outra técnica de IA – de realizar atividades com eficiência crescente em muitos campos de aplicação, mas sem se apoiar em uma base teórica bem definida. É aí onde reside os modelos de IA serem considerados verdadeiras “caixas pretas”, pois sua estrutura muda com o

⁴⁹ Sobre alguns desses conceitos, ver <https://www.enjoyalgorithms.com/>. Acesso em 12 jun. 2021

aprendizado/adaptação com novos e crescentes volumes de dados. Ao hermetismo dos complexos processos internos de aprendizagem dos modelos e opacidade dos algoritmos subjacentes, somam-se vieses implícitos nos dados que se reproduzem nos resultados dos modelos (Bender et al., 2021).

A referência à “caixa preta” (Carabantes, 2020) decorre do hermetismo dos processos internos de aprendizagem. Quanto maior a acuracidade, maior a opacidade dos algoritmos e menor a transparência dos modelos, resultando em um *trade-off* entre desempenho e explicabilidade com impactos potenciais negativos sobre pessoas, economias e sociedades (Alt, 2018). Há autores, contudo, que consideram falaciosa e imprecisa a crença de que a precisão deve ser sacrificada para a interpretabilidade (Rudin e Radin, 2019).

Isso remete ao tema da explicabilidade dos modelos da IA, que não é novo na agenda de pesquisa de sistemas de informação, remontando aos anos 1980, que ressurgiu nos anos 2000 em meio com o paradigma de redes neurais artificiais. Há um reconhecimento de que esses sistemas “não são capazes de declarar inerentemente o conhecimento que contêm, nem de explicar os processos de raciocínio pelos quais passam” (Meske et al., 2020:3). Tal constatação engendrou publicações sobre a capacidade de explicação de sistemas de IA, ou “Inteligência Artificial Explicável” (XAI, sigla da expressão em inglês *explainable artificial intelligence*), termo introduzido formalmente em um artigo apresentado em uma conferência de ciência da computação em 2004 (Van Lent et al., 2004: 900).

Não há uma definição consensual de XAI. Para Adadi e Berrada (2018:52140) “*XAI term tends to refer to the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept*”. Para a DARPA⁵⁰, corresponde à construção de modelos de aprendizado de máquina que sejam explicáveis enquanto mantêm alto nível de desempenho, permitindo que usuários os entendam, confiem e gerenciem com eficácia”⁵¹.

A capacidade de explicar se tornou um pré-requisito para uma IA justa, responsável e confiável, visando impedir comportamentos indesejáveis do sistema. Confiar em um modelo de caixa preta significa confiar nas equações do modelo e todo o banco de dados sobre o qual foi construído e treinado. Uma IA não explicável é potencialmente incontrolável (Meske et al.,

⁵⁰ Disponível em: <https://www.darpa.mil/program/explainable-artificial-intelligence>. Acesso em 5 jul., 2021

⁵¹ Os termos “explicabilidade” e “interpretabilidade” são muitas vezes utilizados como sinônimos, embora Meske et al. (2020) distingan IA interpretável como aquela em que os humanos podem entender diretamente o raciocínio e as ações da máquina sem explicações adicionais (característica passiva do artefato, segundo Rudin 2019), enquanto na IA explicável os humanos precisam de explicações como uma proxy para entender os processos de aprendizado e raciocínio do sistema (Adadi e Berrada, 2018).

2020) e tem motivado a pesquisa em XAI com emprego de diferentes instrumentos para tentar explicar o funcionamento interno do sistema (Ras et al., 2018). As mudanças de pesos e parâmetros e o funcionamento interno não completamente compreensível implicam em uma “autonomia” relativa dos modelos de IA frente à programação original.

Os sistemas de IA são frequentemente classificados por conceitos que remontam à década de 1980 (Searle, 1980). A IA “fraca”, ou “IA estreita” (ANI, sigla de *artificial narrow network*), corresponde aos atuais sistemas existentes, capazes apenas de realizar tarefas específicas, ao passo que a IA “forte”, ou “IA geral”, ou AGI (sigla de *artificial general intelligence*), ainda não existente, que alguns acreditam virá a ter plena autonomia, aplicação ampla e capacidade de generalização no futuro (Goertzel e Pennachin, 2007). O conceito de AGI é puramente teórico, cuja origem é incerta, mas que Goertzel (2011) atribui a Shane Legg, cientista chefe e cofundador da DeepMind. O termo AGI é um dos mais polêmicos e se tornou recorrente no debate desde o surgimento da IA generativa, com pesquisadores da Microsoft tendo documentado comportamento complexo pelo modelo GPT-4 (OpenAI), identificando sinais de raciocínio e “compreensão profunda e flexível” de habilidades humanas, que descrevem como “faíscas” de AGI (Bubeck et al., 2023). Há ainda o conceito hipotético de superinteligência (ASI, sigla de *artificial superintelligence*), capaz de superar a inteligência humana, que teve repercussão com Bostrom (2016), autor que explora o tema há décadas (Bostrom, 1998).

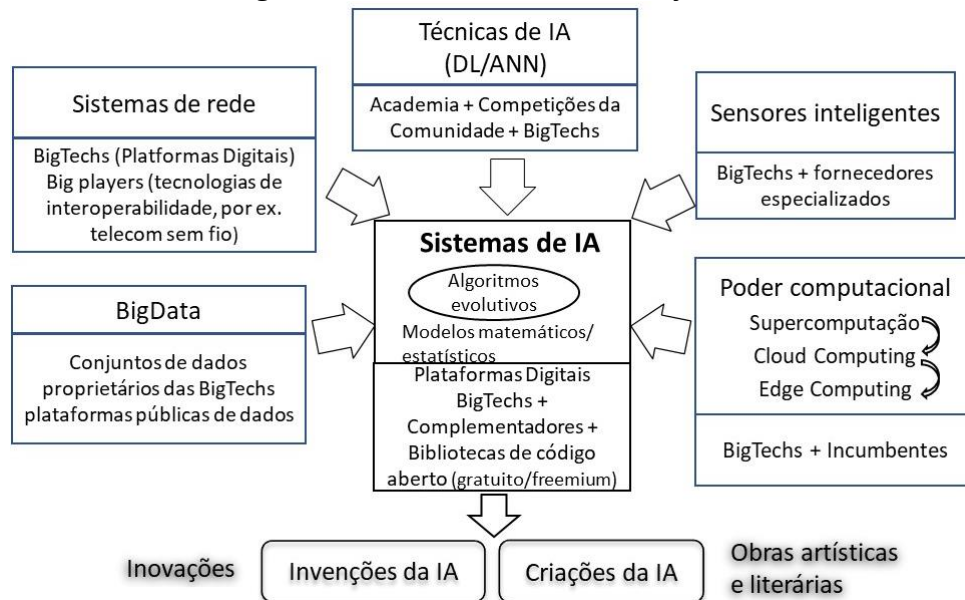
2.3. O ecossistema de inovação da IA preditiva

Os sistemas de redes neurais profundas são desenvolvidos a partir do mais amplo ecossistema privado de inovação conhecido, que contempla fontes diversas de conhecimento, internas e externas às empresas líderes, envolvendo pesados investimentos em P&D e formas inovadoras de interação com múltiplos *players* que transcendem a propriedade direta e ideias difundidas de aquisição maciça de empresas nascentes e relações de subordinação nas plataformas digitais.

O conceito de ecossistema de inovação, que guarda relação com as características de modularidade das TICs, complementaridade e inovação aberta da IA, interessa particularmente com vistas a explorar como as empresas criam valor. Esses ecossistemas em permanente evolução contemplam redes de interconexões cada vez mais amplas e complexas entre *stakeholders* diversos, coordenadas por empresas centrais que se beneficiam das vantagens do pioneiro.

Definitivamente, inovações relacionadas a IA tiveram que esperar até o desenvolvimento de partes importantes de seu ecossistema, cujos recursos e tecnologias que o compõem, além de respectivos provedores, são apresentados na Figura 2.5.

Figura 2.5 – Ecossistema de Inovação da IA



Fonte: Elaboração própria com base na literatura

O ecossistema de inovação da IA abrange avanços nas técnicas de aprendizado, algoritmos autodidatas e BigData analytics para mineração⁵² de gigantesco volume e variedade de dados gerados, coletados e armazenados de diferentes fontes, principalmente pelas plataformas digitais, permitindo que os sistemas de IA e os algoritmos subjacentes aprendam as características dos dados. Este ecossistema inclui, das técnicas de IA e da BigData, extraordinário poder computacional em supercomputadores de alto desempenho com capacidade de armazenamento e, principalmente, velocidade de processamento, em arquiteturas computacionais diferenciadas para realizar cálculos complexos paralelamente, projetadas especificamente para IA, além da integração a sistemas de rede e poderosos sensores⁵³ inteligentes que capturam dados mais confiáveis, bem como tecnologias de interoperabilidade e conectividade (Bastos, 2021). As principais partes componentes do ecossistema de inovação da IA são analisadas nos tópicos seguintes.

⁵² Mineração de textos e dados corresponde ao emprego de técnicas analíticas automatizadas destinadas à análise de texto e dados em formato digital.

⁵³ Sensores são fundamentais para identificação, coleta, transmissão e armazenamento de dados. Algoritmos de IA com capacidade de analisar e se adaptar ao ambiente necessitam de sensores, que são dispositivos eletrônicos que captam dados sobre o ambiente. Um veículo autônomo requer a combinação de tecnologias e sensores que fornecem dados para os algoritmos guiarem seu movimento. Seu papel é central na onisciência digital (Zuboff, 2019).

2.3.1. Sistemas de IA

Sistemas de redes neurais profundas são sistemas computacionais que processam e analisam enormes conjuntos de dados por meio de sofisticados e poderosos recursos computacionais, baseados em modelos estatísticos de probabilidade em funções matemáticas não lineares com pesos e parâmetros de treinamento definidos por programadores, mas que mudam ao longo do processo de aprendizagem. A estrutura do modelo com algoritmos especiais aprende a partir de correlações observadas em massivos conjuntos de dados, desenvolvendo, modificando e adquirindo maior precisão ao longo do processo de treinamento e em vários níveis de abstração.

Contemplam modelos matemáticos/estatísticos com algoritmos especiais capazes de reconhecer padrões, determinar preferências, aprender e fazer previsões apoiadas em correlações em massivos conjuntos de dados que caracterizam a economia digital. O aprendizado (ou a adaptação) em amplos conjuntos de dados possibilita ganhos sucessivos em precisão e acuracidade nos modelos, habilitando capacidades de fazer classificações, recomendações e previsões, com índices crescentes de acuracidade e relativa autonomia em relação à programação original, empregados em diversas aplicações funcionais e prometendo ser a tecnologia mais promissora do nosso tempo.

Algoritmos são regras e procedimentos lógicos que levam à solução de um problema em um número finito de etapas. São instruções, passo a passo, codificadas, como no software de computador tradicional. No entanto, os algoritmos de aprendizagem (ou algoritmos evolucionários) dos sistemas de IA têm a capacidade de, a partir de um dado conjunto de dados e treinamentos sucessivos, atingirem resultados com maior precisão à medida que a aprendizagem avança.

A escolha de modelos específicos e tipos de treinamento depende atualmente de métodos heurísticos, abordagem de resolução de problemas baseada na experiência e na intuição. Não em uma metodologia científica, mas empírica. As redes neurais modificam seu próprio código para encontrar e otimizar *links* entre entradas e saídas.

Esses modelos contemplam dois momentos: um estágio de treinamento, no qual o modelo recebe dados, que podem ser previamente rotulados ou não rotulados por trabalho humano, para identificar preferências e padrões em amplos conjuntos de dados; e um estágio de inferência, no qual o modelo treinado é utilizado para previsões e orientar decisões complexas, como no comércio eletrônico, mecanismos de pesquisa online entre outros.

O êxito da IA preditiva no mercado e da sua apropriação de valor contou com aquela que talvez seja a maior inovação organizacional contemporânea: o modelo de negócios de

plataformas digitais. Ao criar ambientes virtuais que intermediam interações diversas entre indivíduos, empresas e outras instituições, transações e compartilhamento de informações, essa forma organizacional se tornou a peça-chave da economia digital. Há um número expressivo e variado de plataformas digitais, algumas meras usuárias de tecnologias de IA e apoiadas em infraestrutura computacional de terceiros⁵⁴. Embora as plataformas incorporem tecnologias de IA, nem todas são desenvolvedoras (produtoras) de IA.

Existem tipos distintos de plataformas – de transação, de inovação, híbrida e, ainda, de infraestrutura –, algumas das quais são *loci* da inovação em IA, como as plataformas híbridas (Gawer, 2021), estruturadas pelas BigTechs americanas e suas rivais chinesas, que compartilham características econômicas, empresariais e de governança comuns, cruciais para a criação e captura de valor das inovações da IA preditiva, ainda que apoiadas em estratégias empresariais muitas vezes diversas.

As empresas dominantes no negócio de plataformas digitais são firmas tipicamente orientadas por dados (*data-driven*) e se tornaram as líderes da IA preditiva, pois as plataformas digitais asseguram uma posição privilegiada para coleta, armazenamento e análise de massivos conjuntos de dados analisados com tecnologias de BigData e, paralelamente, e tem funções habilitadas ao inserir recursos algorítmicos de IA como insumos cativos (Jacobides, Brusoni e Candelon, 2021; Cutolo e Kenney, 2021), que induzem preferências de clientes por produtos/serviços e maximizam o envolvimento dos usuários.

A IA é, com efeito, a engrenagem chave da maioria das plataformas conhecidas, com algoritmos de recomendação e persuasão inseridos nos seus mecanismos de funcionamento, tornando os efeitos de rede⁵⁵ - mercados onde o valor que os consumidores atribuem a um bem aumenta à medida que outros também consomem o bem – mais potentes do que no contexto de mercados físicos (Choudary, 2016; Parker, Van Alstyne e Choudary, 2016).

A coordenação das plataformas pode ser considerada um pré-requisito e uma forma de reduzir custos de transação, diante de problemas clássicos de coordenação e governança em inovações sistêmicas (transversais) – que nas inovações discretas são alcançadas por meio de transações de mercado e integração vertical -, sujeitas a incertezas substantivas e contingências

⁵⁴ A respeito da economia de plataformas, ver também <https://www.platformeconomy.com/>. Acesso em 15 set. 2024

⁵⁵ Efeitos de rede (*network effect*) ocorrem quando o valor de um bem ou serviço aumenta quanto maior o número de usuários, que atraem novos usuários e novos provedores de serviços. Os mecanismos de recomendação/persuasão dos sistemas de redes neurais profundas contribuem para os efeitos de rede, alcançando maior engajamento de usuários ao fornecer conteúdo personalizado que, por seu turno, contribuem para o maior aprendizado dos sistemas.

imprevistas pela forte interdependência multilateral e co-especialização em ativos (Foss, Schmidt e Teece, 2023).

Ecossistemas de plataformas digitais possuem características específicas como contextos multi-invenção e mercados multilaterais (Jacobides et al., 2018), caracterizados por fortes economias de escala e *network effects* (Cusumano, Gawer e Yoffie, 2019), com fronteiras opacas entre segmentos da “indústria” e um padrão de competição essencialmente indireto.

São ecossistemas marcados tanto por competição quanto cooperação (“*coopetition*”) (Petit e Teece, 2020), onde a criação e captura de valor de produtos, serviços e inovação decorrem da interação entre os líderes das plataformas e complementadores, que cooperam no curto prazo e competem no longo prazo. Uma competição em três níveis, entre as diferentes plataformas, entre elas e seus complementadores e de complementadores entre si. É o caso dos aplicativos móveis desenvolvidos por diferentes complementadores atrelados às plataformas da Apple (*App Store*) e Google (*Play Store*), que funcionam nos respectivos sistemas operacionais Android e iOS. A complementaridade está na essência das plataformas digitais, com distintos ativos nelas incorporados e cujo controle afeta a captura de valor da inovação (Teece, 2018a; Teece, 2016; Petit e Teece, 2020).

A digitalização aumentou muito a criação de valor a partir dos dados, possibilitadas por ecossistemas que contemplam recursos específicos (em especial técnicas da IA, capacidade computacional e, sobretudo, os conjuntos expressivos de dados), cuja apropriabilidade de valor decorreu da implementação exitosa das plataformas digitais, o mais extenso ecossistema de inovação aberta (Chesbrough et al., 2024), que asseguraram o controle e a coordenação de algumas peças do ecossistema de inovação cada vez mais amplo e com governança estendida (Paredes-Frigolett e Pyka, 2022).

As inovações da IA preditiva são geradas em ecossistemas amplos que contemplam, mas não se resumem, aos recursos das plataformas digitais, abarcando conhecimentos internos das suas empresas orquestradoras e de outros *players*, que podem estar vinculados ou não à essa estrutura organizacional peculiar. É nesse sentido que se busca tratar de forma distinta os conceitos de plataforma digital, enquanto modelo de negócios, do conceito mais amplo de ecossistema de inovação, que abarca um conjunto muito mais amplo de redes e interconexões que parece caracterizar o progresso técnico na organização empresarial moderna.

Portanto, as inovações da IA emergem de extensos ecossistemas de inovação aberta, estruturados pelas empresas líderes, que contemplam as plataformas digitais. No entanto, nem os ecossistemas, nem as plataformas, nem mesmo seus “orquestradores”, são um bloco uníssono e homogêneo. Mesmo que exista um comportamento convergente em diversos

aspectos, há diferenças fundamentais entre elas decorrentes de distintas estratégias empresariais das líderes em IA e que se tornarão mais evidentes no cenário de rivalidade aberta da IA generativa, como será visto.

2.3.2. BigData

Grandes conjuntos de dados dão origem à noção de BigData, originalmente descrita pela Tech American Foundation por três fatores (volume, velocidade e variedade) (Grishikashvili et al., 2014), incorporando progressivamente novos atributos e diferentes tipos de dados. Esses dados podem ser estruturados (bancos de dados como planilhas e tabelas), semiestruturados (arquivos de texto, CSV, XML, etc.) ou não estruturados que são obtidos de fluxos rápidos em tempo real (e-mails, imagens, dados gerados em redes sociais, buscadores, compras online e fluxos coletados por sensores conectados em todo o mundo) (Jain et al. 2016; Zuboff, 2019).

Inovações em redes neurais profundas são, em grande parte, produto do massivo e variado conjunto de dados (Jones e Tonetti, 2020). A maior parcela do fluxo atual de cerca de 2,5 quintilhões de dados diários tem como base as plataformas digitais, comandadas pelas BigTechs e os variados produtos e serviços que oferecem. A Internet, fundamental nesse processo, atinge atualmente cerca de 5 bilhões de pessoas (mais de dois terços da população mundial), das quais mais de 90% usuárias de redes sociais, com mais de 400 milhões de terabytes de dados criados diariamente, 90% dos quais nos últimos dois anos e crescendo exponencialmente a cada ano.

O mecanismo de busca do Google processa cerca de 2 trilhões de pesquisas por ano. O Youtube tem 2,5 bilhões de usuários. O Facebook, da Meta, a maior plataforma de mídia social, tem em torno de 3 bilhões de usuários ativos em suas redes, o Instagram e WhatsApp, cerca de 2 bilhões cada um, com cerca de 100 bilhões de mensagens enviadas por dia pelo WhatsApp. Mais de 50% dos gastos globais com anúncios online, que devem atingir US\$ 677 bilhões em 2024, passam pela Meta e Alphabet. A Apple tem mais de dois bilhões de usuários ativos do iPhone. O sistema operacional (OS) Windows 10 da Microsoft tem 1,3 bilhão de usuários, com 80% do mercado de OS em desktops, seguido pelo Chrome (Google), enquanto a liderança no mercado de celulares é do Android, com *market-share* de 71,6% com seus mais de 3 bilhões de usuários ativos. As gigantes chinesas Alibaba, Tencent (WeChat⁵⁶) e Baidu, entre outras, completam o cenário das principais firmas operadoras de plataformas, controladoras de amplos conjuntos de dados e inovadoras em IA.

⁵⁶ A WeChat (Tencent) é mais diversificada de todas as plataformas, contendo aplicativos para absolutamente tudo, desde compra de produtos, rede social, transferências e pagamentos online, etc. Em síntese, um serviço completo que combina aplicativos como Facebook, Twitter, Uber, Instagram.

As empresas investem pesadamente em todas as etapas da cadeia global de valor dos dados, desde a coleta (principalmente por meio das plataformas), o armazenamento (construindo seus próprios centros de processamento de dados em plataformas na nuvem) e até transmissão por meio de satélites e cabos submarinos de longa distância. Mercados de dados na economia digital ainda são incipientes e geralmente baseados na troca de acesso e serviços, em vez da venda explícita de dados específicos. Plataformas de dados comerciais como Acxiom (dados do consumidor), Bloomberg (dados financeiros) ou LexisNexis (dados de seguros), operam como intermediários nos mercados de dados por meio de relações contratuais bilaterais.

É indiscutível a importância dos grandes conjuntos de dados como ativo crucial para criação e captura de valor das inovações da IA. Segundo Petit e Teece (2020:7-8) *“Today, the rate and direction of Amazon, Google, Netflix and Facebook’s diversification is given by the data collected as a byproduct of user engagement with their platforms (including transactions/purchases)”*. Os dados não são um subproduto das atividades dessas firmas, mas parte importante de seu *core business* e ponto focal para a inovação em IA.

Petit e Teece (2020) consideram o controle sobre grandes conjuntos de dados (sobre comportamento e preferências dos usuários) como uma competência chave, que exige o desenvolvimento de habilidades relevantes para aprendê-los e explorá-los, ativos cruciais para a criação e captura de valor na economia digital. Cockburn, Henderson e Stern (2018) destacam que os primeiros entrantes em alguns campos de aplicação de IA podem criar uma vantagem competitiva duradoura sobre rivais em potencial por meio do controle e propriedade sobre os dados, sem necessitar de DPI. De fato, os dados são ativos estratégicos que garantem vantagem competitiva dinâmica para as empresas e fonte da inovação baseada em aprendizado profundo, permitindo aos seus proprietários gerar valor e oportunidades de diversificação.

2.3.3. Investimentos em P&D e aquisição de empresas

A criação de valor da IA por parte das principais empresas, resultante da expansão de ecossistemas, demanda expressivos investimentos em P&D – US\$ 225 bilhões, em 2023, pelas cinco BigTechs americanas – superando concorrentes nos respectivos segmentos/setores de atuação, como IBM (até pouco tempo detentora do maior portfólio de patentes em IA) e algumas se equiparando a líderes históricos em setores intensivos em conhecimento, como farmacêuticas (Tabela 2.2).

Embora tanto o investimento público quanto privado em IA venha crescendo substancialmente, é superior e avança mais rapidamente na indústria. Por exemplo, o investimento das agências governamentais americanas em IA, sem considerar defesa, em 2021, somaram US\$ 1,5 bilhão, comparável ao investimento da Alphabet na DeepMind, em 2019.

Enquanto o investimento privado americano foi de US\$ 52,9 bilhões, em 2021 (que não inclui os investimentos internos em P&D das BigTechs), o investimento do governo federal foi de US\$ 10,8 bilhões, majoritariamente em defesa, pelo DoD (Zhang et al., 2022a).

Tabela 2.2 – Despesas com P&D e intensidade de P&D (pelo lucro bruto ou por receitas), empresas selecionadas

Setor/Industria	Empresa	Gastos em P&D (US\$ bi)	Lucro bruto (US\$ bi)	Receita (US\$ bi)	P&D/Lucro bruto (%)	P&D/ Receita (%)	ano fiscal
Tecnologia/Eletrônicos de consumo	Apple	29,92	169,15	383,29	18%	8%	set/23
	HP	2,35	10,24	29,14	23%	8%	out/23
	Dell	2,72	22,79	102,3	12%	3%	fev/23
Tecnologia/Software-Infraestrutura	Microsoft	27,2	146,2	211,92	19%	13%	jun/23
	Oracle	8,92	37,82	52,96	24%	17%	mai/24
	Service Now	2,12	7,05	8,97	30%	24%	dez/23
Consumidor discricionário/Varejo na Internet	Amazon	85,62	270,05	574,79	32%	15%	dez/23
	Alibaba	52,26	354,85	941,17	15%	6%	mar/24
	JD.com	16,39	95,15	1.084,66	17%	2%	dez/23
	Pinduoduo	9,57	155,92	247,64	6%	4%	dez/23
Serviços de comunicação/Conteúdo e Informações na Internet	Alphabet	43,71	175,02	307,39	25%	14%	dez/23
	Meta Platforms	36,49	108,90	134,90	34%	27%	dez/23
	Snap	1,91	2,49	4,61	77%	41%	dez/23
	Baidu	24,19	69,57	134,60	35%	18%	dez/23
Tecnologia/Serviços de Tecnologia da Informação	IBM	6,78	34,30	61,86	20%	11%	dez/23
	Infosys	-	5,47	18,56	-	-	mar/24
	Accenture	-	21,16	64,90	-	-	ago/24
Saúde/farmacêutica	Pfizer	10,68	34,51	58,50	31%	18%	dez/23
	Ely Lilly	9,31	27,04	34,12	34%	27%	dez/23
	Novo Nordisk	31,03	196,50	232,26	16%	13%	dez/23

Fonte: Elaboração a partir de <https://stockanalysis.com> (acesso em: 20 out.2024)

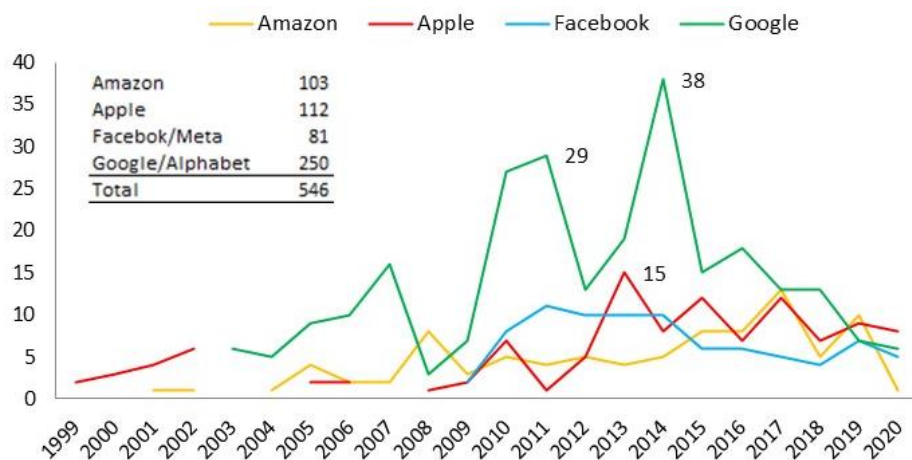
O ecossistema da IA preditiva contempla também firmas de menor porte, provedoras de conteúdo e complementadoras das plataformas digitais, muitas apenas com vínculos contratuais, embora algumas adquiridas pelas BigTechs, possivelmente quando se tornaram estratégicas para a captura de valor da inovação (Gráfico 2.2), ainda que sob forte escrutínio antitruste pela eliminação de concorrentes nascentes, tema tratado na literatura como *killer acquisitions* (Hemphill e Wu, 2020; Cunningham et al., 2021).

De fato, a aquisição de firmas foi importante fonte da inovação e diversificação das BigTechs⁵⁷ em negócios promissores. No caso do Google se destacam as aquisições do Youtube, em 2006, por US\$ 1,65 bilhões; do Android, por módicos US\$ 50 milhões, em 2013, da DeepMind, por US\$ 400 milhões, em 2014 e a aparentemente desastrosa aquisição da Motorola, em 2011, por US\$ 12,5 bilhões, vendida para a Lenovo três anos depois por pouco mais de 20% desse valor, mas que foi importante por fornecer um amplo portfólio de patentes

⁵⁷ As líderes das plataformas são hoje verdadeiros impérios, há anos listadas entre as maiores inovadoras do mundo (Forbes Top100; BCG Top 50), compostas por inúmeras firmas via fusão e aquisição. A Alphabet controla direta e indiretamente cerca de 400 firmas, a principal delas, o Google, com mais de 200 subsidiárias (Franeck, 2021).

essenciais de padrões⁵⁸. Produtos comerciais exitosos como o Google Search contaram com o conhecimento da Akwan Information Technologies, *spin-off* da Universidade Federal de Minas Gerais (UFMG), adquirida em 2005 (Exame, 2008). A Siri⁵⁹ (Apple) resultou da aquisição da Iris, desenvolvida pela *Viv Labs*, empresa depois adquirida pela Samsung em 2017. A Alexa (Amazon) brotou das aquisições da firma polonesa Ivona Software e da Evi Technologies, ambas em 2013. Recentemente, as estratégias de implementação da IA em setores como saúde estão dando origem a uma nova onda de aquisições, como a PillPack (2018) e a OneMedical (2022), pela Amazon, e a FitBit pela Alphabet (2020/21).

Gráfico 2.2 – Número de firmas adquiridas pelas BigTechs americanas, 1998-2020



Fonte: Elaboração da autora com base em dados de U.S. House Digital Markets Report (2020)

Nota: i) A Microsoft não foi incluída porque não fez parte da investigação antitruste do Congresso americano;

ii) mais de metade do total de aquisições (53.9%) ocorreram entre 2013 e 2019

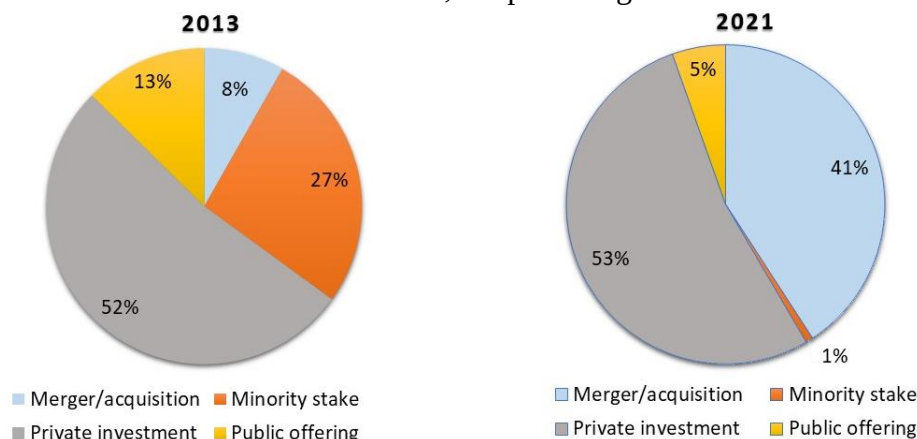
Em 2021, as F&A representaram 41% do total de US\$ 93,5 bilhões dos investimentos corporativos globais em IA (trinta e três pontos percentuais a mais do que em 2013) (Zhang et al., 2022a) (Gráfico 2.3). Antes de “aquisições matadoras”, essas iniciativas pareciam voltadas para a internalização de ativos complementares, fontes de gargalos potenciais, ou ponto de partida para apostas futuras, como as “Bets” do Google, muitas vezes para prover uma resposta rápida à competição entre plataformas (um dos três níveis de competição, ou “coopetição”

⁵⁸ A Alphabet não precisou envolver-se diretamente na fabricação de celulares, ao contrário da Apple, apoiando-se na obtenção desse ativo complementar por acordo com a Samsung, maior fabricante mundial de smartphones, que tem o Google Search como mecanismo de busca padrão dos smartphones, e onde disponibiliza o Google Play. Entretanto, crucial foi o Android, para competir no mercado de sistemas operacionais móveis. Primeiro desenvolveu e controlou a formação do padrão de fato (Android) por meio de um consórcio da indústria, a Open Handset Alliance, em uma abordagem “aberta” em relação à tecnologia (Amadeo, 2018; Herman, 2020) depois desenvolvendo um ecossistema em torno do Android, deixando partes abertas para devs., fechando partes que desenvolve e lança periodicamente com novas versões.

⁵⁹ As assistentes de voz virtuais como Siri, Alexa, Cortana (Microsoft) e Google Now, surgiram entre 2011-14, como produtos no mercado, ainda que a origem dos chatbots remonte a meados dos anos 1960, com o primeiro conhecido, o Eliza, criado por Joseph Weizenbaum, experimental e com muitas limitações. Ver <https://www.masswerk.at/elizabot/>. Acesso em: 15 out.2024.

apontados pela literatura), como no caso do sistema operacional Android, para fazer frente ao iOS da Apple e organizar o *Play Store*, que hoje supera as vendas da *App Store*.

Gráfico 2.3 - Investimento corporativo global anual em IA, por atividade de investimento, 2013 e 2021, em percentagem



Fonte: Our World in Data⁶⁰

2.3.4. Interação com a comunidade científica e acadêmica

As BigTechs mantêm interação crucial com a comunidade acadêmica e de pesquisa, em coautoria de artigos em publicações científicas e em conferências da área, e estão no centro de redes de pesquisa globais diversas, lidendo o *ranking* de publicações científicas (OECD.AI⁶¹; Maslej et al., 2024; Rikap, 2021; Rikap, Lundvall, 2021), somente suplantadas por universidades chinesas⁶².

Quase 40% dos trabalhos apresentados nas principais conferências de IA em 2020 tinham pelo menos um autor funcionário das empresas de tecnologia (Ahmed, Wahed, Thompson, 2023). A Alphabet/Google até pouco tempo era líder absoluta no *ranking* de publicações em duas prestigiadas conferências de pesquisa de IA, *International Conference on Machine Learning* (ICML) e *Neural Information Processing Systems* (NeurIPS) (Chuvpilo, 2020)⁶³. Microsoft, Amazon, Meta, IBM, também se destacam em publicações científicas. A

⁶⁰ Disponível em: ourworldindata.org. Acesso em 06 jul. 2022

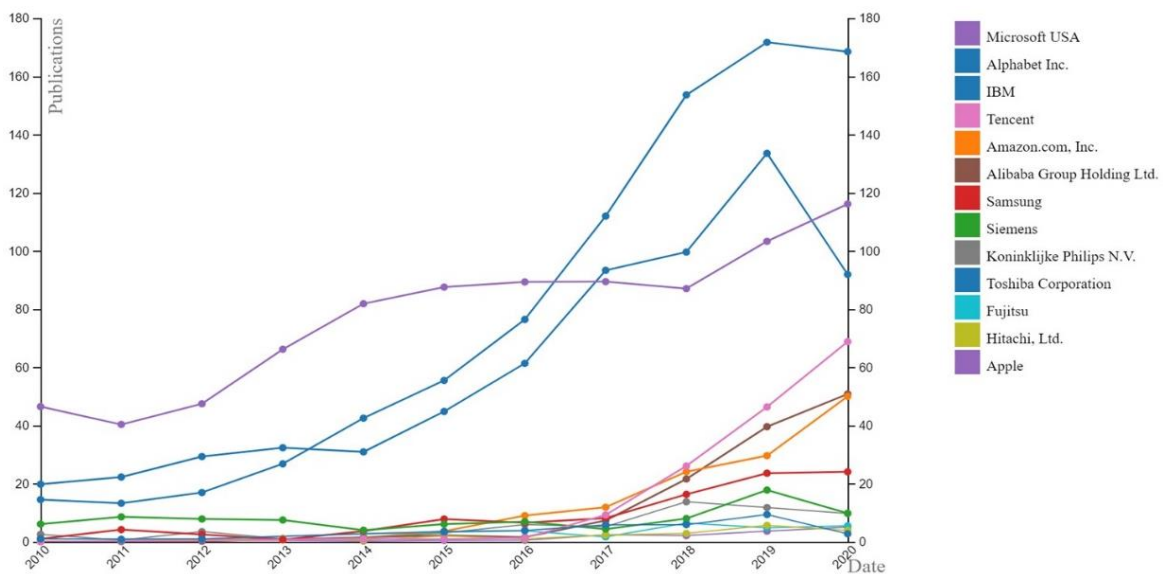
⁶¹ Disponível em <https://oecd.ai/en/>. Acesso em: 16 nov. 2022

⁶² Dados apresentados em WIPO (2024), para publicações relacionadas somente a IA generativa, entre 2014-2023, mostram uma realidade diferente, com predominância de organizações acadêmicas e Alphabet/Google como a única empresa entre as vinte organizações líderes em publicações científicas, além da primeira colocada no *ranking* de citações, com quase o dobro das citações da segunda colocada (Universidade de Berkeley) (WIPO, 2024a). O *ranking* de citações inclui também outras empresas – Meta, DeepMind, Nvidia, OpenAI, Microsoft e Twitter – entre as vinte organizações líderes.

⁶³ Em 2021, além de publicar mais de 750 artigos, quase 600 apresentados em grandes conferências, o Google Research patrocinou mais de 150 conferências e US\$ 59 milhões em financiamento a programas de pesquisa, parcerias e extensão universitária, com programas como o *Research Scholar Program* (Disponível em: <https://research.google/programs-and-events/research-scholar-program/>. Acesso em: 10 fev. 2023), o *Award for Inclusion Research Program* (Disponível em: <https://research.google/programs-and-events/award-for-inclusion-research-program/>. Acesso em: 10 fev. 2023), o *Visiting Researcher* (Disponível em: <https://research.google/outreach/visiting-researcher-program/>. Acesso em: 10 fev. 2023) e o *PhD Fellow*

Apple parece dissonante, talvez limitando seu ecossistema de inovação à plataforma de transação (Gráfico 2.4). Rikap (2023c) destaca a forte presença das BigTechs nos comitês organizadores das principais conferências, podenod influenciar resultados e moldar os rumos da pesquisa.

Gráfico 2.4 - Série temporal de publicações científicas de IA por instituições, da Scopus, 2010-2020



Fonte: OECD.AI.

Dados acumulados de patentes e publicações científicas apresentados por Rikap (2021), a partir das bases Derwent e Web of Science, da Clarivate, evidenciam a participação intensa das BigTechs americanas e chinesas em publicações científicas colaborativas (a maioria dos artigos em coautoria, exceto Microsoft), embora não compartilhem a titularidade de patentes (Tabela 2.3). De fato, para ter acesso a infraestrutura computacional e bases de dados essenciais para desenvolver modelos de IA, pesquisadores acadêmicos frequentemente fazem parcerias com funcionários de tecnologia das empresas, o que reforça o poder dessas últimas (Nix, Zakrzewski e De Vynck, 2024).

Tabela 2.3 – Publicações científicas e patentes, BigTechs selecionadas, 2014-2019

Firmas	Publicações científicas			Patentes		
	número de publicações	Publicações em co-autoria	%	Depositas e concedidas	Co-titularidade	%
Amazon	824	766	93%	10063	13	0,1%
Microsoft	17405	4025	23%	76109	160	0,2%
Google	6447	3397	53%	25538	65	0,3%
Tencent	643	366	57%	5462	13	0,2%
Alibaba	685	427	62%	3532	0	0,0%

Fonte: Rikap (2021)

A coautoria de artigos das BigTechs com a academia é expressiva, bem como dessas

firmas entre si, evidenciando a clara colaboração e compartilhamento do conhecimento entre elas (Tabela 2.4).

Tabela 2.4 – Top 10 organizações de coautores em publicações científicas - BigTechs e universidades, 2014-2019

Microsoft	Google	Amazon	Tencent	Alibaba
Univ of California	Univ. of California	Univ of California	CAS	Univ of S&T of China
Univ of Washington	Stanford Univ	Microsoft	Chinese Univ of Hong Kong	CAS
Univ of S&T of China	Microsoft	Univ. of Washington	Tsingua Univ	Zhejiang Univ
MIT	MIT	Google	Peking Univ	Tsingua Univ
Tsingua Univ	Harvard	IBM	Shanghai Jiao Tong Univ	Peking Univ
Univ of London	Carnegie Mellon Univ	Georgia Inst of Tech	Sun Yat-Sen Univ	Shanghai Jiao Tong Univ
Carnegie Mellon Univ	Univ of Illinois	Carnegie Mellon Univ	Zhejiang Univ	Microsoft
Google	Univ. of Washington	Univ of Texas	Harbin Inst of Tech	Fudan Univ
Stanford Univ	IBM	MIT	Beihang Univ	Nanyang Tech Univ
ETH Zurich	New York Univ	Indian Inst of Tech	Nanyang Tech Univ	Nanjing Univ

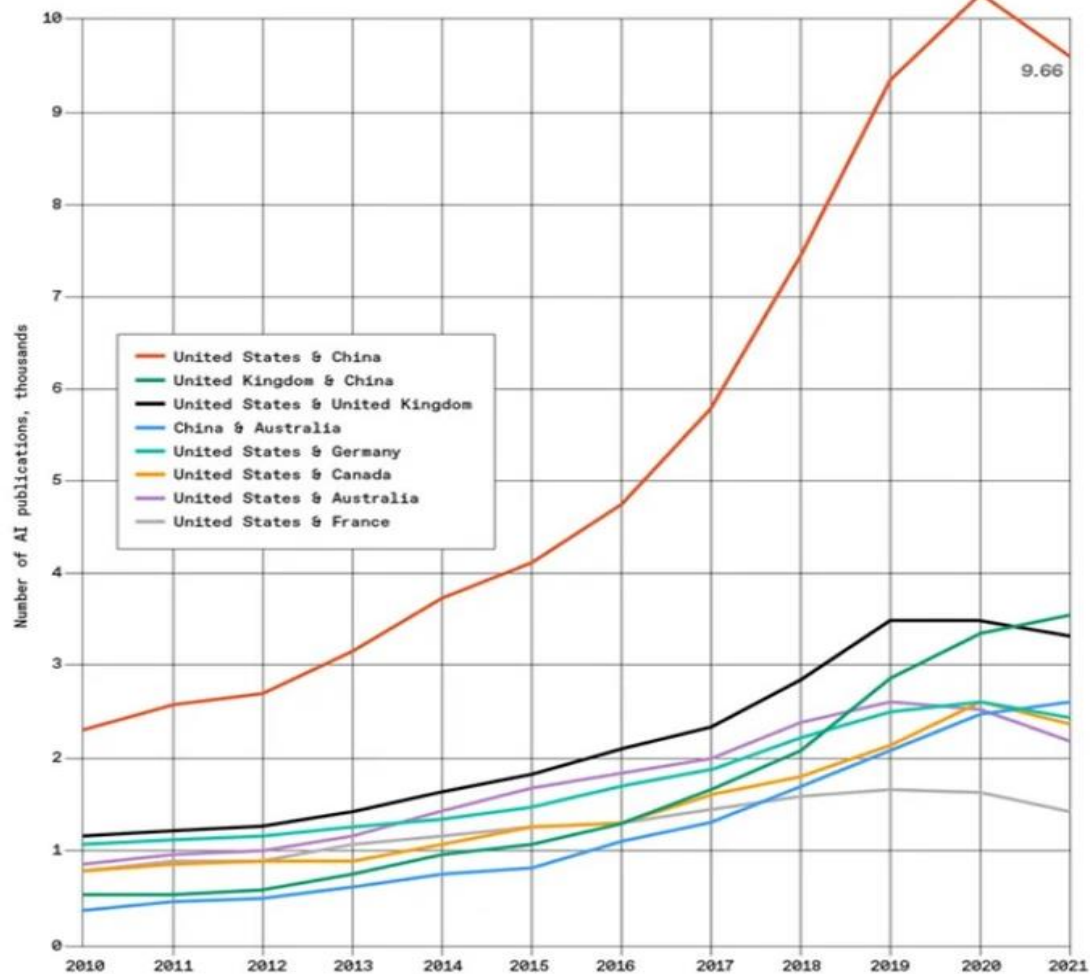
Fonte: Rikap, Lundvall (2021)

De fato, apesar da disputa tecnológica acirrada em IA entre EUA e China, a colaboração dos países em publicações de IA é intensa (Gráfico 2.5), envolvendo firmas BigTechs americanas e chinesas e universidades dos dois países.

A partir da construção de um mapa de redes dos coautores mais frequentes, Rikap (2023a) constata a importância das BigTechs americanas, em particular Microsoft, que ocuparia posição central nessas redes, tendo publicado em coautoria, entre 2012/20, com mais de 4.500 organizações espalhadas pelo mundo – e parcerias frequentes com quase 500 instituições, com as quais teria publicado dez ou mais vezes - ocupando posição central nos elos que conectam diferentes *players* ao redor do mundo (BigTechs americanas e chinesas, renomadas universidades americanas e universidades chinesas). Participou de 88% de artigos em coautoria entre 2012/21 e apenas 1% da cotitularidade em patentes (Rikap, 2023a:156).

De fato, as fronteiras entre indústria e academia são tênues. Importantes pesquisadores acadêmicos em IA, inclusive, ganhadores do Turing Award, são também cientistas-chefes das principais BigTechs. Geoffrey Hinton, pesquisador da Universidade de Toronto, por mais de uma década liderou o Google Brain, do qual se desligou no início de 2023 por preocupações sobre os riscos da nova IA generativa; Yann LeCun, pesquisador da New York University é também Vice-presidente e cientista-chefe de IA da Meta. Mais que casos isolados, esse papel duplo parece a tendência predominante desde os primórdios da IA, quando os nomes do emergente campo de pesquisa faziam parte dos quadros de empresas como IBM, AT&T Bell Labs, RAND Corporation do então ecossistema governamental da DARPA.

Gráfico 2.5 - Colaboração entre países em publicações de IA, 2010-2021



Fonte: Zhang et al. (2022a)

A interrelação entre academia e empresas parece ser a marca registrada da IA, objeto de críticas sobre o poder das BigTechs, privatização e monopólio do conhecimento (Rikap, 2021). De fato, a indústria controla os três principais ingredientes da moderna pesquisa em IA: grandes conjuntos de dados, pesquisadores altamente qualificados e, mais recentemente, poder computacional (Ahmed, Wahed e Thompson, 2023).

Uma tendência que deve se aprofundar considerando que a proporção crescente de novos PhDs em IA que se empregam na indústria, superando a média de outros campos da ciência, com riscos de reproduzir no quesito recursos humanos a assimetria gigantesca em capacidade computacional empregada pela academia e indústria no desenvolvimento de modelos de IA.

Bibliotecas abertas de dados e treinamento de modelos de IA

As plataformas de empresas privadas detêm o controle sobre mais dados do que qualquer outro *player*, gerados em tempo real em volume, variedade e velocidade

incomparavelmente superiores. Uma base pública, a ImageNet⁶⁴, lançada em 2009 por pesquisadores da Universidade de Stanford com imagens extraídas da web e rotuladas por meio de substantivos do banco de dados lexical sem vieses (*WordNet*), contém cerca de 14 milhões de imagens, classificadas em mais de 20.000 categorias. Esses números contrastam com os usuários das plataformas privadas, cujas bases de dados privadas não são abertas, com poucas exceções, como a *Open Images*, lançada em 2016 pelo Google, composta por conjunto de dados gratuitos com 9 milhões de imagens (Krasin e Duerig, 2016) e, em 2021, o “*More Inclusive Annotations for People*” (MIAP) da coleção *Open Images Extended* (mais completo e com menos vieses) (Schumann e Ricco, 2021)⁶⁵. Outras iniciativas, entretanto, parecem avançar na recente fase da IA generativa, como será visto.

As chamadas bibliotecas de treinamento de modelos de aprendizado de máquina de código aberto (*AI Open-source software libraries*), disponíveis para “devs.” ou qualquer usuário cadastrado, em versões *free*, gratuitas com funcionalidade limitada, ou “*freemium*”, com maiores funcionalidades, mas pagas (Pujol, 2010), começaram a ter um papel importante no ecossistema de inovação da IA.

Mesmo distinta do software tradicional, a IA é frequentemente expressa em programas de computador e sua implementação prática depende cada vez mais do software, inclusive *open-source software* (OSS) (Gráfico 2.6). Nesse sentido, como as implementações práticas de IA dependem de software, este se torna um importante meio de identificação de desenvolvimentos e aplicativos de IA.

Praticamente todas as BigTechs vêm adotando a IA de código aberto, embora o *open-source* para as líderes da IA, não seja um fim em si mesmo, mas um meio para um fim: criar um ecossistema em torno das bibliotecas de treinamento usadas como elemento de atração para devs., fabricantes de hardware, etc. Notavelmente, o TensorFlow⁶⁶ do Google, o Microsoft Azure⁶⁷, o Amazon SageMaker⁶⁸ do Amazon Web Services (AWS), o PyTorch⁶⁹ parcialmente desenvolvido pela Meta/Facebook e até incumbentes como IBM, com o Apache SystemML⁷⁰

⁶⁴ Disponível em: <https://www.image-net.org/>. Acesso em 23 dez. 2022.

⁶⁵ O Google menciona que essa abordagem foi lançada “(...) to help users discover new and useful datasets, wherever they are on the Web” (Google Research). Disponível em <https://datasetsearch.research.google.com/help>. Acesso em 2 jan. 2023. Mas compreende um conjunto limitado de imagens frente a totalidade dos dados variados coletados pelo Google que só no mecanismo de busca gera 2 trilhões de buscas por ano, que fornece um fluxo de dados em tempo real mais amplos e variados do que qualquer das bases gratuitas, públicas ou privadas – da mesma forma que as demais plataformas digitais.

⁶⁶ Disponível em: <https://www.tensorflow.org/>. Acesso em: 10 fev.2023

⁶⁷ Disponível em: <https://azure.microsoft.com/pt-br/>. Acesso em: 11 fev.2023

⁶⁸ Disponível em: <https://aws.amazon.com/sagemaker/>. Acesso em: 09 fev. 2023

⁶⁹ Disponível em: <https://pytorch.org/>. Acesso em: 10 fev. 2023

⁷⁰ Disponível em: https://researcher.watson.ibm.com/researcher/view_group.php?id=3174. Acesso em: 11 fev.2023

são plataformas abertas que disponibilizam recursos diversos. Embora os usuários tenham acesso a modelos de aprendizado de máquina de código aberto para treinar seus próprios modelos e dados, as ferramentas podem camuflar estratégias de coleta de dados e soluções de “devs.” que acabam por auxiliar no aperfeiçoamento dos modelos proprietários da biblioteca, colaborando, de forma gratuita, para futuras invenções, inclusive protegidas por PI. Com isso, as grandes empresas compartilham parte do conhecimento, mas mantendo sigilo sobre detalhes técnicos dos modelos, executados em infraestrutura computacional exclusiva (Metz, 2015).

Gráfico 2.6 - Inteligência Artificial, uso como OSS



Fonte: Baruffaldi et al. (2020)

Funcionam, assim, como uma fonte complementar de conhecimento, até recentemente pouco conhecida, fundamental para interação com “devs.” individuais e startups. Alguns autores consideram essas estruturas de código aberto como “rampas de acesso” para as infraestruturas proprietárias de grandes empresas de IA (Dyer-Witthford et al., 2019) e, por outro lado, bases para construção de comunidades de “devs.” treinados com suas ferramentas (Srnicsek, 2020) como parte das estratégias empresariais para consolidação de suas tecnologias como *design* tecnológico dominante pelos efeitos de rede (Tomic, 2018). Certamente, são também um funil de recursos humanos conhecedores das ferramentas das empresas, sinalizando talentos e *startups* promissores, inclusive para futura contratação e aquisição pelas BigTechs.

Desde meados da década passada é significativa a expansão do número de usuários de bibliotecas de software de IA de código aberto como o GitHub⁷¹, com destaque para o

⁷¹ O GitHub é a maior plataforma de hospedagem de OSS e arquivos de controle de versão, permitindo que devs cadastrados ao redor do mundo divulguem trabalhos e contribuam em projetos privados ou OSS. Não é a única plataforma online de código aberto, mas a maior do mundo, com 100 milhões de usuários em 2023 (majoritariamente devs. de fora dos EUA) e mais de 420 repositórios. Desde 2014 ocorre um rápido crescimento nos repositórios relacionados à IA. Nos três anos seguintes, o software de código aberto de IA cresceu cerca de três vezes mais do que o restante dos OSS (Baruffaldi et al. 2020:32). Em 2018, foi adquirida pela Microsoft

TensorFlow, seguido pelo OpenCV, Keras, PyTorch e outras. Simultaneamente, as BigTechs passaram a ter participação direta nos principais repositórios, em particular no GitHub, adquirido pela Microsoft em 2018 (Lardinois e Lunden, 2018), seguindo a estratégia do Google que comprou o australiano Kaggle um ano antes (Li, 2017). O franco-americano HuggingFace é outro destaque. Em maio de 2020, a Meta AI disponibilizou um modelo de idioma com 175 bilhões de parâmetros treinados, o *Open Pretrained Transformer* (OPT-175B) (Zhang et. al., 2022), além do código para treinamento/uso, bem como o conjunto de dados, buscando engajamento da comunidade - pesquisadores do governo, sociedade civil e academia, ao redor do mundo, sob licença não comercial para uso em pesquisa. Esses repositórios são peças importantes no desenvolvimento de modelos da IA, com as pesquisas, metodologias e bases de dados compartilhadas nos repositórios por devs de diversas partes do mundo.

2.3.5. Supercomputação e computação na nuvem

Elemento central para as invenções/ inovações baseadas em sistemas de redes neurais profundas é a capacidade computacional⁷², principalmente para treinamento dos modelos e cuja evolução foi extraordinária a partir dos anos 2000, como foi visto na figura 2.1, possibilitando cálculos complexos a partir de funções não lineares com crescente número de parâmetros e representações digitais de dados.

Alguns supercomputadores começaram a ser desenvolvidos para uso exclusivo em modelos de IA, como o Selene, da NVIDIA, operando em 2020, voltado para veículos autônomos e aprendizado profundo de uso geral. A Mobileye, Intel, Qualcomm e Samsung estão desenvolvendo *chips* incorporando ConvNets para permitir aplicativos de visão computacional em tempo real em *smartphones*, robôs e veículos autônomos. A Microsoft, em parceria com a OpenAI, entrou em supercomputação, sem um dispositivo físico, na sua plataforma na nuvem Azure, dedicada a sistemas de IA amplos. A Meta, por meio do seu AI Research Super Cluster (RSC), em parceria com NVIDIA, Pure Storage e Penguin, visando desenvolver um dos maiores supercomputadores para treinamento conjunto de dados de imagem, vídeo e texto para metaverso.

A crescente utilização da computação na nuvem (*cloud computing*)⁷³ permite a

(US\$ 7,5 bilhões) que, no mesmo ano liderou o número de projetos de código aberto no GitHub (com 19.000 devs), Facebook (10.000) e Google (9.300) (Rikap, 2020).

⁷² Sobre os supercomputadores mais rápidos do mundo ver levantamento anual da Top500. Disponível em: <https://www.top500.org/lists/top500/2024/06/>. Acesso em 27 set. 2024.

⁷³ Sobre os conceitos relevantes sobre computação na nuvem, ver Mell e Grance (2011). Cabe mencionar que existem diferentes modelos de plataformas na nuvem: a) nuvem privada, em que a infraestrutura em nuvem é fornecida para uso exclusivo por uma única organização composta por vários usuários, como unidades de negócios, podendo pertencer, ser gerenciada e operada pela organização, por terceiros ou por alguma combinação deles, e

terceirização mais ampla de serviços de computação, desde infraestrutura física básica (*Infrastructure as a Service - IaaS*), gerenciamento da infraestrutura subjacente para firmas (planejamento de capacidade, manutenção, correção etc.) (*Platform as a Service - PaaS*) e “*Software as a Service*” (SaaS). Por conseguinte, está se tornando um mercado mundial de um subconjunto das BigTechs Amazon (AWS Cloud)⁷⁴, Microsoft Azure⁷⁵ e Google Cloud⁷⁶. Essas infraestruturas constituem importante fonte de receitas – o AWS é atualmente o negócio mais lucrativo da Amazon – e prometem se tornar o mais importante ativo especializado integrado às suas operações de prestação de serviços. As chinesas também possuem poderosas infraestruturas na nuvem, como Alibaba Cloud⁷⁷, Tencent Cloud⁷⁸, a Huawei Cloud,⁷⁹ a Baidu Cloud⁸⁰. Incumbentes do setor de tecnologia como IBM, Oracle, Qualcomm, Oracle, HP, Dell e outras estão reposicionando suas estratégias por meio da infraestrutura na nuvem (The Economist 2021a; The Economist 2021b), além de provedores de sistemas para clientes, como Salesforce, e até a Nvidia, projetista de chips, com seu DGX Cloud.

A infraestrutura de computação na nuvem é o fornecimento ao usuário de capacidade de processamento, armazenamento, redes e outros recursos computacionais fundamentais na nuvem possibilitando que seja capaz de implantar e executar modelos que incorporem seus aplicativos (Mell e Grance, 2011:3). Esta infraestrutura é não controlada pelo usuário, mas pelos provedores, ainda que lhe assegure acesso a sistemas operacionais, capacidade de armazenamento e aplicativos implantados, além de componentes de rede selecionados. O mercado de mundial de infraestrutura na nuvem anualizado é estimado em US\$ 300 bilhões em 2024, que cresce exponencialmente, contemplando PaaS, IaaS e hospedagem de serviços de nuvem privada, com dados da Sinergy Group. Oito empresas respondem por 80% desse mercado, com dois terços do total nas mãos da Amazon, Microsoft e Google, tornando-se os efetivos **Cloud Service Providers (CSPs)** (Gráfico 2.7).

existir dentro ou fora das instalações; b) nuvem pública, onde a infraestrutura é fornecida para uso aberto pelo público, pertencendo, gerenciada e operada por organização empresarial, acadêmica ou governamental, ou alguma combinação, nas instalações do provedor de nuvem; c) nuvem comunitária, quando a infraestrutura é fornecida para uso exclusivo por uma comunidade específica de usuários que compartilham sobre requisitos de segurança, conformidade, missão, comuns, gerida e operada por uma ou mais organizações da comunidade, um terceiro, ou alguma combinação, podendo existir dentro ou fora das instalações; e d) nuvem híbrida, na qual a infraestrutura é uma composição de duas ou mais infraestruturas de nuvem distintas (privada, comunitária ou pública), para entidades unidas por tecnologia padronizada ou proprietária, permitindo portabilidade de dados e aplicativos.

⁷⁴ Disponível em: <https://aws.amazon.com/>. Acesso em 21 jan. 2023

⁷⁵ Disponível em: <https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-cloud-computing/> Acesso em 21 jan. 2023

⁷⁶ Disponível em: <https://cloud.google.com/>. Acesso em 20 jan. 2023

⁷⁷ Disponível em: <https://www.alibabacloud.com/>. Acesso em 10 ago. 2024

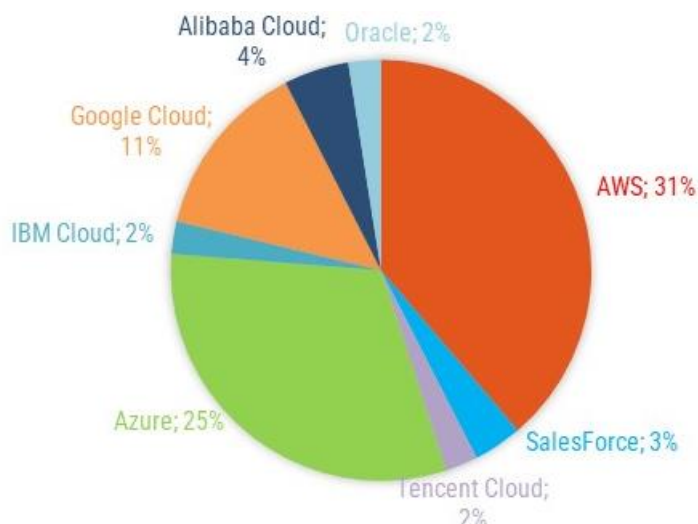
⁷⁸ Disponível em: <https://www.tencentcloud.com/>. Acesso em 10 ago. 2024

⁷⁹ Disponível em: <https://www.huaweicloud.com/intl/en-us/>. Acesso em 10 ago. 2024

⁸⁰ Disponível em: <https://intl.cloud.baidu.com/>. Acesso em 10 ago. 2024

A internalização de recursos computacionais ao ambiente das plataformas tornou-se um ativo crucial para a inovação em IA por possibilitar acesso a grande capacidade de processamento/armazenamento de dados, permitindo até mesmo que a IA passe a ser fornecida como serviço (*Artificial Intelligence as a Service - AIaaS*), alimentando suposições de que as técnicas de aprendizado de máquina se tornariam *commodities* da nuvem (Taddy, 2018).

Gráfico 2.7 – Participação de mercado mundial dos principais provedores de serviços de infraestrutura em nuvem, Q1 2024



Fonte: Synergy Research Group apud Richter (2024)

As antigas abordagens de aprendizado de máquina exigiam centralização de dados de treinamento em máquina ou datacenters ou, mais recentemente, nas robustas infraestruturas na nuvem, ao passo que novos modelos começam a ser treinados a partir da interação com o usuário em dispositivos móveis, introduzindo novas abordagens (*federated learning*) onde telefones celulares aprendem em modelos de previsão compartilhados, com parte dos dados de treinamento mantidos no dispositivo (*edge computing*) (McMahan e Ramage 2017; Wang et al., 2019), empregadas em técnicas como *Federated Reconstruction* (Singhal, 2021).

A computação na nuvem pode representar um novo estágio das plataformas digitais, com amplos e crescentes conjuntos de atividades prestadas na nuvem como serviços, tais como IaaS, PaaS, SaaS e, agora também AIaaS e *Data as a Service* (DaaS) e *Models as a Service* (MaaS)⁸¹. Firms de tecnologia, incumbentes e BigTechs, estão se transformando em CPSs, a partir de pesados investimentos físicos e em P&D de plataformas com um conjunto robusto de serviços em nuvem, executados sobre dados e IA, além de *Internet of Things* (IoT). Serviços fornecidos na nuvem a custo marginal próximo de zero, cujos códigos não são disponibilizados aos clientes (são bloqueados), possibilitam lucros expressivos aos provedores. Rikap (2022)

⁸¹ A ZhipuAI, spin-off da Tsinghua University, dedica-se a criar uma nova geração de modelos universais de inteligência cognitiva e a introduzir o conceito de negócios de Modelo como Serviço (MaaS).

resume: “*Since the same lines of code can be simultaneously used by many, the reproduction costs of selling artificial-intelligence (AI) algorithms as cloud services tend to zero*”.

Serviços em nuvem híbrida começam a ser providos de forma acelerada por meio de *kubernetes*⁸² e *containers*⁸³, pelas BigTechs, firmas incumbentes ou mesmo nascentes (Amazon Kubernetes, Microsoft Azure Kubernetes Services - AKS, Google Kubernetes, Dell Kubernetes, Alibaba ACK, Tencent Kubernetes Engine, Huawei CCE, IBM Kubernetes, Kyndryl Kubernetes, etc.). Prometem ser um ativo essencial na fase de implementação da IA, em modelos de inovação aberta, que envolvam parcerias entre firmas de plataforma e incumbentes dos setores *downstream*, coordenadas por meio de DPIs e relações contratuais complementares sobre AIaaS e DaaS.

2.4. A liderança na inovação da IA: ecossistemas e modelo de negócios

Assim, as inovações relacionadas à IA preditiva resultaram de um extenso ecossistema de plataformas digitais, fundamental para a compreensão da criação e captura de valor dessa poderosa tecnologia. Os elementos chaves desse ecossistema relacionam-se à operação em plataformas digitais híbridas, mas não só. O forte ritmo da inovação das BigTechs resulta da sua alta intensidade de P&D, da rede de interações com a pesquisa acadêmica e de um posicionamento central na relação com “devs.” e startups atrelados aos ambientes das plataformas digitais e/ou bibliotecas *free/freemium* de treinamento de modelos de aprendizado e bases de dados, consolidada no negócio de fornecimento de serviços na nuvem.

Tal arranjo institucional peculiar engloba ainda a aquisição direta de empresas menores responsáveis por suprir partes estratégicas do conhecimento e a crescente atuação das BigTechs como verdadeiros capitalistas de risco, que funcionam como fonte estendida de inovação e “radar” da identificação de negócios promissores e gargalos potenciais em ativos complementares. O crescente recurso a estratégias de inovação aberta pelos principais *players* é combinada com o uso de mecanismos formais de DPI para captura de valor da inovação em IA, uma vez que versões mais avançadas dos modelos e o conjunto completo dos dados não parecem ser compartilhados.

A abordagem de criação de valor por extenso ecossistema de inovação da IA é

⁸² Originalmente projetado pelo Google, hoje mantido pela *Cloud Native Computing Foundation*, Kubernetes é um sistema de orquestração de contêineres *open-source* que automatiza a implantação, o dimensionamento e a gestão de aplicações em contêineres.

⁸³ Contêiner é um único pacote de software que agrupa o código de um aplicativo junto com todos os elementos do ambiente de computação necessários para executá-lo como códigos de configuração. A containerização é uma forma de empacotar tudo em uma única unidade que pode ser movida entre ambientes computacionais. Essa abordagem evita problemas que devs geralmente encontram quando pegam o código escrito para um sistema operacional ou ambiente de hospedagem e o transferem para outro (Bawcom et al., 2022).

exemplificada com o caso da Alphabet/Google, que mostra os esforços internos de P&D e apostas (Bets) estratégicas para o grupo, as plataformas de transação, que engloba inúmeras empresas do grupo, a biblioteca Tensorflow, a intensa interação com a academia e participação em repositórios *open-source*, investimentos em F&A e em fundos próprios de capital de risco - Google Ventures (GV)⁸⁴, Capital G⁸⁵ e o recente Gradient Ventures⁸⁶ com foco específico em IA – e o crescente negócio em computação na nuvem (Figura 2.6). Outras BigTechs possuem fundos próprios de venture capital, como Microsoft, Amazon, Tencent, Baidu, Alibaba, JD.com e até mesmo a Nvidia, líder no *design* dos chips de IA⁸⁷.

Figura 2.6 – Representação gráfica das fontes de dados, inovação e capacidades dinâmicas da Alphabet/Google



Fonte: Elaboração própria a partir de pesquisa em vários Websites da empresa e na literatura.

Nesse sentido, a abordagem de inovação da IA é baseada em um extenso modelo que combina inovação aberta⁸⁸, fontes internas e externas, e se traduz em um modelo de apropriabilidade complexo e híbrido, combinando amplos mecanismos formais de DPIs sobrepostos, controle sobre ativos complementares e que convivem harmoniosamente com o regime *open-source*, como será visto.. No entanto, parece experimentar tendências divergentes

⁸⁴ Disponível em: <https://www.gv.com/>. Acesso em: 3 fev. 2024

⁸⁵ Disponível em: <https://www.capitalg.com/>. Acesso em: 29 jan. 2024

⁸⁶ Disponível em: <https://www.gradient.com/>. Acesso em: 2 mar. 2024

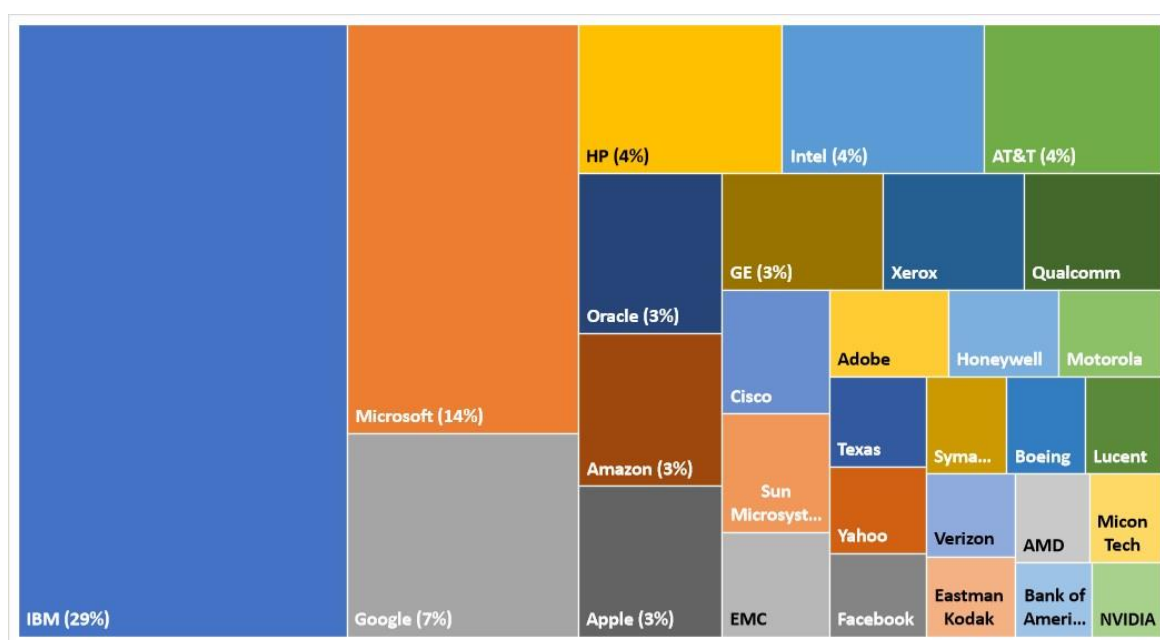
⁸⁷ Esses investimentos em startups rivalizam ou mesmo superam capitalistas de risco tradicionais como *Sequoia Capital*, *Andreessen Horowitz* e *SoftBank*. Em seu conjunto, os investimentos da Microsoft, Alphabet e Amazon com empresas emergentes da IA generativa representaram dois terços dos US\$ 27 bilhões por elas captados em 2023 (Hammond, 2023).

⁸⁸ O website de patentes do Google contempla módulos de Programa de Patentes, Patentes de Código Aberto e Envio de Oportunidade de Patente, Disponível em: <https://www.google.com/patents/opnpledge/>. Acesso em 7 jun.2023.

e aparentemente contraditórias na nova fase da IA generativa, a partir de estratégias empresariais distintas e até mesmo destoantes.

O quinteto das BigTechs americanas Apple, Amazon, Alphabet/Google, Meta/Facebook e Microsoft parecia representar as efetivas vencedoras da era das redes neurais profundas, estruturando complexos ecossistemas de inovação e respondendo pelos exitosos sistemas da IA preditiva que chegaram ao mercado, principalmente em produtos/serviços de plataformas digitais. Até poucos anos atrás, muitas dessas firmas lideravam a inovação da IA, como mostram os números expressivos de pedidos de patentes depositados no USPTO (Gráfico 2.8).

Gráfico 2.8 - Trinta principais proprietários de patentes de IA concedidos pelo USPTO, em percentagem, 1976-2018



Fonte: PatentsView⁸⁹

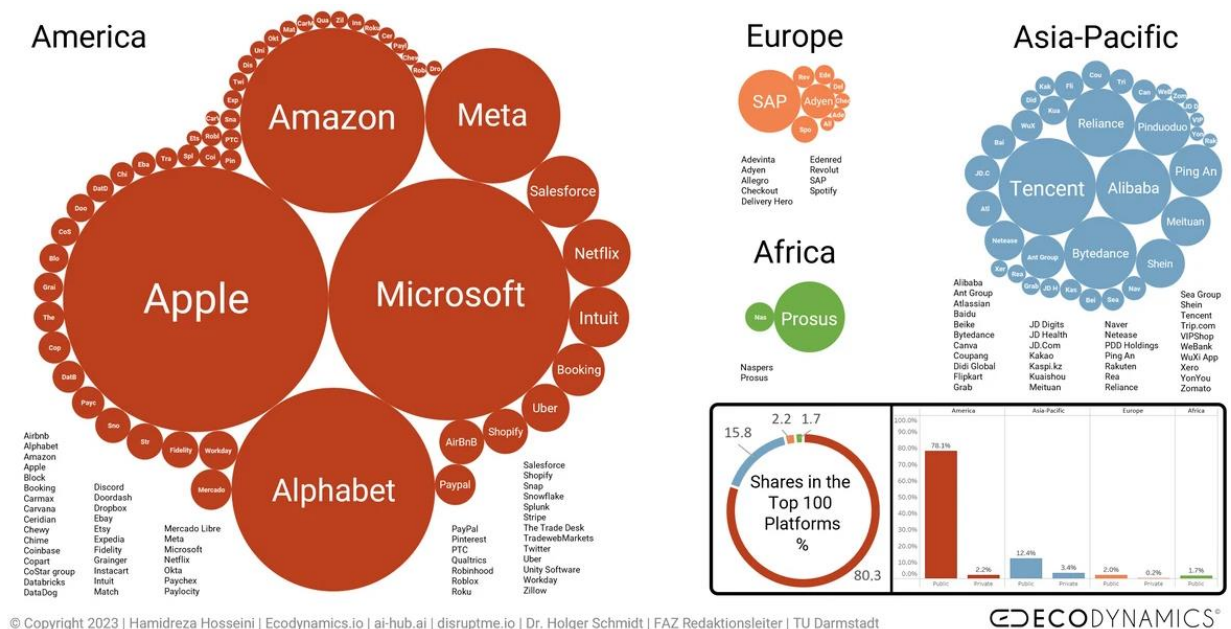
A inovação em IA pelas BigTechs, que emerge desses ecossistemas é sintetizada por Rikap (2021) e Rikap e Lundvall (2021) por meio da metáfora de um quebra-cabeças, onde as empresas líderes incorporam capacidades externas que fornecem parte das peças, mas onde a peça central – conjuntos de dados treinados e a parte mais importante dos códigos dos modelos de IA – é por elas mantida em sigilo, possibilitando a captura da maior parcela do valor da inovação.

De fato, as firmas que comandam plataformas digitais apresentaram ascensão meteórica desde a década passada. As BigTechs americanas somaram, em 2022, US\$ 6,9 trilhões em

⁸⁹ A PatentsView é uma plataforma online gratuita sobre dados de patentes do USPTO, que utiliza vários algoritmos de desambiguação, com métricas de classificação, similaridade e clusterização para análise de dados de patentes <http://www.patentsview.org/>. Acesso em 04 jan. 2023

capitalização de mercado e uma receita agregada de quase US\$ 1,5 trilhão, superando o PIB de muitos países, mesmo com a retração frente ao ano anterior. As maiores empresas do mundo são aquelas que empregam modelos de negócios de plataforma digital, regionalmente concentradas nos continentes asiático e americano, sobretudo nos EUA e na China, com um número muito menor na Europa (Figura 2.7). A concentração de poucas firmas em capitalização de mercado é crescente, ampliando o *gap* entre a maior e a menor firma entre as 100 listadas (que era 5,4 vezes em 2010 e passou a quase 16 vezes em 2021) (PWC, 2021). Não por acaso, a liderança global da inovação da IA também se concentrou nos EUA e China, como mostrado em estudo da OMPI, que identificou 340 mil famílias de pedidos de patentes e mais 1,6 milhão de artigos científicos relacionados à IA entre 1960-2017 (Gráfico 2.9).

Figura 2.7 – Cem maiores plataformas digitais por capitalização de mercado, agosto 2023

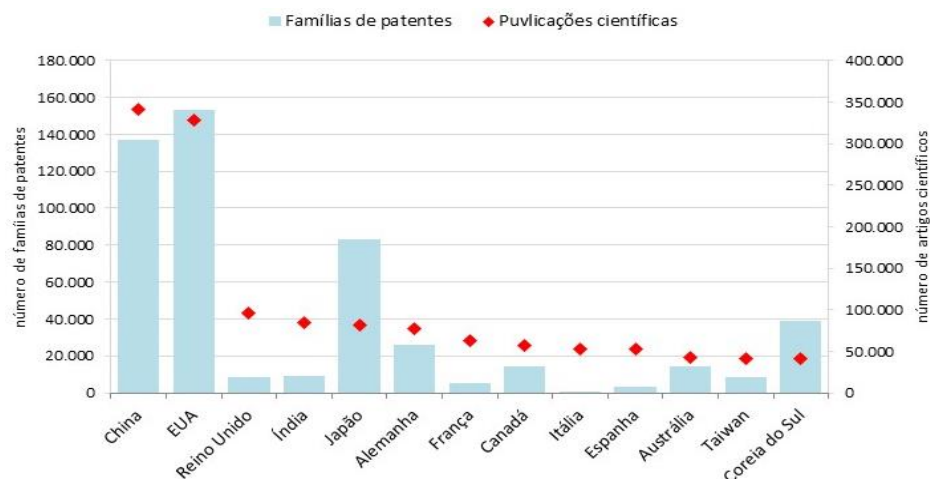


© Copyright 2023 | Hamidreza Hosseini | Ecodynamics.io | ai-hub.ai | disruptme.io | Dr. Holger Schmidt | FAZ Redaktionsleiter | TU Darmstadt

ECODYNAMICS®

Fonte: Ecodynamics

Gráfico 2.9 – Número de famílias de patentes de IA e publicações científicas, países selecionados, 1960-2017



Fonte: WIPO (2019)

CAPÍTULO 3 – A INTELIGÊNCIA ARTIFICIAL GENERATIVA: AVANÇOS EM REDES NEURAIS E A IMPORTÂNCIA CRESCENTE DE ATIVOS TANGÍVEIS

Os sistemas computacionais de redes neurais profundas, baseados em modelos que dependem de amplos conjuntos de dados e de avanços computacionais para alcançar precisão e acuracidade em diferentes aplicações funcionais, estiveram até recentemente restritos a tarefas isoladas.

Há alguns poucos anos, entretanto, surgiu uma nova classe de modelos de fronteira da chamada IA generativa, inicialmente adstritos à comunidade de devs. até começarem a ser disponibilizados no mercado. Isso ofuscava a tendência de algumas empresas líderes integrarem verticalmente ativos de hardware, tecnologias complementares cada vez mais essenciais para o avanço dos modelos. O ano de 2023 foi apontado como um ponto de inflexão e de mudança de paradigma da IA, com uma explosão de modelos do tipo generativo, desenvolvidos por empresas incumbentes e empresas emergentes.

Abrangem uma nova geração de modelos de aprendizado profundo com habilidade para criar conteúdo de texto, imagem, vídeo, áudio e até mesmo código de programação, em alguns casos indistinguíveis dos criados por um humano, a partir de demandas (consultas) na forma de texto e, mais recentemente, também por voz, em linguagem natural. Não exigem conhecimento de programação, ampliando assim a interação humana com as máquinas. Esses modelos possibilitam sistemas com aplicações muito mais amplas em tarefas distintas, cuja combinação será exigida para a implementação das poderosas tecnologias da IA em múltiplos setores econômicos a jusante.

De fato, o ano de 2023 pareceu singular em vários aspectos. Primeiro, pela enorme popularidade alcançada pelo aplicativo ChatGPT (da novata OpenAI), logo seguido pelas empresas líderes, além dos surpreendentes modelos de difusão lançados por empresas até então desconhecidas, como Midjourney e Stability AI.

Segundo, foi marcado pela proliferação desenfreada de modelos (e inúmeros aplicativos) que despontam diariamente em rápida disseminação e atualização, muitos em regime *open-source*, o que agregou novos *players* ao contexto antes praticamente restrito ao seleto grupo das BigTechs, exigindo uma maior compreensão do que são modelos abertos.

Terceiro, simultaneamente ao entusiasmo provocado pelas maiores funcionalidades dos novos modelos, houve uma ampliação acentuada de inquietações em temas como segurança, riscos e danos potenciais para a sociedade, alastrando uma onda de iniciativas regulatórias nos planos nacional, regional e global, paralelamente ao acirramento das disputas geopolíticas que parecem reproduzir um padrão de nova guerra-fria. É preciso reconhecer o papel que a

regulação ou qualquer outra forma de intervenção pública nos mercados representa, como mostrou a história na indústria de TICs sobre o padrão de competição e arquitetura da indústria, com vistas a vislumbrar possíveis efeitos sobre a evolução futura da IA, seus *players* e a apropriabilidade da inovação.

Quarto, como os novos modelos aumentam exponencialmente em escala, construídos sob expressivo número de parâmetros e treinados com colossal quantidade de dados, passaram a demandar extraordinária e crescente capacidade computacional. É justamente nesses recursos que parece residir a mais profunda dimensão da mudança de paradigma da IA, tratada no presente capítulo que analisa a importância do hardware – aqui chamado de ativo tangível da IA – que antes era timidamente explorado pelos estudos sobre IA, à exceção dos voltados para segurança nacional e cibernética (NSCAI, 2021; SCSP, 2023; Bresnick, Luong e Curlee, 2024).

Isso demandou a compreensão das razões que explicam o ingresso de novos *players*, da diferença entre desenvolvimentos a partir de modelos pré-treinados e de modelos efetivamente inovadores da IA, entre outros temas intrigantes. O capítulo procura abordar essas questões emergentes observando as características dos modelos e as distintas estratégias dos *players* da IA generativa, a partir de dados empresariais de produção da IA e dos ativos físicos que se tornaram cada vez mais cruciais na corrida por novos modelos e para a liderança no setor. Esses ativos são representados pelos recursos computacionais e sua cadeia de valor digital, incluindo unidades de processamento gráfico (GPUs), *Tensor Processing Units* (TPUs), chips para IA, softwares e equipamentos ultra especializados e serviços de computação na nuvem, quase monopolizados organizacional e geograficamente embora pareçam vitais para criação e captura de valor da inovação e definição de possíveis vencedores da nova fase da IA.

A despeito da dificuldade de teorizar sobre um fenômeno tão recente, cuja evolução ainda está em curso, com todas as incertezas decorrentes, procura-se avançar o debate sobre a inovação e a apropriabilidade que parecem ir além de meros DPI e diferenciados na IA generativa. A contribuição do capítulo é, portanto, aprofundar aspectos técnicos da IA e seu ecossistema, buscando um melhor entendimento sobre essa promissora tecnologia a partir da identificação dos ativos complementares essenciais não somente para captura de valor, como para sua criação. Nesse sentido, contribui para a compreensão das vantagens competitivas das inovadoras líderes em IA, complementando análises de plataformas vencedoras (Yoffie, Gawer e Cusumano, 2019).

A chegada da IA generativa trouxe consigo novos desafios econômicos, para a inovação e para o progresso técnico, que se somam a temas relacionados ao simples uso de sistemas de

IA, de natureza ética e de segurança, que mobilizam a mídia e estudos técnicos e começam a nortear as recentes políticas públicas em diversas partes do mundo.

3.1. O passo adiante com a IA generativa

Nos últimos cinco anos as fronteiras dos sistemas de aprendizado profundo avançaram em velocidade e escopo por uma série de avanços nas arquiteturas, no tamanho e no treinamento dos modelos e na capacidade computacional empregada.

O termo IA generativa – genérico, impreciso, mas popularizado - corresponde à classe de modelos de aprendizado profundo generativo que se concentra na criação de novos conteúdos ou dados semelhantes ao conjunto de dados com que foram treinados (Suresh, 2023). Isso implica que o papel dos dados se modifica nos modelos generativos, continuando a constituir uma entrada e, simultaneamente, também uma saída, criando um elemento circular na cadeia de valor desses modelos, com a entrada gerando uma saída que é posteriormente reciclada para gerar outra saída (Tyagi, 2024:6).

Esses modelos foram apontados como divisor de águas por representaram uma mudança significativa em relação aos modelos tradicionais, abrangendo diferentes domínios, mas com destaque para os grandes modelos de linguagem (LLM, sigla de *Large Language Models*) (Ganguli et al., 2022), modelos multimodais (Xu, Zhu e Clifton, 2023) e modelos de fundação (Bommasani et al., 2022).

Essas diferentes categorias de modelos da IA generativa se diferenciam em escopo. Os LLMs, os mais evidentes da nova onda da IA, avançaram espetacularmente nos anos recentes, dando origem a aplicativos conversacionais. São modelos que preveem qual palavra (na realidade, *tokens*) vem a seguir em uma sequência a partir de probabilidades em vetores de dados. Seu valor é devido, em parte, a empregar um novo tipo de rede neural (Transformers) que supera as limitações de antigas redes neurais, como as *Recurrent Neural Networks* (RNN), como as *Long-Short Term Memory* (LSTM) no campo de linguagem conhecido como processamento de linguagem natural (NLP, sigla de *natural language processing*).

Na realidade, a nova onda da IA teve início com outros modelos unimodais geradores de conteúdo de imagem, baseados em outros tipos de redes neurais, como as *Generative Adversarial Networks* (GANs), que suplantaram redes neurais artificiais, como ConvNets, que dominavam o campo de imagem na fase da IA preditiva ou classificativa.

Modelos multimodais, por seu turno, são categorias de aprendizado de máquina treinados com conjuntos de dados variados, pessoais e não pessoais, com capacidade de aprender uma ampla gama de tarefas e habilidades. Podem processar e fazer inferências em

diferentes domínios pelo treinamento com dados de diferentes modais, apresentando melhoria no desempenho e compreensão contextual, transcendendo limitações usuais dos LLMs treinados somente com dados unimodais linguísticos, ao incorporar novos modais de dados ao treinamento.

Em função da multimodalidade⁹⁰, que é essencial para a “inteligência”, os modelos apresentaram funcionalidade mais ampla e a geração de conteúdo diverso a partir de instrução textual, uma vez que aprendem padrões e a estrutura de dados de treinamento de diferentes domínios de entrada, gerando saídas de novos dados com características semelhantes, combinando dados visuais com dados de texto ou áudio (Singh, 2023)⁹¹. Com isso, respondem perguntas e criam legendas descritivas de imagens, combinando visão computacional e processamento de linguagem natural, tradução de informações visuais etc. A aprendizagem de máquina multimodal (MML) é uma abordagem geral para a construção de modelos de IA que podem extrair e relacionar informações de dados multimodais. Não constituem um conjunto homogêneo, alguns são modelos discriminativos, enquanto outros são generativos.

Modelos de fundação (“*foundation models*”) são definidos como “qualquer modelo treinado em dados amplos (geralmente usando auto supervisão em escala) que pode ser adaptado (ajustado) para uma ampla gama de tarefas *downstream*” (Bommasani et al., 2022:3). São modelos promissores de fronteira, cuja nomenclatura visa seu caráter central, mas incompleto, de tamanho expressivamente maior, derivados de enormes *corpora* de imagens, texto, vídeo e, mais recentemente, código, passíveis de flexibilidade, generalização e adaptação a ampla gama de tarefas complexas – razão pela qual foram considerados com maiores riscos potenciais e necessidade regulatória, como será visto.

⁹⁰ Um problema de pesquisa é caracterizado como multimodal quando inclui múltiplas modalidades (linguagem, visual, auditiva etc.) e com isso é capaz de realizar distintas tarefas (tradução, reconhecimento de imagem, reconhecimento de fala etc.). O mecanismo fundamental da percepção sensorial humana é a capacidade de trata conjuntamente múltiplas modalidades de dados, a fim de possibilitar envolvimento mais realista com o mundo, com cada modalidade servindo como uma fonte de informação distinta, caracterizada por diferentes propriedades estatísticas. Uma imagem tem aparência visual de uma cena através de milhares de *pixels*, enquanto um texto descreve uma imagem com uma frase usando palavras discretas. Redes Transformers podem abranger uma ou múltiplas sequências de tokens e o atributo de cada sequência (por exemplo, o rótulo da modalidade, a ordem sequencial), permitindo o aprendizado multimodal sem modificação arquitetônica (Baltrusaitis, Ahuja e Morency, 2017).

⁹¹ Arquiteturas Transformers podem ser usadas/adaptadas para modalidades (entradas e saídas), encontrando uma maneira de “tokenizar” a modalidade. Um método comum para criar modelos multimodais a partir de um LLM é “tokenizar” a saída de um codificador treinado. Obviamente, ampliam os problemas de “interpretabilidade” do modelo em função da heterogeneidade dos dados.

3.2. Os primeiros modelos e aplicativos generativos

A IA generativa não surgiu com o lançamento do ChatGPT, da OpenAI⁹², em novembro de 2022, uma aplicação do tipo específico de modelo de aprendizado profundo que emergiu dos avanços no campo de redes neurais Transformers e abordagens de aprendizado auto supervisionado e por reforço, caracterizando o limiar do novo benchmark da IA.

Os avanços foram muitos e ao longo de décadas, em termos de diferentes tipos de redes neurais, cujas principais são apresentadas na Tabela 3.1, inaugurando a IA generativa a partir das redes adversárias generativas (GAN). Entre essas novas redes neurais, que respondem pelo progresso notável dos recentes modelos, está a arquitetura *Transformers*⁹³, divulgada em artigo de pesquisadores do *Google Brain* (Vaswani et al., 2017) apresentada na Conferência NIPS 2017, empregadas em LLMs e permitindo superar grandes limitações das redes neurais existentes, como por exemplo as RNNs, sobretudo em NLP. O Google também foi responsável por inovações algorítmicas importantes como *Layer Normalization* (Ba, Kiros e Hinton, 2016). A OpenAI, por seu turno, contribuiu para o avanço dos LLMs baseados em redes Transformers desenvolvendo importantes técnicas pós-treinamento ou de ajuste fino dos modelos, como aprendizagem no contexto (Brown et al., 2020) e ajuste de instruções com feedback humano (Ouyang et al., 2022), sem esquecer da Meta AI (divisão da MetaPlatforms) e da DeepMind⁹⁴.

No campo de imagens, os avanços surgiram de novos recursos que asseguram maior estabilidade e eliminação de ruídos em modelos de difusão baseados em redes adversárias

⁹² A OpenAI é uma instituição criada em 2015, originalmente sem fins lucrativos, por Elon Musk, Sam Altman, Dario Amodei, Ilya Sutskever entre outros, dedicada à pesquisa em IA e a meta de desenvolver AGI. Em 2019, a OpenAI Inc., sem fins lucrativos, criou a subsidiária OpenAI LP (*Limited Partnership*), com natureza limitada com fins lucrativos (lucro limitado a 100 vezes o investimento), assim viabilizando o primeiro aporte de recursos da Microsoft. Isso levou a saída de co-fundadores, primeiro Musk e Amodei até recentemente Sutskever após divergências internas sobre a segurança dos modelos, restando atualmente apenas três dos seus onze cofundadores (Zitron, 2024b) e especulações de se transformar em empresa com fins lucrativos.

⁹³ O Transformers é uma arquitetura de modelo de aprendizado profundo que se baseia no mecanismo de auto-atenção (*self-attention*) para computar representações de entradas e saídas, pesando a influência de diferentes partes dos dados de entrada. Ao contrário de outras redes neurais que lidam com dados de entrada sequenciais, essa nova rede não exige que sejam processados em ordem, identificando o contexto para qualquer posição na sequência de entrada, possibilitando maior paralelismo, maiores conjuntos de dados e menor tempo de treinamento (DeepLearning Book, 2022).

⁹⁴ A DeepMind, criada em 2010 e adquirida pelo Google em 2014 (por US\$ 400 milhões), passou a chamar Google DeepMind, em 2023, quando foi fundida com as antigas divisões de pesquisa Google Brain e Google Research. É uma das empresas mais proficientes na criação de modelos revolucionários e central na pesquisa do grupo Alphabet/Google. Como mencionado, iniciou no campo de games, como Atari, despoitando com os modelos AlphaGoLee, AlphaGoZero, AlphaGo, AlphaGoMaster, AlphaGoFun e o recente AlphaStar. Recentemente se destacou com os premiados AlphaFold (2018) e AlphaFold2 (2020), com aplicação em biologia (especificamente, previsão de estrutura de proteínas), além do AlphaFold3, parceria com a Isomorphics Labs, com grande potencial no desenvolvimento de medicamentos (Abramson et al., 2024), do AlphaCode (Li et al., 2022), de geração de código de programação, e do AlphaGeometry voltado para raciocínio quantitativo (Trinh et al., 2024), visando superar limitações dos LLMs para uso em problemas matemáticos (Romera-Paredes et al., 2024).

generativas (GAN)⁹⁵ (Goodfellow et al., 2014), além das VAEs (*Variational Auto-Encoders*) (Kingma e Welling, 2013)⁹⁶ e, principalmente, os avanços em novos modelos de difusão alimentados por Transformers usados em imagem (Croitoru et al., 2023). Destacam-se o Midjourney, notório desde a criação da imagem do Papa Francisco usando um *puffer* Balenciaga, e o Stable Diffusion⁹⁷, ambos lançados em 2022, baseados em difusão latente e Transformers de difusão (DiT) (Rombach et al., 2021)⁹⁸ e abertos ao público pelas empresas emergentes Midjourney e Stability AI, criando entusiasmo similar ao ChatGPT⁹⁹.

Um número crescente de modelos brotaram no mercado com arquiteturas variadas, empregando técnicas distintas, inclusive combinadas, em arquiteturas híbridas. Por exemplo, o recente Sora (OpenAI) é construído em uma arquitetura Transformer de difusão (DiT), com codificador e decodificador VAE, enquanto o JEPA (*Joint Embedding Predictive Architecture*) (Meta), também um modelo de vídeo, não é generativo e faz previsões no espaço de representação. Mas o *breakthrough* científico parece ser os Transformers em processamento de linguagem e, posteriormente, em modelos de texto para imagem (Dosovitskiy et al., 2021), entre outros domínios.

⁹⁵ As GANs (sigla de *Generative Adversarial Networks*) funcionam através de duas redes neurais concorrentes, uma rede geradora e outra discriminadora. A rede geradora cria amostras aleatórias e a rede discriminadora tenta distinguir entre amostras geradas artificialmente e amostras reais. Sua função objetivo é treinar a rede geradora para criar amostras cada vez mais convincentes que possam enganar a rede discriminadora. Essa técnica é a base do software FaceApp que causou alvoroço nas redes sociais há alguns anos, ao criar imagens e promover o envelhecimento/rejuvenescimento de fotos. Existem outras redes neurais com potencial ainda pouco explorado, como as Wavenets, redes probabilísticas e autorregressivas preditivas de áudio, desenvolvidas pela DeepMind (van den Oord et al., 2016).

⁹⁶ As VAEs, da família de modelos gráficos probabilísticos e métodos bayesianos variacionais, aprendem propriedades estatísticas de um conjunto de dados, como imagens, para gerar novos dados. Modelos codificadores automáticos incluem codificadores (rede neural que aprende a codificar e compactar e aprender a representar a natureza dos dados de entrada em uma representação intermediária), código (sequência de números) e decodificador (outra rede neural que usa o código, que aprendeu a descompactar, reconstrói dados no formato de entrada na saída). Há atualmente muitas variantes de autoencoders, que introduziram várias melhorias, os mais populares os VAEs, usados na geração de imagem (WIPO, 2024a). Sobre vários desses conceitos da IA generativa, ver também Suresh (2023).

⁹⁷ O modelo de imagem LDM (sigla de *Latent Diffusion Model*), lançado em licenciamento permissivo em código aberto, é baseado em um tipo de rede neural generativa profunda utilizada em poderosos auto-encoders pré-treinados e GANs, desenvolvida junto com grupo CompVis/Universidade de Munique e a startup Runway, com apoio da EleutherAI e LAION (*Large-scale Artificial Intelligence Open Network*, organização sem fins lucrativos que fornece dados de imagem, com a inovação levada ao mercado pela Stability AI).

⁹⁸ Um modelo de difusão para geração de imagens envolve uma rede neural para prever e remover ruído em uma imagem; primeiro aplicando pixels aleatórios (ruídos) a uma imagem para depois usar a rede neural para sua remoção, daí resultando uma nova imagem, controlada por mecanismos adicionais de aprendizado de máquina. Os modelos de difusão fizeram progressos consideráveis nos últimos anos e agora são muito bem-sucedidos para as gerações de texto para imagem, como o DALL-E da OpenAI. Ver também Ho, Jain e Abbeel (2020).

⁹⁹ Os avanços em síntese de texto para imagem estiveram baseados em GANs, modelos autorregressivos (AR) - modelos estatísticos preditivos para prever valores futuros em uma série temporal com base em valores passados, uma representação de um tipo de processo aleatório (conceito semelhante à análise de regressão, baseado em aprendizado de máquina), e nos mais recentes modelos de difusão, que estão na vanguarda dos avanços em imagem - a difusão estável é usada principalmente para criar imagens condicionadas a descrições de texto -, gerando imagens mais diversificadas e mais eficazes em parâmetros do que os modelos AR (Pan et al., 2022).

Os primeiros LLMs surgiram em 2018, disponibilizados para comunidades de “devs.”, não para o público em geral, como o BERT (*Bidirectional Encoder Representations from Transformers*) da Alphabet/Google, e o GPT (*Generative Pre-trained Transformer*) (Radford et al., 2018), da OpenAI, bem como as versões subsequentes do GPT-2 (fevereiro de 2019), do GPT-3 (junho de 2020) e do GPT-3,5 (2022).

Tabela 3.1 – Sumário de redes neurais artificiais de aprendizado profundo

	Redes neurais artificiais	Pontos-chave descritivos	Artigos
IA preditiva	RvNN - Redes neurais recursivas	Usa uma estrutura semelhante a uma árvore. Uso preferencial em NLP	Goller e Kuchler (1996) Socher et al. (2011)
	RNN – Redes neurais recorrentes	Bom para informações sequenciais. Rede neural artificial bidirecional, o que significa que permite que a saída de alguns nós afete a entrada subsequente para os mesmos nós. Uso preferencial em NLP e processamento de fala	Cho et al. (2014) Li e Wu (2015)
	CNN – Redes neurais convolucionais	Originalmente para reconhecimento de imagem. Estendido para NLP, processamento de fala e visão computacional	LeCun e Bengio (1995) Krizhevsky, Sutskever e Hinton (2012) Kim (2014) Abdel-Hamid et al. (2014)
	DBN – Redes de crenças profundas	Aprendizado não supervisionado. Conexões direcionadas	Hinton (2009) Hinton et al. (2012)
	Long Short-Term Memory (LSTM)	Variação de rede recorrente	Hochreiter e Schmidhuber (1997)
	DBM - Máquinas de Boltzmann Profundas	Aprendizado não supervisionado. Modelo composto de conexões não direcionadas de RBMs	Salakhutdinov e Hinton (2009) Salakhutdinov e Hinton (2012)
IA generativa	GAN - Redes Adversariais Generativas	Aprendizado não supervisionado. Estrutura teórica dos jogos	Goodfellow et al. (2014) Radford et al. (2015)
	VAE - Auto-codificadores variacionais	Modelo gráfico probabilístico de aprendizado não supervisionado	Kingma e Welling (2013)
	Transformers (Attention Networks)	Originalmente empregado em NLP, estendido para visão computacional etc.	Vaswani et al. (2017)

Fonte: Pouyanfar et al. (2018:924) complementado com Van Veen e Leijnen (2019)

Este último é o modelo base do ChatGPT, treinado com gigantescos e variados conjuntos de dados da Internet coletados até setembro de 2021, sem pré-rotulagem, finalmente disponibilizado ao público, de forma gratuita, no final de 2022. Posteriormente vieram as versões aprimoradas disponibilizadas por assinatura paga, o GPT-4 (março de 2023), que aceita entradas de texto e imagem e tem capacidade de processamento multimodal, e o GPT-4 Turbo (novembro de 2023)¹⁰⁰, com novas funcionalidades, mais dados de treinamento e maior número de parâmetros (variáveis que as redes neurais otimizam continuamente durante o treinamento para minimizar erros, cujo número é usado como *proxy* do tamanho e complexidade do modelo) e GPT-4V (com componente de visão). A nova versão multimodal GPT-4o (omni) (maio de

¹⁰⁰ Os modelos CLIP e DALL-E (OpenAI) já tinham capacidades similares, revolucionando a multimodalidade empregando aprendizado auto supervisionado.

2024) promete revolucionar os assistentes virtuais, inclusive no modo gratuito, com capacidades significativamente aprimoradas de texto, visão e áudio, além de interação de texto e voz, com visualização de imagem em tempo real pela câmera do dispositivo, com baixa latência e uma efetiva interação com o usuário.

Mesmo que parte de uma estratégia midiática, o lançamento do ChatGPT foi comparado ao surgimento do primeiro computador pessoal nos anos 1970 e à primeira interface amigável de sistemas operacionais com o Windows 1.01 na década de 1980. O fundador da Microsoft reforçou o entusiasmo em seu blog:

"The development of AI is as fundamental as the creation of the personal computer. It will change the way people work, learn, and communicate--and transform healthcare. But it must be managed carefully to ensure its benefits outweigh the risks. I'm encouraged to see this early exploration of the opportunities and responsibilities of AI in medicine." (Gates, 2023).

De fato, o aplicativo conversacional ChatGPT representou um avanço extraordinário em termos de habilidades de conversação das assistentes existentes, como Alexa, Siri, Google e outras capazes de fornecer respostas predefinidas limitadas. Mas seu sucesso estrondoso foi muito além de substituir produtos existentes (como assistentes ou mecanismos de busca online) e mesmo da arquitetura Transformers, em função de quatro razões básicas: 1) melhorar o modelo com maior treinamento, número de parâmetros e refinamento técnico pela OpenAI; 2) empregar uma interface simples e revolucionária com o usuário, baseada em instrução textual (*prompts*), que dispensa conhecimento em programação¹⁰¹; 3) ser disponibilizado de forma gratuita para o público em geral, não somente para “devs.”, como era o padrão precedente, facilitado pelo uso dos *prompts*; e 4) ter sido lançado apoiado em uma estratégia de marketing arrojada, para a qual contribuiu Microsoft, acionista minoritária e membro do conselho de administração da OpenAI, ainda que sem direito a voto, mas com influência inequívoca como demonstrado no episódio de embate entre o CEO e o Conselho de administração da empresa, além de grande suporte financeiro e computacional, como será visto (The Economist, 2024).

De fato, há tempos as BigTechs buscavam desenvolver *chatbots* mais flexíveis. Com esse intuito a Microsoft adquiriu a startup Semantic Machines, em 2018 (Johnson, 2018), sem muito êxito no mercado, até a parceria com o ChatGPT, do qual obteve licença exclusiva para uso do modelo GPT-3,5 (Scott, 2020) incorporado ao mecanismo de busca Bing (Marshall,

¹⁰¹ *Prompts* são textos que passam instruções ou contextos - uma interface de linha de comando (CLI, sigla de *command-line interface*) - a um programa para que uma resposta seja gerada com base na solicitação feita por texto. Os sistemas funcionam a partir de uma instrução textual do usuário, cuja saída resultante, que pode ser textual, visual e outras, é determinada pelo modelo e material com que foi treinado. Em sistemas operacionais como Windows, o *prompt* de comando é um programa que emula o campo de entrada em uma tela de interface do usuário baseada em texto com a interface gráfica do usuário (GUI, sigla de *Graphic User Interface*) do Windows, e representou um grande avanço em relação aos sistemas MS-DOS que exigiam programação.

2023), lançado em 2009 para competir com os buscadores do Google e Yahoo!, até hoje concorrente inexpressivo do Google Search¹⁰².

O Google, responsável pela inovação disruptiva da arquitetura Transformers e grandes avanços em tipos de aprendizado – base do ChatGPT, junto a importantes melhorias incrementais da OpenAI - e desenvolveu inúmeros modelos disponibilizados para “devs.” desde o BERT, como os modelos de linguagem MEENA (Adiwardana et al., 2020), LaMDA (Thoppilan et al., 2022), GLaM (Du et al., 2021), Minerva 540B (Lewkowycz et al., 2022), entre outros. Chegou a lançar Meena como *chatbot*, em 2020 (Agarwal, 2020) com capacidade computacional e treinado com mais dados do que o então principal concorrente (GPT-2) obtidos de conversas em redes sociais, mas sem grande êxito.

O Google parecia seguir uma estratégia mais difusa, não focada exclusivamente em LLMs, mas em domínios variados – visão, imagem, robótica, games etc., - até que o surgimento do ChatGPT provocar reação imediata, com o lançamento do BARD (março de 2023), baseado no modelo LaMDA, com muitas falhas iniciais, recentemente atualizado com o Google Gemini (fevereiro 2024) (Gemini Technical Report, 2024), que na última versão, lançada como parte do projeto Astra (Sérvio, 2024) quase em simultâneo com o GPT-4o, em maio de 2024, demonstrou poder usar texto e imagem como entrada.

A estratégia do Google de não lançar no mercado um serviço similar ao ChatGPT para o público em geral, parece decorrente dessa abordagem diversificada (ou possíveis limitações técnicas remanescentes nos modelos), ao invés de um “efeito Kodak”¹⁰³ como atribuído por muitos analistas, com um *chatbot* conversacional visto como concorrente do exitoso mecanismo de busca online atrelado à principal fonte de receita de publicidade online do grupo Alphabet. Apesar do sucesso do ChatGPT, a OpenAI ainda carece de um modelo de negócios que possibilite a monetização da inovação. Sua receita anualizada é de apenas US\$ 1,6 bilhão, em 2023, frente à receita de US\$ 175 bilhões do Google Search Engine, que responde por 58% das receitas Google - apenas a título de ilustração, já que são produtos de mercados distintos. O Google justificava o não lançamento de um produto para o público em geral em

¹⁰² O Google Search é líder absoluto de mercado entre os mecanismos de buscas online, desde 2015, a partir do seu pioneiro e exitoso esforço de mapear e extrair padrões e relações da Web. Atualmente com *market share* global de 90,8% (maio/2024) (93,11%, em maio/2023), é seguido de longe pelo Bing (que apresentou pequeno crescimento desde a incorporação do ChatGPT, passando do *market share* de 2,77%, em maio/2023, para 3,72%, maio/2024). Statcounter. Disponível em: <https://gs.statcounter.com/search-engine-market-share>. Acesso 22 jun. 2024. O Bing chegou a suplantear o Yahoo!, em 2016, com 21% do mercado, mas distante do Google Search, então com 64% do mercado.

¹⁰³ A expressão é atribuída ao fato de a Kodak ter desenvolvido, mas não lançado no mercado, a primeira máquina fotográfica digital face ao risco de corromper sua principal fonte receita de revelação de filmes (Lucas Jr., Goh 2009).

função de razões éticas e reputacionais por serem modelos ainda experimentais e sujeitos a erros e “alucinações”¹⁰⁴.

De fato, além desses problemas, os aplicativos conversacionais ainda são inseguros, incluindo riscos de vazamento de informações confidenciais, o que levou muitos dirigentes de empresas usuárias, como Amazon, IBM, Apple, JPMorgan Chase e Samsung, que teve informações vazadas, a orientar funcionários a não compartilhar informações. Segundo um funcionário da Amazon: *“This is important because your inputs may be used as training data for a further iteration of ChatGPT, and we wouldn’t want its output to include or resemble our confidential information (and I’ve already seen instances where its output closely matches existing material)”* (Hurler, 2023).

Além do Google, outras empresas rivais lançaram produtos similares ao ChatGPT, como a chinesa Baidu com o ERNIE Bot (agosto 2023) (CNA, 2024), hoje uma família de vários modelos (que teria alcançado 200 milhões de usuários em abril de 2024), da mesma forma que Alibaba (modelo de linguagem Qwen; visão e linguagem Qwen-VL; e linguagem de áudio Qwen-Audio, todos sobre o modelo Tongyi Qianwen) (início de 2023) (Wang et al., 2024). O aplicativo Hunyuan da Tencent, criado para uso no WeChat (setembro de 2023) aberto para empresas chinesas e criarem aplicativos por meio de APIs na Tencent Cloud (Tencent, 2023). A Amazon anunciou seu *chatbot* de compras Rufus (fevereiro de 2024) (Ovide, 2024), além de disponibilizar modelos Anthropic, StabilityAI e AI21 Labs e até HuggingFace (Bass, 2023), na nuvem no Amazon Bedrock, serviço destinado a construir e dimensionar aplicativos de IA da AWS. Por fim, a Meta AI lançou seu novo AI Assistant baseado nos modelos LLaMA (fevereiro de 2023) (Heath, 2023) recentemente atualizado para o LLaMA 3. Todas essas empresas anunciam *chatbots* integrados aos seus serviços de plataforma digital, enquanto a OpenAI ainda luta para desenvolver um modelo de negócios que permita a monetização da sua inovação.

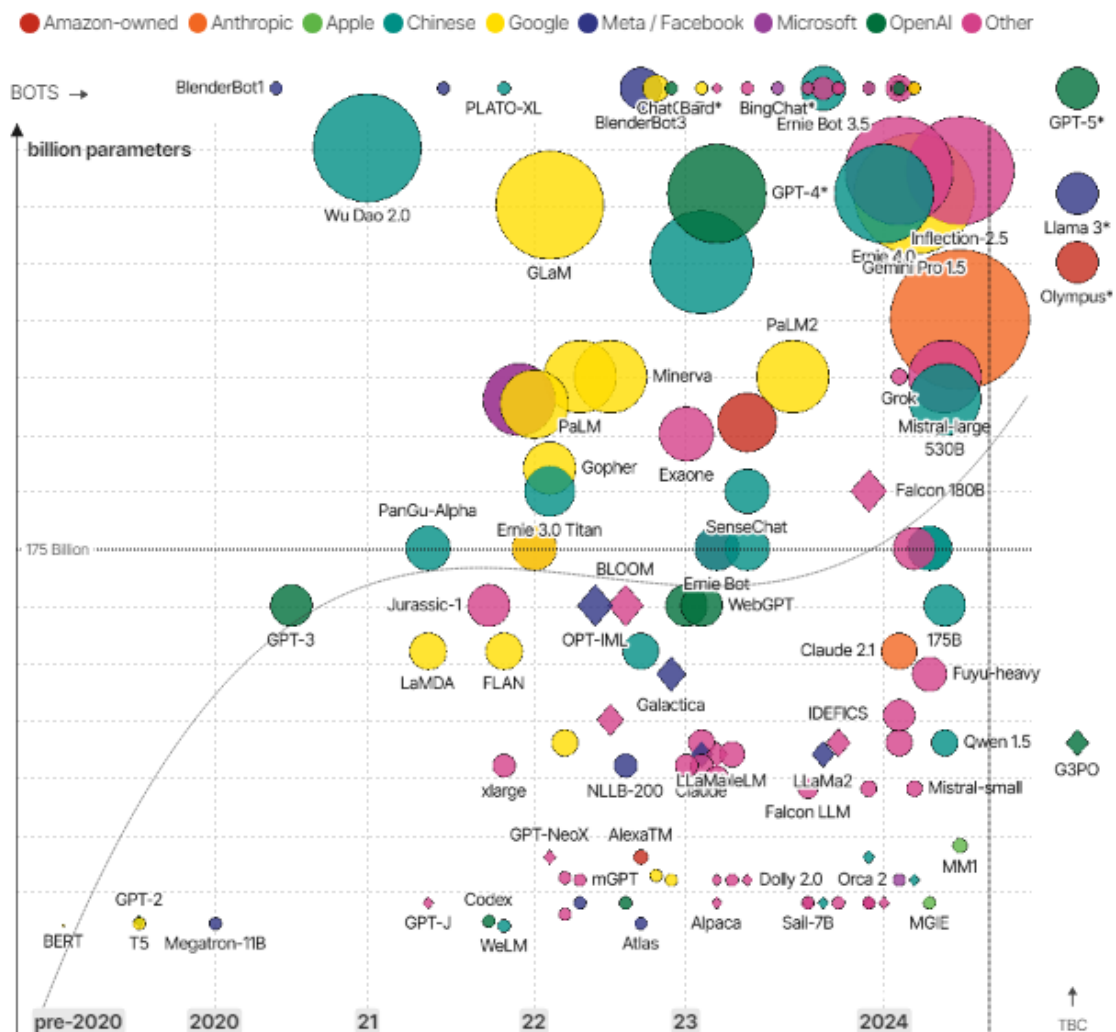
A mudança de paradigma não parou por aí. O mês de março de 2023 foi prolífico no início da onda de modelos abertos. A HuggingFace lançou o HuggingChat, baseado no LLM de código aberto ajustado para conversa, o Open Assistant; a Stability AI lançou o StableLM, LLM de código aberto e, logo em seguida, o aplicativo StableVicuna otimizado para conversa; a Universidade de Stanford lançou o Alpaca. Empresas entrantes da IA, como Databricks - empresa de software que se especializou na criação e venda de bancos de dados - lançou o Dolly

¹⁰⁴ Alucinação é um termo técnico cunhado em artigo de pesquisadores do Google (Lee et al. 2018), um fenômeno inerente a esse tipo de modelos que gera saídas “erradas completamente desvinculadas do material de origem, mesmo que semântica ou sintaticamente plausíveis.

até o recente DBRX, enquanto a startup de chips, Cerebras, lançou o Cerebras-GPT. Na sua maioria, esses e outros foram construídos sobre o LLaMA, da Meta AI, e conjuntos de dados e modelos da EleutherAI.

A proliferação de LLMs e aplicativos conversacionais foi frenética nos últimos anos (Gráfico 3.1), por gigantes do setor, empresas de menor porte e organizações acadêmicas ou institutos sem fins lucrativos (como o Allen Institute for AI). O Gráfico apresenta os principais LLMs (e aplicativos, na linha superior “Bots”), por número de parâmetros dos modelos. As cores representam os criadores, o tamanho dos círculos os parâmetros dos modelos e o formato de diamante mostra modelos de acesso aberto, outra novidade do paradigma da IA generativa. Modelos grandes têm mais de 175 bilhões de parâmetros.

Gráfico 3.1 – Ascensão e Crescimento dos principais LLMs e bots associados, por número de parâmetros, principais desenvolvedores (posição até 20 de março de 2024)



Fonte: Information is Beautiful¹⁰⁵

¹⁰⁵ Disponível em: <https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/>. Acesso em: 20 abr. 2024

Embora alguns *chatbots* sejam construídos sobre modelos próprios, como no caso das BigTechs (com exceção da Microsoft, que se baseia nos modelos da OpenAI) e emergentes como OpenAI, Anthropic PBC¹⁰⁶ e Inflection AI, há muitos meros APIs construídos a partir de modelos pré-treinados de terceiros disponibilizados em *open source*. Fei-Fei Li, ex-cientista chefe do Google Cloud AI, descreve APIs como interruptores que ativam imediatamente o componente inteligente do modelo de IA e a maneira mais fácil de sua colocação no mercado, (Bhatia, 2017), embora meras interfaces com o usuário que não envolvem os enormes desafios técnicos de desenvolver um modelo.

3.3. Muito além de LLMs e aplicativos conversacionais

A proliferação em massa não se restringiu aos chatbots e modelos de criação de texto (LLMs) por incumbentes e novos *players*, mas um ritmo acelerado de modelos de texto para imagem (DALL-E/OpenAI, CogView/BAAI, Imagen/Google, Stable Diffusion/StabilityAI, Midjourney), texto para vídeo (Make-me-a-Video/Meta, Imagen Video/Google), áudio (MusicLM/Google, HarmonAI, WaveNet/DeepMind, DeepVoice/Baidu), código de programação (Codex/OpenAI, AlphaCode/DeepMind, Copilot/GitHub¹⁰⁷) e tarefas de raciocínio lógico - limitação antiga dos LLMs¹⁰⁸, ainda que com menor desempenho, com exceções como Minerva540B e PaLM2 (ambos Google) que demonstraram capacidades aprimoradas em lógica, e recentes modelos da DeepMind -, além de modelos de biologia, robótica e outros domínios, além dos versáteis e flexíveis modelos multimodais e os poderosos e promissores modelos de fundação.

A Figura 3.1 evidencia os tipos de modelos de IA generativa e os principais dados de entrada e saída desses modelos.

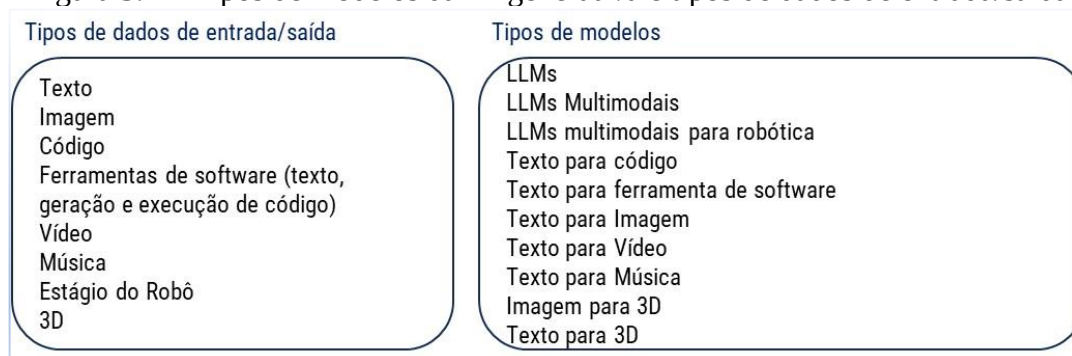
Embora a maioria dos modelos predominantes sejam LLMs, também avançam modelos de domínios distintos, como visão, jogos e outros (Gráfico 3.2), sendo notável o crescimento dos modelos dos domínios da biologia, robótica e, particularmente, multimodais (Gráfico 3.3).

¹⁰⁶ A Anthropic foi a primeira criada, em 2021, por ex-cofundadores da OpenAI, como *Public-Benefit Corporation*, pelo ex-vice-presidente Dario Amodei que deixou a empresa por discordar da parceria com a Microsoft. Tem parcerias com o Google (investimento de US\$ 1,5 bilhão) e Amazon (US\$ 4 bilhões), com modelos disponibilizados no Amazon Bedrock AWS.

¹⁰⁷ O Copilot, desenvolvido por Microsoft/OpenAI/GitHub está incorporado em produtos Microsoft, como word, planilhas excel etc. oferecendo novas funcionalidades.

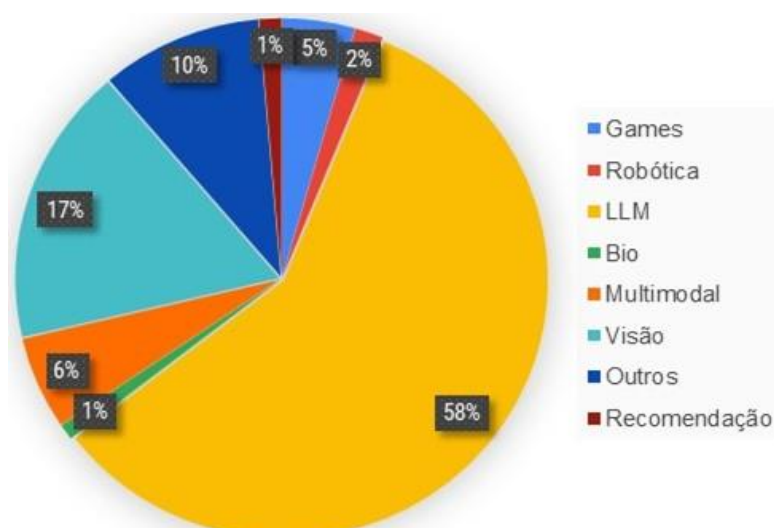
¹⁰⁸ Capacidades de problemas complexos como raciocínio quantitativo para solução de problemas eram as principais limitações dos LLMs (Lewkowycz et al., 2022), testados em Olimpíadas Internacionais de Matemática. O desempenho desses modelos está abaixo do nível de um estudante de pós-graduação (Frieder et al., 2023) embora não seja claro se tais limitações serão relevantes para o uso em pesquisa e aplicações nesse campo (Davies, 2023).

Figura 3.1 – Tipos de modelos da IA generativa e tipos de dados de entrada/saída



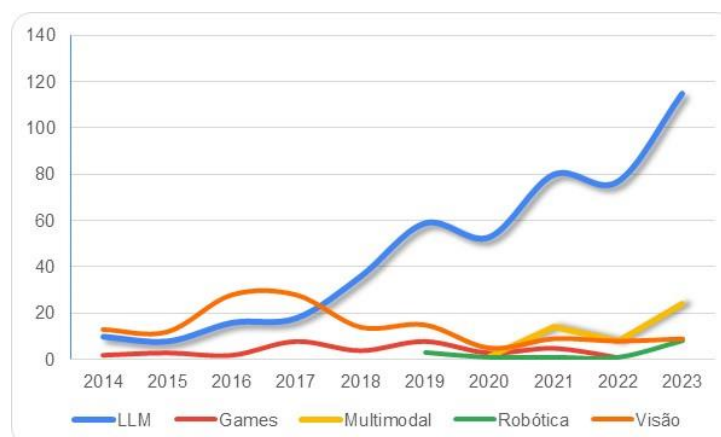
Fonte: Benaich (2023)

Gráfico 3.2 – Modelos de IA generativa, por domínio (percentual do número total), acumulado 2014-2023



Fonte: Elaborado a partir da base de dados da Epoch

Gráfico 3.3 – Evolução do número de modelos destacados da IA generativa, por domínio, 2014-2023



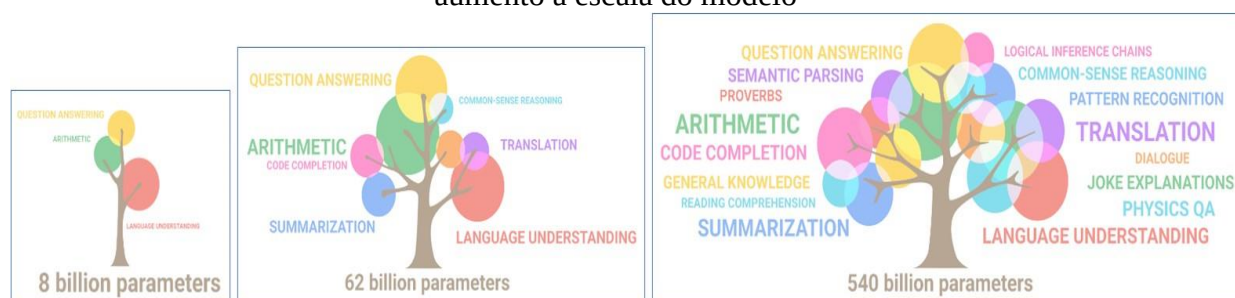
Fonte: Elaborado a partir da base de dados da Epoch

Küspert, Moës, Dunlop (2023) observam que as principais características técnicas dos modelos são o grande tamanho, a opacidade e o potencial de desenvolver capacidades inesperadas além das pretendidas pelo desenvolvedor. São construídos sobre um número

elevado de parâmetros, que crescem 2,8 vezes por ano desde 2018 (Azhar et al., 2024), atingindo ordens de grandeza de bilhões ou mesmo trilhões (medida usada como *proxy* do seu tamanho e complexidade), treinados com extensas bases de dados e intensivos computacionalmente. Empregam aprendizado auto supervisionado, paradigma que aproveita a *estrutura inerente dos próprios dados* para gerar rótulos para treinamento, seguido por ajuste fino em cada tarefa específica - por meio de instrução (*Self-Instruct*) ou de aprendizado por reforço (*RLHF*, sigla de *Reinforcement Learning from Human Feedback*, técnica de aprendizado por reforço desenvolvida pela OpenAI e DeepMind/Google), entre outras (Ramponi, 2023).

De fato, há uma clara corrida por modelos gigantes, cada vez mais poderosos, com melhor desempenho e incorporação de múltiplas tarefas à medida que aumentam em escala (Figura 3.2).

Figura 3.2 – Representação da melhoria do desempenho e ampliação de tarefas com o aumento a escala do modelo



Fonte: Narang e Chowdhery (2022) (blog do Google Research)

A Tabela 3.2 apresenta informações de importantes LLMs e multimodais, sem a pretensão de ser exaustiva diante do ritmo acelerado de lançamento, não incluindo os modelos recentes Gemini (Google DeepMind, 2024a; Google DeepMind, 2024b) Claude 3 (Anthropic), Sora (OpenAI, 2024a), V-JEPA (Meta), LLaMA 3 70B (Meta), GPT-4o (OpenAI), Grok (xAI), etc. Os complexos modelos da IA generativa, são igualmente empíricos, cuja opacidade parece aumentar com o aumento de parâmetros, os montantes de dados e os cálculos computacionais internos, tornando-os “caixas pretas” cada vez mais inescrutáveis. Os modelos começaram a exibir recursos surpreendentes, que não haviam sido explicitamente programados por seus devs (Chen et al., 2021), principalmente a partir de 2022, um fenômeno que é chamado “emergência”¹⁰⁹. Não são poucos os exemplos. O SAM (*Segment Anything Model*), da Meta AI é capaz de “cortar” qualquer objeto em uma imagem ou vídeo sem ter sido treinado para isso.

¹⁰⁹ Emergência significa que o comportamento de um sistema é induzido implicitamente, em vez de construído explicitamente, que ao mesmo tempo que gera entusiasmo científico é fonte de inquietação sobre consequências não intencionais e riscos potenciais (Bommasani et al., 2022). Outros problemas recentes envolvem dados, como o “engano imitativo” (Lin, Hilton e Evans, 2022).

O Grover-Mega-1,5B, desenvolvido por pesquisadores da Universidade de Washington para detectar notícias falsas, também pode gerá-las (*Neural Fake News*) (Zellers, 2019; Zellers et al., 2020). O GPT-3, treinado para processar linguagem natural, adquiriu capacidade de escrever programas. O AutoML (Google) automaticamente descobre a melhor arquitetura para uma tarefa de visão computacional. Hacker, Engel e Mauer (2023:2) esclarecem que isso decorre dos modelos “encontrarem padrões e relacionamentos nos dados *por conta própria*, sem serem explicitamente informados sobre o que procurar” (grifos da autora).

Tabela 3.2 – Principais LLM e multimodais, por número de parâmetros e desenvolvedor

LLMs	Nº Parâmetros	Desenvolvedor	País	Referência
LaMDA (<i>Language Models for Dialog Applications</i>)	137 bilhões	Google	US	Thoppilan et al. (2022)
JURASSIC-1 Jumbo	178 bilhões	AI21 Labs	IL	Lieber et al. (2022)
GPT-3 175B (<i>Generative Pre-trained Transformer</i>)	175 bilhões	OpenAI	US	Brown et al. (2020)
OPT-175B (<i>Open Pretrained Transformer</i>)	175 bilhões	Meta	US	Zhang et al. (2022)
Gopher (280B)	280 bilhões	DeepMind	UK	Rae et al. (2022)
Megatron-Turing NLG 530B (MT-NLG)	530 bilhões	NVIDIA, Microsoft	US	Smith et al. (2022)
Minerva (540B)	540 bilhões	Google	US	Lewkowycz et al. (2022)
PaLM-540B (<i>Pathways Language Model</i>)	540 bilhões	Google	US	Chowdhery et al. (2022)
GLaM (<i>Generalist Language Model</i>)	1,2 trilhão	Google	US	Du et al. (2022)
Claude 2 (<i>AnthropicLM v4-s3</i>)	52 bilhões	Anthropic	US	Anthropic
Chinchilla	70 bilhões	DeepMind	UK	Hoffman et al. (2022)
ERNIE Titan (Enhanced Representation from Knowledge Integration)	260 bilhões	Baidu	CN	Ernie 2023
BLOOM (BigScience Large Open-science Open-access Multilingual Language Model)	176 bilhões	BigScience	n.a.	Blog BigScience
Multimodais	Nº Parâmetros	Desenvolvedor		Referência
M6-10T (Multi-Modality-to-Multi-Modality Multitask Mega-transformer for Unified Pretraining)	10 trilhões	Alibaba Dharma Institute	CN	https://damo.alibaba.com/Lin et al. (2021)
GPT-4	n.d. ¹	OpenAI	US	https://arxiv.org/abs/2303.08774
ERNIE 4.0	n.d.	Baidu	CN	https://cdn.openai.com/papers/GPTV_System_Card.pdf
BaGuaLu	174 trilhões	Alibaba, Zhejiang Lab, Tsinghua Univ, BAAI	CN	https://dl.acm.org/doi/abs/10.1145/3503221.3508417
Gemini Ultra	n.d.	Google	US	https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
ChatGLM3	130 bilhões	Zhipu AI	CN	https://www.zhipuai.cn/en/news/76
Wu Dao 2.0	1,75 trilhão	BAAI	CN	https://www.engadget.com/china-s-gigantic-multi-modal-ai-is-no-one-trick-pony-211414388.html
PaLI-X	55 bilhões	Google	US	https://arxiv.org/pdf/2305.18565.pdf
Flamingo	80 bilhões	DeepMind	UK	https://arxiv.org/abs/2204.14198
Ferret (13B)	13 bilhões	Columbia Univ, Apple	US	https://arxiv.org/abs/2310.07704
ImageBind	932 milhões	Meta	US	https://arxiv.org/abs/2305.05665

Fonte: Atualizado de Maslej et al. (2023) com a base de dados Epoch e referências citadas. A tabela não é exaustiva.

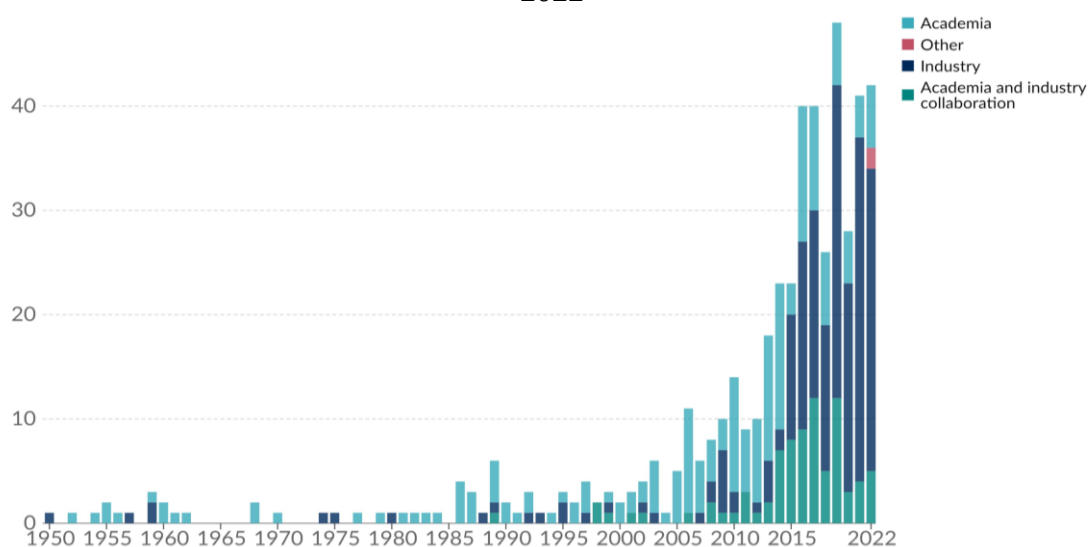
¹ Estimativas apontam o número de parâmetros ajustáveis entre 1 trilhão (Semafor) e 100 trilhões (SSC NR)

O aumento da escala dos modelos passou a depender de três elementos básicos: desenvolvimento de arquiteturas que possam ser executadas em paralelo, disponibilidade de

grandes quantidades de dados para treinamento e melhorias no *hardware*. Algumas pesquisas recentes, contudo, contestam a hipótese do tamanho dos modelos como decisiva para seu desempenho, dado que modelos menores, como Chinchilla (DeepMind), Claude2 (Anthropic), PALI-X (Google), etc. demonstraram desempenho similar aos grandes modelos a partir de treinamento mais extenso e maiores/melhores conjuntos de dados (Hoffman et al., 2022). Não obstante, o padrão predominante para melhor desempenho continua sendo modelos grandes. Entre os modelos notáveis identificados pela Epoch, que no seu conjunto cresceram dez vezes mais rápido após 2018, a maioria fica abaixo dos 20 bilhões de parâmetros ou acima dos 70 bilhões de parâmetros, com poucos na faixa 20B-70B (Villalobos et al. 2022; Hobbhahn, Heim e Avdos, 2023).

Os modelos maiores são um campo quase exclusivo de empresas, sobretudo de maior porte, e umas poucas emergentes “eleitas”. Em 2021, os modelos grandes representaram 96% do total de modelos lançados (frente a 11%, em 2010), com crescentes custos de desenvolvimento e treinamento, proibitivos para empresas menores, academia e organizações sem fins lucrativos (Sevilla et al., 2022). As empresas desenvolveram 84% dos modelos em 2022, que são, em média, 29 vezes maiores que os modelos acadêmicos, evidenciando uma assimetria, crescente, entre os dois tipos de organizações (Maslej et al., 2023). A predominância dos modelos criados por pesquisadores da indústria é também evidenciada por dados da Epoch Database (Gráfico 3.4).

Gráfico 3.4 - Número de modelos de AI notáveis, por afiliação e ano de publicação, 1950-2022



Fonte: Epoch Database apud OurWolrdinData¹¹⁰

¹¹⁰ Disponível em: <https://ourworldindata.org/grapher/affiliation-researchers-building-artificial-intelligence-systems-all>. Acesso em: 24 mai. 2024

Há uma concentração dos melhores talentos nas empresas, comparado a academia e governos. Em 2021, quase 70% dos PhD em IA nos EUA seguiram para a indústria (21%, em 2004), não para a universidade, que absorveu apenas 20% desses talentos, e o governo apenas 0,76% (Ahmed, Wahed e Thompson 2023; Maslej et al., 2024). Isso confere vantagens indiscutíveis às empresas, frente a governo e academia. A incomparável capacidade das empresas para atrair talentos é decorrente não somente de maiores e crescentes níveis salariais (na Meta, por exemplo, da ordem de US\$ 335 mil, em 2023, frente a US\$ 256 mil, em 2020, podendo chegar a US\$ 5 milhões/ano no caso de PhDs experientes), mas também dos seus sofisticados recursos computacionais.

A assimetria em capacidade computacional entre indústria e academia é talvez a mais dramática. O maior modelo criado pela academia usou apenas 1% da computação usada para treinar o maior modelo da indústria (Besiroglu et al., 2024). A Meta anunciou o objetivo de ampliar significativamente sua infraestrutura, adquirindo 350 mil novos chips de IA somente em 2024, ao passo que o importante grupo de Processamento de Linguagem Natural da Universidade de Stanford dispõe de apenas 68 GPUs (Nix et al., 2024)¹¹¹

Infraestrutura computacional robusta e capacidade de atrair talentos se tornam progressivamente exclusivos de grandes empresas, em especial as BigTechs, levantando preocupações sobre a concentração do progresso técnico em organizações motivadas pelo lucro (Ahmed, Wahed, Thompson 2023; Maslej et al. 2023; Rikap, 2021).

Cottier, Besiroglu e Owen (2023), entretanto, consideram inexorável essa concentração do progresso técnico nas empresas, pois modelos da IA generativa exigem extensivo treinamento, demandando expressiva capacidade computacional. De fato, a indústria é responsável por quase todas as execuções de treinamento com uso intensivo de computação.

3.4. Protagonistas e coadjuvantes da IA generativa

A IA generativa realmente parece representar um novo capítulo da IA e do aprendizado profundo. Não existe precedente histórico de uma tecnologia que avance em velocidade tão rápida com intenso lançamento de novos modelos, tanto por empresas líderes, quanto emergentes. As três variáveis principais para o desenvolvimento dos modelos recentes (elevados parâmetros, amplos dados de treinamento e recursos computacionais) tornam enigmático o surgimento de tantos *players* de menor porte.

¹¹¹ No State of the Union de 2024, a pesquisadora e codiretora do HAI-Stanford, Fei-Fei Li, solicitou ao governo Biden financiamento para uma estrutura nacional de poder computacional e conjuntos de dados para que pesquisadores acadêmicos de IA dos EUA possam acompanhar as gigantes da tecnologia. Há um projeto de lei em tramitação no Congresso americano, o Create AI Act, vai ao encontro dos anseios de “democratizar a IA”.

O ecossistema que alimenta o progresso técnico da IA se tornou mais diversificado nos últimos dois ou três anos, deixando de se limitar aos LLMs e às BigTechs americanas, ainda que Google (e a subsidiária britânica DeepMind), Meta e Microsoft se destaquem. A elas se somaram as BigTechs chinesas e um surpreendente conjunto de empresas emergentes no lançamento de modelos generativos disruptivos. Além da notória OpenAI, merecem destaque os modelos de imagem da empresa americana Midjourney Inc. e da britânica Stability AI; o Inflection-1 e Inflection-2 e o *chatbot* Pi da Inflection AI - criada por ex-fundadores da DeepMind e do LinkedIn¹¹²; os nove modelos identificados da família Claude (Claude3Opus, Claude3Haiku e Claude3Sonnet, com performance superior ao GPT-4 em testes no campo do direito como o Multistate Bar Examination) lançados pela Anthropic - criada por dois ex-membros sênior da OpenAI - em 2023 e 2024; a americana Runway AI Inc., com quatro modelos lançados desde 2021, incluindo o Stable Diffusion (LDM-KL-8-G), de texto para imagem, em código aberto, em parceria com CompVis Group da Ludwig Maximilian University of Munich (Rombach et al., 2021). Modelos acadêmicos menores, como o Koala, na versão BAIR (*Berkeley Artificial Intelligence Research*) e na versão HuggingFace Koala (13B-HF), com pesos do modelo abertos ao público.

Empresas de setores afins também invadiram o campo da IA generativa, como a MosaicML, da empresa americana de software Databricks¹¹³, que se autodenomina uma *lakehouse* (mistura de "*data warehouse*" e "*data lake*"), com modelos disponibilizados na plataforma de nuvem da Oracle Cloud; Adobe, que possui um banco de dados proprietário com centenas de milhões de imagens e investiu no modelo próprio *Firefly* (The Economist, 2023a). Gigantes de chips, como Nvidia, que abraçou rapidamente a arquitetura Transformers desenvolvendo ativos (tecnologias) complementares (chips especializados ou co-especializados) e integrando para modelos de IA desde 2017.

A avalanche de empresas emergentes não se restringe aos Estados Unidos. Muitas são chinesas, como SenseTime Inc., Inspur Group, LightYear AI (adquirida pela gigante de *food delivery* Meituan), Zhipu AI (com o modelo multimodal ChatGLM3), China Unicom (modelo Honghu Graphic), 360 Security Technology (modelo 360 Smart Brain), JD.com (ChatRhino), ByteDance (dona do TikTok) e recentemente a Kuaishou Technology (dona da Kwai, concorrente do TikTok, que lançou o recentíssimo modelo de texto para vídeo Klingai, que vem

¹¹² O LinkedIn foi adquirido pela Microsoft em 2016

¹¹³ A Databricks possui muitos dados e conexões corporativas, com mais de 7 mil clientes, tendo adquirido a MosaicML (US\$ 1,3 bilhão) que se destaca no campo dos modelos generativos, ainda assim dependente de GPUs das gigantes.

sendo chamada de Sora chinês). Muitas empresas surgiram em 2023 em meio ao *boom* de startups chinesas¹¹⁴ promovendo a rápida disseminação de LLMs e outros modelos de IA generativa, como DeepSeek, 01.AI (da Sinovation Venture, do renomado cientista e capitalista de risco Kai-Fu Lee), Kunlun Inc., Baichuan, iFlytek, DeepLangAI, etc.. Isso sem esquecer a academia, com a Tsinghua University¹¹⁵, e laboratórios sem fins lucrativos, como BAAI (*Beijing Academy of Artificial Intelligence*), com vários modelos que constam da Epoch Database.

A desconcentração geográfica da IA fora dos EUA e China parece ter início com as israelenses AI21Labs (com a família de LLMs Jurassic) e a Deci AI, a canadense Cohere, a coreana Naver Corp (HyperClova X), a japonesa Sakana AI e empresas europeias como as francesas Mistral AI (fundada em 2023 por ex-pesquisadores do Google/DeepMind e Meta, com LLMs disponibilizados por licença permissiva gratuita ou assinatura paga, disponível para download por vários meios, acesso remoto ou acesso por API)¹¹⁶ e LightOn (com seu modelo de proteínas RITA) (Hesslow et al., 2022) e a alemã Aleph Alpha (cujo LLM Luminous-supreme atraiu investimentos de incumbentes como Bosch, SAP, Hewlett Packard, que apoiam modelos de fundação dentro do princípio de soberania tecnológica¹¹⁷).

O florescimento de organizações sem fins lucrativos, repositórios *open-source* e grupos de pesquisa de IA têm avançado pelo mundo, com destaque para o mencionado HuggingFace e projeto alemão Large European AI Models (LEAM.AI) que desenvolveu o OpenGPT-X, a partir de infraestrutura de dados aberta Gaia-X, fundada por grandes grupos econômicos

¹¹⁴ Apesar da dificuldade para obter informações seguras, especula-se sobre uma explosão de *startups* chinesas de IA que lideram investimentos mundiais em venture capital (Shen, 2023). Peng (2023) menciona que existiam mais de 100 LLMs (até julho de 2023) na China (79 deles com mais de 1 bilhão de parâmetros), muitos ainda em fase de testes (embora a indústria trabalhe com a hipótese de que 85% dos modelos nunca saem do laboratório). Outra fonte identificada apenas ao final da pesquisa lista vários LLMs chineses (LifeArquite.ai, 2024), mas que não foi possível compatibilizar com os dados aqui apresentados pela necessidade de verificar sua fidedignidade (por exemplo, nesta base aparecem 5 modelos da Tencent que não constam das demais bases).

¹¹⁵ A Tsinghua University desenvolveu diversos modelos de IA, entre LLMs, robótica e modelos multimodais, sozinha ou em parceria com organizações chinesas, como BAAI, Alibaba, Huawei, Tencent e Zhipu AI, além de importantes empresas e universidades americanas, como Microsoft, Apple, Carnegie Mellon, Berkeley e Stanford. A BAAI parece privilegiar o desenvolvimento autônomo, com poucas parcerias (apenas Alibaba e Tencent).

¹¹⁶ Os modelos parecem ser um refinamento de modelos de linguagem "pequenos" como o LLaMA2 (Meta), iniciado no repositório GitHub e com canal no Discord para colaboração de devs, com alto desempenho e menor custo computacional. A estratégia francesa é ampla e híbrida, com apoio público e privado. O Kyutai Labs, organização sem fins lucrativos com foco em modelos de fundação, recebeu aportes do governo e grandes grupos privados da ordem de US\$ 300 milhões, inclusive a disponibilização a preço de custo de mil GPUs Nvidia adquiridos pela *Scaleway*, divisão na nuvem do grupo Iliad (Dillet, 2023). Esta empresa europeia de infraestrutura na nuvem lançou um programa de startups no grande no grande campus de startups *Station Fem Paris*.

¹¹⁷ O princípio de soberania tecnológica (Edler et al., 2021) é a missão precípua e princípio norteador da política digital e industrial europeia, em especial na Alemanha (Edler et al., 2023; EFI Report, 2022), mas também França.

franceses e alemães, com apoio governamental e da Comissão Europeia¹¹⁸. Também a comunidade transfronteiriça de pesquisa aberta EleutherAI, que opera em servidor Discord¹¹⁹ com vários LLMs, incluindo o BLOOM, primeiro LLM multilíngue, desenvolvido por mais de mil pesquisadores no âmbito do projeto BigScience (Le Scao et al., 2023) treinado no supercomputador francês Jean Zay, com doação de €3 milhões das agências francesas CNRS e GENCI, entre outras instituições de código aberto.

Surpreendentemente, países do Oriente Médio também avançam em IA generativa, em particular Arábia Saudita e Emirados Árabes¹²⁰. O *Techological Innovation Institute* (TII), dos EAU, desenvolveu os modelos Falcon40B e o Falcon180B, considerados mais poderosos que o LLaMA 2 (Meta), enquanto a Universidade King Abdullah de Ciência e Tecnologia (Kaust)¹²¹ da Arábia Saudita, o LLM AceGPT treinado em árabe, chinês e inglês, desenvolvido em colaboração com a Universidade Chinesa de Hong Kong, Shenzhen (CUHK-SZ) e o Instituto de Pesquisa de Big Data de Shenzhen¹²². A maior parte desses avanços tem sido atribuída a parceria com a China, embora a própria Microsoft tenha diversas parcerias com universidades chinesas, como Tsinghua e Peking University¹²³, realizando pesquisas na China desde 1998 (Rikap, 2024) e tendo instalado seu maior centro de pesquisa fora dos EUA, o Microsoft Research Asia (MSR Asia)¹²⁴.

O terreno caudaloso do Oriente Médio parece ser o principal campo de batalha EUA-China, com os principais países mantendo relações com ambos os países. A “ambiguidade” regional parece ter sido o motivo que impediu a região de ser incluída nas restrições à

¹¹⁸ O LEAM.AI é uma iniciativa da Associação Alemã de IA (KI Bundesverband) e um dos principais representantes da indústria no desenvolvimento de grandes modelos de IA. <https://leam.ai/>. Acesso 10 ago. 2023

¹¹⁹ O Discord é um repositório americano da empresa com mesmo nome, originário do segmento de games, que tem despertado o interesse de BigTechs, com propostas de aquisição pela Microsoft (por US\$ 10 bilhões, em 2021), Amazon e Epic Games.

¹²⁰ Os EAU almejam ter papel importante no cenário internacional, tendo definido, em 2019, a Estratégia Nacional de IA até 2031, projeto liderado pelo ministro da IA, Omar Sultan Al Olama, que visa integrar a tecnologia em aspectos críticos do governo, transformar o país em líder mundial em pesquisa da IA e fornecer dados às empresas. Conta com investimentos do Mubadala, o fundo soberano de Abu Dhabi, e o grupo de *private equity* norte-americano Silver Lake. Os EAU escolheram a Huawei para instalar a primeira infraestrutura 5G nacional em 2019, justificado pela não disponibilidade de tecnologia semelhante de parceiros ocidentais, embora as empresas estatais tenham assinado acordos com a sueca Ericsson, na política de diversificação de parceiros de telecomunicações.

¹²¹ Disponível em: <https://cemse.kaust.edu.sa/ai>. Acesso 03 ago. 2024

¹²² A Kaust.ai, liderada pelo cientista da computação alemão Jürgen Schmidhuber, está desenvolvendo um poderoso supercomputador, Shaheen III, objetivando ampliar em 20 vezes o poder de computação atual, a partir da contratação da Hewlett Packard Enterprise, com chips Nvidia. Um membro da equipe da Kaust afirma que as universidades chinesas ofereciam estagiários e estudantes mais baratos que poderiam realizar o trabalho mecânico e triagem de dados, mas nega que os alunos testem GPUs (trabalho restrito aos fornecedores e funcionários da Kaust) (Kerr et al, 2023).

¹²³ Os esforços da OpenAI em desenvolver chips próprios em parceria com empresas do Oriente Médio gerou apreensão por parte do governo americano, que menciona o risco de ser um mecanismo de buscar o acesso da China aos chips americanos.

¹²⁴ Disponível em: <https://www.microsoft.com/en-us/ard/overview>. Acesso em: 05 ago. 2024

exportação americanas de semicondutores. Resta saber se a disputa geopolítica – e as supostas prioridades de segurança nacional do governo americano - serão fortes o suficiente para sobrepujar interesses corporativos no ecossistema coopetitivo global de inovação da IA.

É intrigante como tantas empresas, inclusive emergentes de pequeno porte, estão conseguindo entrar para o clube seletivo dos inovadores de uma tecnologia tão complexa e que demanda ativos intangíveis e tangíveis complementares. Como explicar essa onda tão acelerada de modelos da IA generativa?

Em primeiro lugar, a tradição de divulgação dos detalhes técnicos dos modelos quase concomitantemente ao seu lançamento, por meio de artigos científicos publicados e/ou apresentados em conferências da área, certamente contribuiu para a ampla e rápida disseminação do conhecimento, sem paralelo histórico em outras áreas. Esse avanço tem ocorrido por meio de ampla pesquisa colaborativa e se expressa em variadas parcerias, desde a coautoria em artigos entre incumbentes BigTechs entre si (incluindo as chinesas) e com empresas integrantes do seu ecossistema de inovação, incluindo emergentes da IA, fornecedoras de ativos tangíveis (como chips), intangíveis (como dados) e instituições de ensino e pesquisa (americanas, chinesas e outras espalhadas pelo mundo), além das mencionadas organizações de pesquisa aberta.

Essa prática da IA corrobora a ideia de um padrão de competição atípico de natureza “coopetitiva” - marcado por competição e cooperação - entre os principais *players* (Teece 2018; Petit e Teece, 2020) dos ecossistemas digitais, que desde a década passada cultivam um padrão colaborativo de progresso técnico com a revelação dos modelos, mesmo que competindo entre si no lançamento de produtos (aplicações) no mercado, como evidencia a corrida de OpenAI, Google, Meta, etc., no momento de definição do *design* tecnológico dominante.

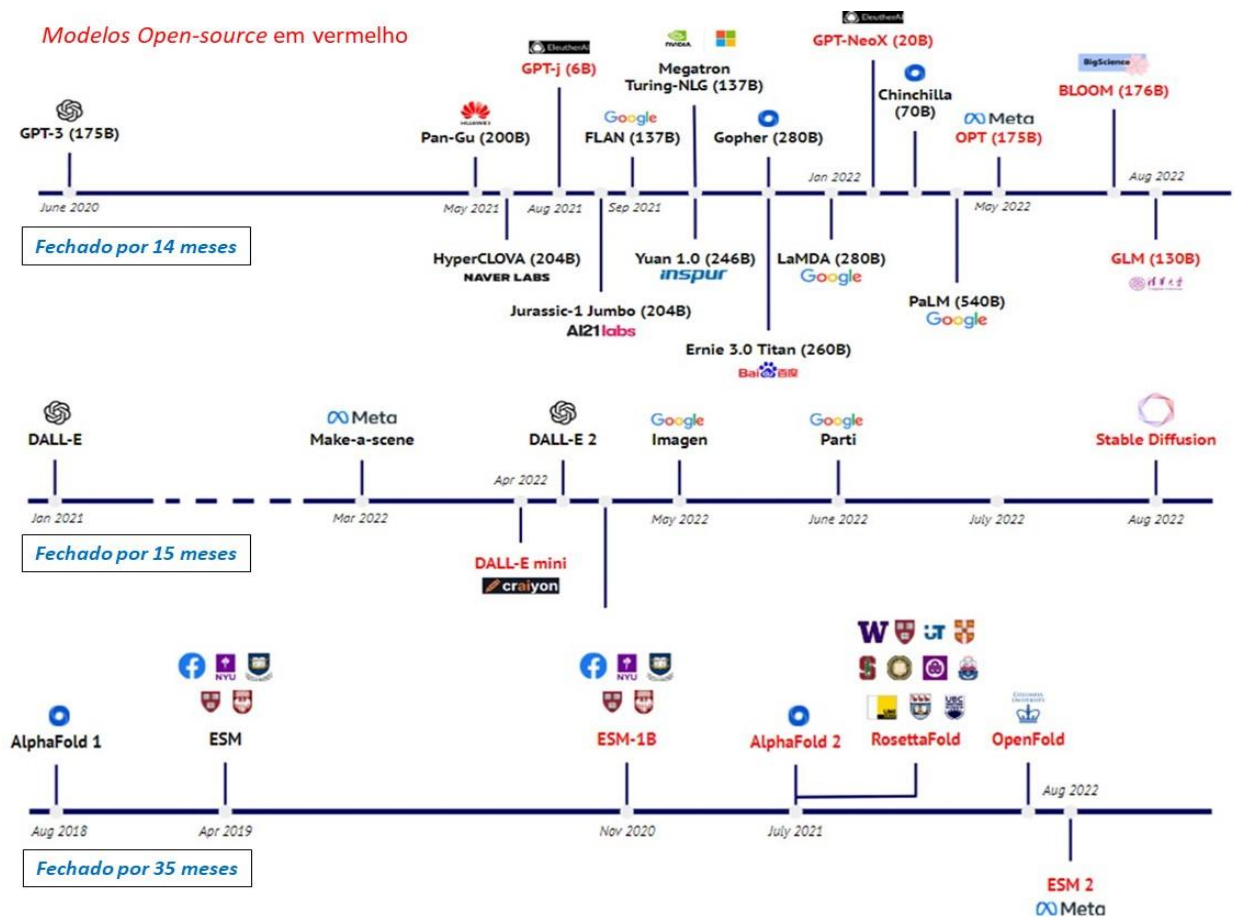
Em segundo lugar, muitos modelos estão sendo lançados por empresas emergentes fundadas por antigos talentos do Google (todos os autores do artigo dos *Transformers* deixaram a empresa e fundaram as americanas Adept, Near, CharacterAI e Inceptiv, a canadense Cohere e a japonesa Sakana AI), da subsidiária DeepMind (Inflection AI, Isomorphic Labs, Shift Lab, etc.)¹²⁵ e até mesmo da OpenAI (Anthropic, Covariant, etc.) (Benaich, 2023). Além dessas, um número crescente de *startups* que são financiadas não somente pelo venture capital tradicional, mas cada vez mais pelas próprias BigTechs.

Em terceiro lugar, o lançamento dos primeiros modelos da OpenAI e do Google

¹²⁵ DeepMind/Google perderam vinte pesquisadores nos últimos oito anos, como Arthur Mensch, cofundador da Mistral AI, Cyprien de Masson d'Autume, que cofundou a Reka AI e Michael Johanson, que cofundou a startup canadense Artificial.Agency, entre outros que migraram para Meta, HuggingFace e Anthropic (Smith, 2023)

disponibilizados por meio de licenças *open-source* entre membros da comunidade de “devs.”, constitui uma razão para a proliferação de modelos por novos *players*. Mesmo modelos originalmente lançados com código fechado, como GPT-3 e DALL-E (OpenAI) e AlphaFold1 (Google DeepMind), foram abertos em tempo recorde (Figura 3.3), enquanto outros foram lançados diretamente *open source*, como o OPT-175B (Meta), o Stable Diffusion (Stability AI) e o Alpa (desenvolvido pelo Google em parceria com Sky Lab, UC Berkeley, AWS, com código aberto para treinamento de redes neurais de grande escala). O vazamento do LLaMA (Meta AI), divulgado para *download* irrestrito no site 4chan poucos dias após seu lançamento (Vincent, 2023) foi certamente o divisor de águas. Em 2023, foram lançados 149 modelos de fundação, 65,8% deles em *open source* (44% em 2022 e 33% em 2021) (Maslej et al., 2024).

Figura 3.3 – Tempo em número de meses em que os modelos de referência GPT-3 (LLM da OpenAI), DALL-E (Modelo de imagem da OpenAI) e AlphaFold1 (Modelo de Proteínas da DeepMind) permaneceram fechados até serem implementados, clonados ou melhorados pela comunidade de código aberto



Fonte: Benaich e Hogarth (2022)

3.5. As incertezas na escala dos modelos: custos e grau de abertura

O avanço da IA generativa por novos *players* graças à disseminação de modelos pré-treinados *open source*, possibilitou que contornassem uma das principais barreiras à entrada do

setor que é o requisito de elevada capacidade computacional e crescentes custos de treinamento dos modelos, desenvolvendo modelos derivados e aplicativos para usos específicos, sem precisarem começar do zero (Küspert, Moës e Dunlop, 2023), automatizando o treinamento com algumas linhas de código (Zheng et al., 2022).

O cenário da IA começou a apresentar mudanças a partir de 2023 com tendências contraditórias. Primeiro, pela expansão descomunal no tamanho dos modelos, que evoluíram de unimodais, principalmente os LLMs, para modelos multimodais e a aposta das empresas líderes em modelos de fundação. Segundo, pela cisão entre as líderes do setor com respeito à abertura dos modelos e o debate disseminado acerca das abordagens de código aberto ou fechado dos modelos. Finalmente, pela agitação global relacionada aos modelos da IA generativa, envolvendo regulação, políticas antitruste, inclusão de salvaguardas em tratados internacionais e políticas “industriais” e comerciais relacionadas ao ecossistema (ou cadeia de valor) da IA. Esses aspectos e suas implicações são analisados no restante do capítulo e no seguinte.

3.5.1. A importância estratégica da capacidade computacional

Por mais tautológico que pareça, os sistemas de IA, enquanto sistemas computacionais, demandam recursos computacionais que crescem exponencialmente. O desempenho, a precisão e a habilitação de novas tarefas dos modelos de IA aumentam com a escala, refletindo as leis da escala.

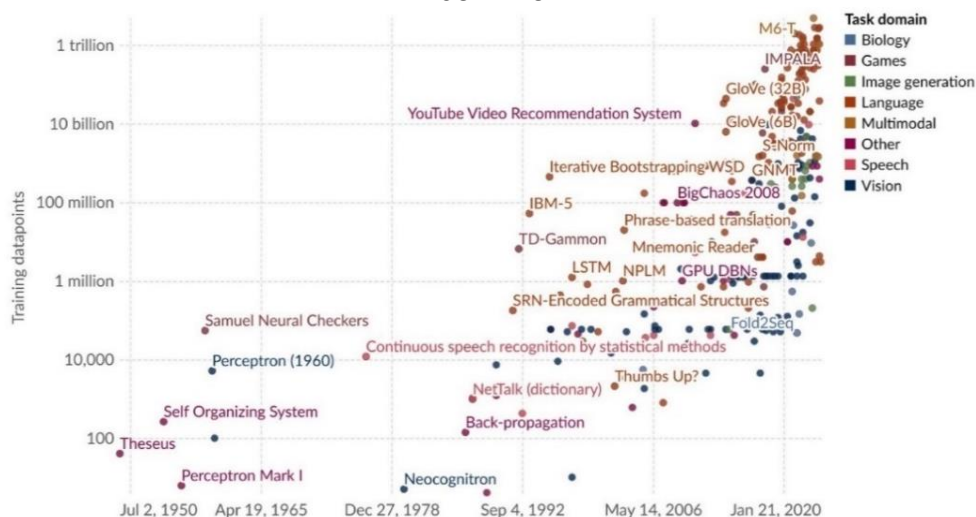
Os modelos são estruturados com número crescente de parâmetros e treinados com quantidades cada vez maiores de dados, o que requer extraordinário poder computacional. Os três gráficos a seguir mostram a evolução concomitante dessas três variáveis-chaves dos modelos de redes neurais profundas e, em particular, seu crescimento exponencial na fase da IA generativa. O Gráfico 3.5 mostra o crescimento do tamanho dos modelos em número de parâmetros, enquanto o Gráfico 3.6 a evolução na quantidade de dados usados no treinamento dos modelos, ao passo que o Gráfico 3.7 deixa claro o avanço em capacidade computacional, acelerando e mudando de patamar desde os Transformers com mudança na inclinação das retas da figura menor na parte superior esquerda - a reta azul correspondendo aos modelos não-Transformers, cujo crescimento em capacidade computacional em petaFLOPS (10^{15} operações de ponto flutuante) aumentou oito vezes a cada dois anos; a reta verde mostra que esse aumento passou para 275 vezes a cada dois anos.

Gráfico 3.5 - Crescimento no número de parâmetros dos modelos de IA no tempo, 1950-2023



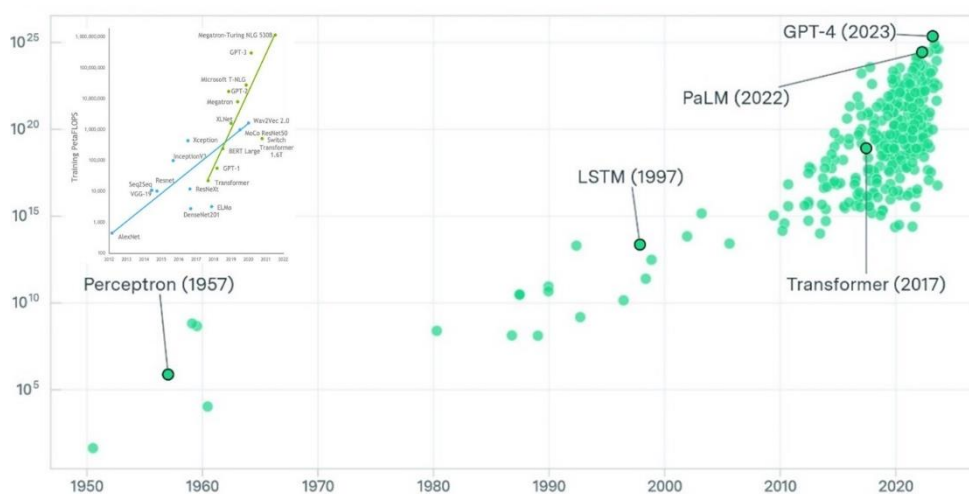
Fonte: Epoch AI

Gráfico 3.6 – Quantidade de dados usados para treinar sistemas de IA notáveis, por modelo e domínio



Fonte: Epoch apud OurWorldinData

Gráfico 3.7 - Crescimento na capacidade computacional, em petaFLOP, ao longo do tempo



Fonte: Epoch AI e Moore (2022)

3.5.2. Custos crescentes de treinamento dos modelos de IA

Modelos gigantes, com imenso número de parâmetros e amplos conjuntos de dados de treinamento mudaram a escala de recursos computacionais utilizados, com aumento expressivo dos investimentos exigidos. Azhar et al. (2024) detalham as diferenças entre a capacidade computacional exigida nos diferentes estágios de treinamento, ajuste fino e inferência. O treinamento dos grandes modelos, em particular de fundação, demandam grandes recursos computacionais, concentrados em grandes datacenters por exigir latência mínima e hardware especializado, consequentemente implicando em custos muito mais elevados, restritos às gigantes do setor e a umas poucas emergentes. O ajuste fino é computacionalmente menos intensivo, representando no máximo 10% dos custos de pré-treinamento de modelos de fundação. A inferência, que corresponde a “operação” do modelo, como no processo de perguntas e respostas de um chatbot como ChatGPT, demanda elevada capacidade computacional, que poderão futuramente superar até mesmo o treinamento em função da necessidade de dar suporte a um elevado número de usuários por longo período de tempo, embora não requeira hardware especializado, nem infraestrutura centralizada de datacenter, podendo ser realizada de forma distribuída. No presente momento, entretanto, as preocupações da corrida por modelos de grande escala recaem sobre o treinamento, mesmo que no longo prazo os maiores custos com hardware possam migrar para a inferência.

Valores exatos do custo de desenvolvimento (incluindo P&D, treinamento, hardware e gerenciamento) não são normalmente divulgados pelas empresas, mesmo com o custo de treinamento dos modelos apontado como a parcela que mais cresce. Na ausência de informações reais, um conjunto de trabalhos fazem estimativas desses custos. Não há uma métrica padronizada e os métodos de estimação são distintos, com graus de incerteza decorrentes das hipóteses feitas. No entanto, são indicadores relevantes para prever restrições econômicas ao progresso dos modelos e potenciais *players* capazes de arcar com os altos custos das corridas de treinamento, em que se destacam custos do hardware e de eletricidade durante sua execução.

Cottier (2023) detalha o cálculo do custo do hardware pela razão do tempo de treinamento (horas de GPUs usadas por unidade de hardware para treinamento, em geral semanas ou meses) e do número total de horas de GPU que a unidade de hardware é usada antes de ser substituída por novo hardware, multiplicada pelo custo do hardware pelo tempo de sua amortização. Os custos de energia, por seu turno, são estimados pelo tempo de treinamento do número de hardware multiplicado pelo consumo de energia do hardware (em kW) e a taxa de energia (valor por kWh) - consumo máximo de energia listado na ficha técnica dos chips.

Outra alternativa de mensuração é o preço da computação em nuvem – que inclui custos

de hardware, energia e manutenção não individualizados, além do lucro adicionado pelo fornecedor - expressos em valor por hora de uso e por unidade de hardware. O custo de treinamento é, assim, calculado pelo número de horas de GPU usadas multiplicadas pelo número de unidades de GPUs e o preço da computação na nuvem, que corresponde ao limite superior do custo para empresas menores contratarem infraestrutura de terceiros.

Maslej et al. (2024) apresentam os custos de computação de treinamento estimados para os dez principais modelos de IA (Tabela 3.3), permitindo visualizar sua significativa expansão (Gráfico 3.8) e os valores impressionantes dos modelos mais recentes, GPT-4 e, principalmente, Gemini. Rahman, Owen e You (2024) rastrearam na base da Epoch os modelos de IA com uso intensivo de computação no treinamento (acima de 10^{23} operações de ponto flutuante) e, igualmente, mostraram elevados custos.

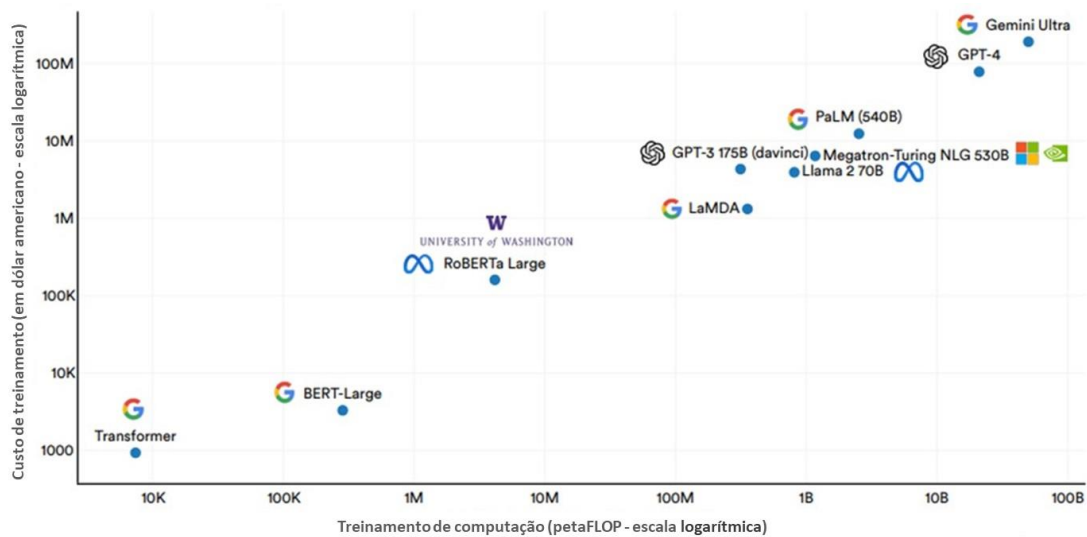
Cottier et al. (2024) decompuseram os custos de treinamento para quatro importantes modelos de IA (Gráfico 3.9). Esses custos isolados - com P&D, chip acelerador, outros componentes do servidor, cabos de interconexão do *cluster* de hardware e energia - mostram diferenças interessantes entre os modelos, como os maiores custos de P&D do Gemini e do OPT; custos menores com chips, outros componentes de servidor e interconexão do Gemini, possivelmente resultado das internalizadas unidades de processamento (TPUs) do Google; e os maiores custos de energia dos modelos da OpenAI, em especial do GPT-4.

Tabela 3.3 – Custo de treinamento estimado de modelos de IA selecionados

Modelo	Criador	Lançamento	Custo (em US\$)
Transformer	Google	2017	930
BERT Large	Google	2018	3.288
RoBERTa Large	Meta AI	2019	160.018
GPT-3 175B (davinci)	OpenAI	2020	4.324.883
Megatron-Turing NLG 530B	Microsoft/Nvidia	2021	6.405.653
LaMDA	Google	2022	1.319.583
PaLM (540B)	Google	2022	12.389.056
GPT-4	OpenAI	2023	78.352.034
LLaMA2 70B	Meta AI	2023	3.931.897
Gemini Ultra	Google	2023	191.400.000

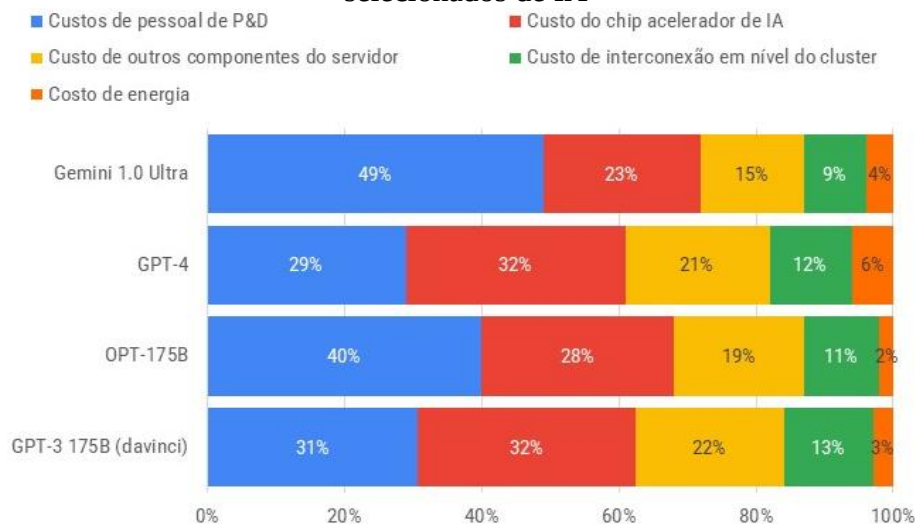
Fonte: Epoch (2023) apud Maslej et al. (2024)

Gráfico 3.8 – Custo estimado de treinamento e uso de computação, modelos de IA selecionados



Fonte: Maslej et al. (2024)

Gráfico 3.9 - Desagregação dos custos de treinamento e desenvolvimento, modelos selecionados de IA



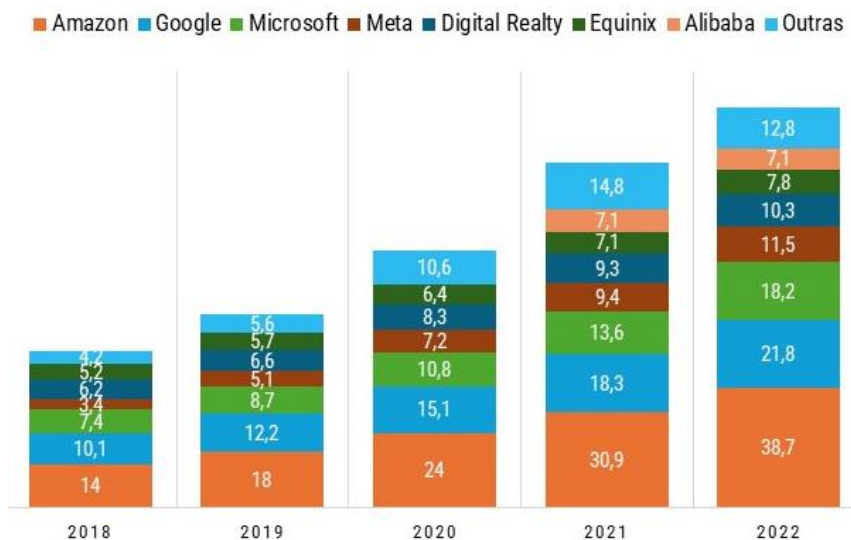
Fonte: Cottier et al. (2024)

Os enormes recursos computacionais usados no treinamento dos modelos têm levantado preocupações relacionadas ao seu impacto ambiental, tanto pela emissão de CO², quanto pelo consumo de água para resfriamento dos equipamentos (Maslej et al., 2023). O consumo de água dobrou no LLaMA2 (Meta) comparado ao LLaMA1, aumentou 34% para resfriamento dos datacenters da Microsoft que executam o ChatGPT, cujo consumo é de 500 ml para uma conversa de 20-50 perguntas/respostas (Li et al., 2023).

Relatório recente da agência das Nações Unidas sobre Comércio e Desenvolvimento disponibilizou um conjunto amplo de informações sobre os impactos ambientais da digitalização, com recomendações para uma digitalização ambientalmente sustentável e inclusiva. A evolução do consumo de eletricidade, medido em terawatt/hora, pelas principais

empresas de datacenters – BigTechs americanas, Alibaba e empresas gestoras de datacenters, como Digital Realty e Equinix – são apresentados no Gráfico 3.10.

Gráfico 3.10 - Consumo anual de eletricidade por empresas de data centers, em terawatt hora, 2018–2022



Fonte: UNCTAD (2024)

Nota: Outros inclui provedoras chinesas de serviços de datacenters e Tencent, Apple e Baidu

Os dados mostram que o consumo mais que dobrou entre 2018 e 2022, com destaque para Amazon, maior operadora de centros de processamento de dados do mundo, junto a Microsoft e ao Google, os líderes globais em computação na nuvem. A maioria da energia relacionada a data centers é consumida por servidores (80%), seguidos por dispositivos de armazenamento (18%) e equipamentos de rede (3%), enquanto a maior parte do uso de energia da infraestrutura é destinado ao resfriamento, influenciando esforços das empresas no uso de técnicas mais eficientes de resfriamento, localização de data centers em países de clima frio (devido ao menor aquecimento e a possibilidade de uso de ar livre), além de abordagens inovadoras (data centers subaquáticos, softwares para melhoria de eficiência energética do hardware e aquisição de energia renovável pelos maiores operadores de data centers em contratos de longo prazo que garantem a oferta e os preços) (UNCTAD, 2024). Os maiores operadores de data centers compram energia renovável, principalmente compromissos de longo prazo, visando reduzir emissões. Amazon, Google, Microsoft e Meta são os maiores compradores de energia renovável nos EUA.

A eficácia do uso de energia (PUE, sigla de *power usage effectiveness*) é a medida de quão eficientemente um data center usa energia. A PUE média global em data centers é hoje cerca de 1,6 (para cada 1,6 kWh de eletricidade usada, 1 kWh é usado para equipamentos de TI e 0,6 kWh para resfriamento e outros equipamentos não TI). A PUR dos data centers do

Google¹²⁶ e Meta, os mais energeticamente eficientes do mundo, é de cerca de 1,1 (UNCTAD, 2024).

Goldman (2024) esclarece que os impactos ambientais não se limitam aos produtores de IA, afetando toda a cadeia de produção/ecossistema, inclusive na produção de chips, cujas informações não divulgadas pelas empresas, além dos danos ambientais da exploração de minerais estratégicos e geração de lixo eletrônico. A produção de chips requer quantidades enormes de energia e água – a TSMC, utiliza quase 5% de toda a eletricidade do país e libera resíduos perigosos. A mineração do silício a partir da retirada de areia pode provocar danos ambientais pela intensificação da erosão dos solos (Belton, 2021). Todas essas preocupações estão levando a medidas legais, como a proposta do Artificial Intelligence Environmental Impacts Act¹²⁷, em 2024, por senadores americanos, propondo o desenvolvimento de padrões de mensuração/acompanhamento dos impactos ambientais da IA.

Com efeito, as empresas estão investindo em hardware eficiente e com menor impacto ambiental. O modelo gigante M6-10T da Alibaba foi anunciado como o primeiro a reduzir significativamente o consumo de energia no treinamento (Lin et al., 2021). O GLaM-1,2T, do Google, sete vezes maior que o GPT-3 (175 bilhões de parâmetros), consome um terço da energia no treinamento e requer metade dos FLOPS de computação (Du et al. 2022). Google e NVIDIA projetam chips otimizados para supercomputação e IA que reduzem substancialmente o consumo de energia.

3.5.3. Divulgação dos modelos em regime *open-source*

A divulgação extensiva dos modelos LLaMA e OPT-175B se tornou o principal pilar da expansão da Meta, estratégia que começa a colidir com outras gigantes da IA e emergentes notórias que retrocederam na prática de divulgar detalhes técnicos dos modelos, omitindo informações sobre arquitetura, hardware, códigos, dados e até número de parâmetros nos artigos sobre modelos recentes, cada vez revestidos de maior sigilo. As informações disponibilizadas do GPT-4 foram drasticamente reduzidas frente às versões anteriores, justificado por razões do ambiente competitivo e segurança (OpenAI, 2023), como no PaLM2, cujas informações divulgadas são reduzidas comparado ao PaLM1, do Google (Goldman, 2024). A publicação de artigos de pesquisa da Alphabet/Google passou a ocorrer somente depois do trabalho em laboratório se transformar em produtos, desde que documento interno mostrou apreensão da

¹²⁶ Além disso, o ASIC personalizado do Google é 30–80 vezes mais eficiente em termos de energia do que unidades de processamento central de uso geral

¹²⁷ Disponível em: https://www.markey.senate.gov/imo/media/doc/artificial_intelligence_environmental_impacts_act_of_2024_-_020124pdf.pdf. Acesso em: 05 set.2024

empresa com a disseminação de modelos menores *open-source* com custo infinitamente inferior, como o Vicuna-13B, desenvolvido a partir do LLaMA (Patel e Ahmad, 2023).

Algumas análises questionam a qualificação desses modelos como *open source*, considerando o termo mal definido e empregado erroneamente para descrever modelos com graus de abertura e transparência muito distintos. O LLaMA-2 (Meta) é considerado fechado para os padrões *open-source* do software convencional. Pode ser baixado, modificado e implantado gratuitamente, mas impõe restrições ao uso comercial para certos usuários e uso do modelo para determinados fins, como treinar outros modelos de linguagem, casos em que exige licença especial para aplicativos/serviços com mais de 700 milhões de usuários/dia (LLaMA Community License Agreement).¹²⁸ Por outro lado, modelos GPTNeo e Pythia (Eleuther), são considerados com máxima abertura (Maffulli, 2023; Widder, West e Whittaker, 2023; Kak, West e Whittaker, 2023).

O *open source* seguiu uma trajetória de internalização típica das BigTechs, que mesmo disponibilizando os modelos, não incluem a totalidade das informações relevantes e vedam seu uso para fins comerciais, bloqueando competidores potenciais. Widder, West e Whitakker (2023) alertam para a vulgarização do termo e o mito dos *benefícios da IA em open source*, pois muitos desses modelos são apenas “*open for business*”. Há uma variedade de situações, com grandes diferenças no grau de abertura dos modelos. A efetiva abertura do modelo requer disponibilização de informações sobre parâmetros, bases de dados e seu tratamento, treinamento, algoritmos, calibração e avaliação, código efetivamente aberto e reusável, como nas licenças permissivas como Apache2¹²⁹.

Há níveis muito díspares de abertura das informações dos modelos, conforme as categorias mostradas na Figura 3.5, como apontado em vários estudos. Bommasani et al. (2023a), por exemplo, desenvolveram um índice de transparência dos modelos, constatando níveis distintos de “abertura”¹³⁰ e transparência menor do que se supõe, como Gemini 1.0 Ultra (47% de abertura), GPT-4 (49%) e LLaMa2 (60%), nem mesmo por empresas emergentes, com o Mistral 7B (55%) e o Claude3 (51%). Xiang (2023) argumenta que além de dificultar o avanço do conhecimento e o aprendizado pelos “seguidores”, a não divulgação de detalhes técnicos impede a identificação de vieses, riscos e custos ambientais, comprometendo a avaliação de impactos. Rikap (2023b) questiona a real capacidade dos modelos *open-source* constituírem alternativas reais, argumentando que os LLM *open-source* são menos abrangentes

¹²⁸ Disponível em: <https://github.com/facebookresearch/llama/blob/main/LICENSE>. Acesso em: 12 set.2024

¹²⁹ Disponível em: <https://www.apache.org/licenses/LICENSE-2.0>. Acesso em: 12 set.2024

¹³⁰ Disponível em: <https://crfm.stanford.edu/fmti/May-2024/index.html>. Acesso em: 12 set.2024

(menor número de parâmetros) do que os modelos proprietários, não podem ser usados diretamente (demandando desenvolvimento de interface com o usuário) e exigem adaptação para atender necessidades de terceiros *downstream*, o que demanda capacidades e talentos escassos.

Figura 3.5 – Nível de acessibilidade e estratégias de liberação de modelos de fundação

Considerações						
	Somente pesquisa interna Controle de alto risco Baixa auditabilidade Perspectivas limitadas					Pesquisa comunitária Controle de baixo risco Alta auditabilidade Perspectivas mais amplas
Nível de acesso	Totalmente fechado	Liberação gradual/escalada	Acesso hospedado	Fechado ao público Baseado em nuvem/Acesso API	Para download	Totalmente aberto
Sistema (desenvolvedor)	PaLM (Google) Gopher (DeepMind) Imagen (Google) Make-A-Video (Meta)	GPT-2 (OpenAI) Stable Diffusion (Stability AI)	DALL-E (OpenAI) Midjourney (Midjourney)	GPT-3 (OpenAI)	OPT (Meta) Craiyon (Craiyon)	BLOOM (BigScience) GPT-J (Eleuther AI)

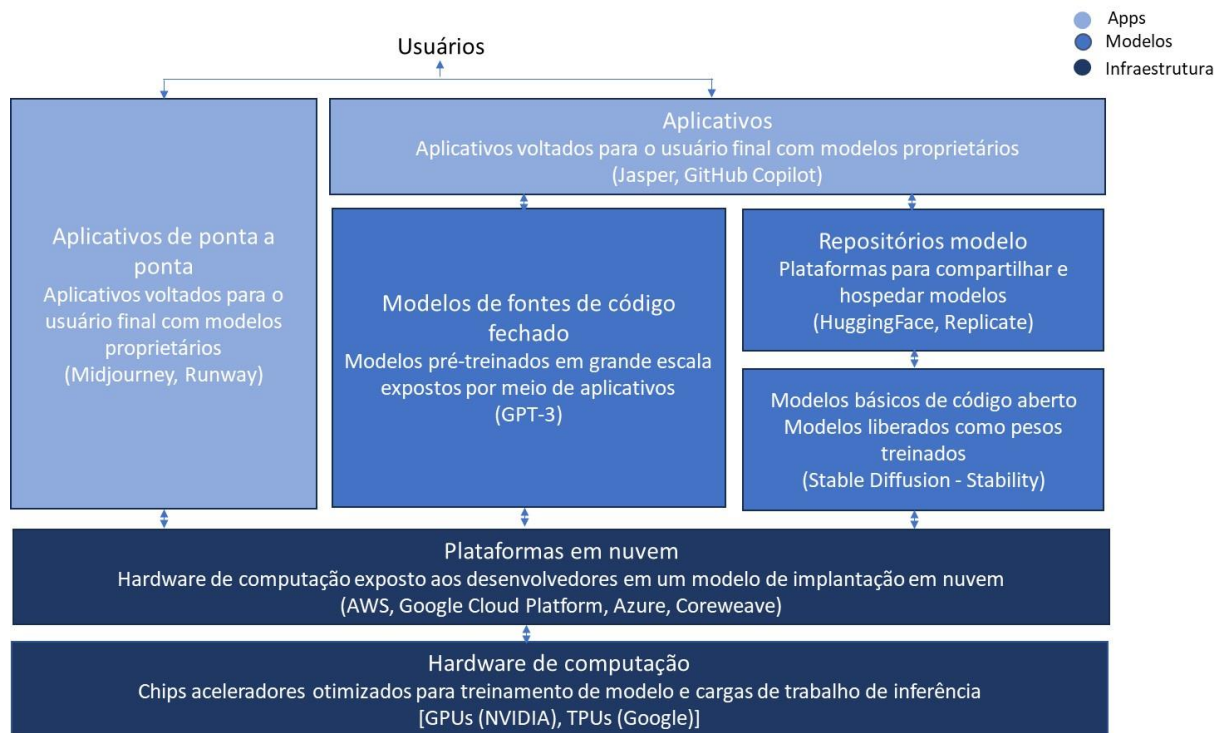
Fonte: Solaiman (2023)

Widder, West e Whitakker (2023) observam que os recursos necessários para construir do zero e implantar grandes sistemas de IA em escala permanecem “fechados”, pois mesmo “abrindo” parcialmente os modelos, as infraestruturas em que operam, como PyTorch/Meta e Tensorflow/Google, são proprietárias e não interoperáveis, até mesmo no framework do HuggingFace, que oferecem ferramentas e bibliotecas de IA amplamente acessíveis, em que “devs.” se especializam em modelos atrelados a essas infraestruturas e respectivos produtos proprietários. Os autores citam o caso da Meta, que aproveita o HuggingFace como acelerador de startups para estimular a adoção de seus modelos de código aberto no programa de startups no campus *Station Fem Paris* pela *Scaleaway*. Rikap (2021) e Rikap e Lundvall (2021) já faziam referência ao sigilo dos sistemas de IA por meio da metáfora do quebra-cabeças, pela qual as BigTechs fornecem parte das peças, mas mantêm em sigilo a peça central.

A pilha da tecnologia da IA generativa, discriminada na Figura 3.6, expressa a principal assimetria entre empresas emergentes e BigTechs em aspectos do hardware/computação na nuvem, o que afeta todos os modelos (fechados ou abertos) e aplicativos desenvolvidos por empresas que não dispõem de ativos de infraestrutura.

É no desenvolvimento de modelos multimodais - predominantemente resultantes de retreinamento dos LLMs com dados de múltiplos domínios – e de fundação que a assimetria entre líderes incumbentes e startups e organizações de código aberto se torna mais evidente, em função dos recursos computacionais extremos demandados por esses modelos; uma assimetria em quantidade e eficiência em escala. A divisão entre incumbentes e emergentes se transformou em uma assimetria por porte em GPUs.

Figura 3.6 – Pilha de tecnologia IA generativa



Fonte: Mateos-Garcia (2023)

3.6. Os ativos tangíveis da IA: onde o vencedor leva tudo

Os crescentes requisitos de hardware dos modelos criam grandes desafios para as empresas emergentes de menor porte, causando perplexidade sobre como essas empresas conseguiram dispor de recursos computacionais e arcar com os elevados custos da atual corrida da IA.

Esta seção procura tratar especificamente dos ativos tangíveis da IA. Esses ativos compreendem a intensidade de computação exigida pelos modelos, incluindo chips personalizados de IA, data centers e cabos submarinos, computação na nuvem, que parecem caracterizar um mercado em que o “vencedor leva tudo” (Frank e Cook, 2013). A compreensão dessa situação, entretanto, requer uma dimensão geopolítica do que apenas fatores econômicos.

3.6.1. O complicado quebra-cabeça geopolítico da cadeia de valor dos semicondutores

A produção de semicondutores¹³¹ foi reorganizada na década de 1990 em cadeias de valor globais caracterizadas pela dispersão geográfica das várias etapas de produção envolvidas.

¹³¹ Semicondutores (também conhecidos como circuitos integrados, chips microeletrônicos ou chips de computador) são pequenos dispositivos eletrônicos (baseados principalmente em silício ou germânio) compostos de bilhões de componentes que podem processar, armazenar, detectar e mover dados ou sinais. Existem vários tipos de chips semicondutores – incluindo lógicos, memória, analógicos, optoeletrônicos, sensores e discretos – que executam funções diferentes e exigem projeto e fabricação especializados. Os chips lógicos, que respondem pela maior parcela das vendas globais de semicondutores (42%), incluem as unidades centrais de processamento (CPUs) usadas para computação de uso geral, e as unidades de processamento gráfico (GPUs) amplamente usadas nos modelos da IA (Grossmn, Blevins e Sutter, 2023).

Empresas sediadas nos EUA lideram o *design* de chips lógicos (64% da receita global, em 2021), empresas que não fabricam, apenas projetam os chips, chamadas “*fabless*”, com destaque para AMD, Qualcomm, Broadcom e, principalmente, Nvidia. As instalações de produção foram terceirizadas para fundições localizadas em outros países, como Taiwan e mesmo China. As empresas de *design* de chips lógicos projetam chips de ponta, com “nós” de 3 e 5 nanômetros (nm), ao passo que empresas integradas como a americana Intel e a fundição chinesa *Semiconductor Manufacturing International Corporation* (SMIC)¹³² estariam nas gerações anteriores de nós (chips para a Huawei, de 7 nm).

As fabricantes contratadas (fundições ou Fabs) respondem por mais de 80% da capacidade mundial de produção de chips lógicos, com destaque para a taiwanesa TSMC (*Taiwan Semiconductor Manufacturing Company*), primeira fundição pura no mundo, que fabrica a partir de *design* de terceiros (Chiao, 2022), detendo metade da capacidade global de fundição, em 2022 (Khan, Mann e Peterson, 2021). A fabricação nos EUA, com somente 7% da capacidade mundial, ocorre por produtores de dispositivos integrados (IDMs, sigla de *integrated device manufacturers*)¹³³, como Intel, *Global Foundries* e *Samsung Foundries*.

Os Estados Unidos lideram em *design*, equipamentos e softwares dedicados dos chips de IA, com a produção na Ásia, que responde por 70% da produção mundial, com destaque para Taiwan, Coreia do Sul e China no caso dos chips lógicos.

Após a fabricação, os chips são enviados em *wafers* para outras instalações para finalização de montagem, teste e embalagem (ATP, sigla de *Assembly, Test, and Packaging*) por fabricantes contratados, que cortam, testam o desempenho e embalam de forma a proteger o chip e permitir sua integração em dispositivos eletrônicos acabados, anexando as interconexões elétricas. A capacidade de produção de ATP localiza-se na China (38%), Taiwan (19%) e Coreia do Sul (9%), com 5% nos EUA, embora avanços técnicos em embalagens estejam promovendo o *reshoring* americano. As instalações ATP exigem uma variedade de matérias-primas e materiais processados, produtos químicos (fluoreto de hidrogênio e ácido nítrico) e gases hélio e néon (a Ucrânia fornece 90% do gás neon usado na fabricação de

¹³² Os chips de memória que armazenam dados (memória dinâmica de acesso aleatório - DRAM e flash NAND), são produzidos em massa, *commoditizados*, envolvendo economias de escala e margens de lucros mais estreitas, com competição por preço, não por potência e desempenho. Sua principal aplicação é no gerenciamento de energia (fabricados com materiais semicondutores compostos, como carboneto de silício (SiC) e nitreto de gálio (GaN)).

¹³³ As IDMs são empresas integradas de semicondutores que projetam, fabricam e vendem circuitos integrados, modelo prevaiente até as cadeias de valor globais, quando fábricas contratadas (fundições) passaram a fabricar chips a partir do projeto das empresas “*fabless*”. Hoje poucas empresas, como Samsung, fabricam a partir de seu próprio *design*, além de fornecer serviços de fundição para outras empresas (Palma et al., 2022). A Intel estabeleceu a Intel Foundry Services, em 2021, que ao invés de produzir apenas *wafers* como o modelo tradicional de fundição, fornece silício, embalagens, software e chips.

semicondutores nos EUA e cerca de 50% do mercado global, o que explica a escassez de oferta e aumento de preços após a invasão russa, tal como em 2014).

Peça importante da cadeia de produção de chips são equipamentos complexos, fabricados por poucas empresas, americanas e também japonesas e europeias. As máquinas de fotolitografia¹³⁴, componente essencial da produção de semicondutores, são monopólio da empresa holandesa ASLM, enquanto outros equipamentos são fabricados por empresas japonesas ou chinesas (máquinas de manuseio de *wafer*, para embalagem, montagem e testes). Apesar de empreender uma política industrial ativa e forte financiamento governamental para produção local de chips e sua cadeia de valor, a China ainda é dependente de chips e equipamentos dos Estados Unidos e países aliados, ainda que avanços importantes tenham sido alcançados pela fundição chinesa SMIC e pela fabricante de equipamentos (litografia imersão DUV) *Shanghai Micro Electronics Equipment* (SMEE).

3.6.2. Processamento paralelo e chips de IA

Com o aumento na escala dos modelos, o progresso técnico da IA passou a depender de avanços em três elementos básicos: arquiteturas que possam rodar em paralelo; disponibilidade de grandes quantidades de dados para treinamento; e hardware. Os avanços no hardware computacional subjacente aos modelos correspondem a quantidades significativas de *clusters*¹³⁵ de semicondutores, circuitos integrados ou simplesmente chips, que envolvem cadeias de fornecimento amplas.

Existem tipos diferentes de semicondutores e aqui interessam particularmente os chamados chips para IA ou aceleradores, que oferecem maior poder de processamento (Khan e Mann, 2020). Os chips de IA, como são chamados (Harris, 2023), que são chips especializados que incorporam tecnologia de IA e são usados para aprendizado de máquina, compõem um mercado com grande potencial de crescimento (de US\$ 14,9 bilhões em 2022 para US\$ 383,7 bilhões até 2032, segundo a Allied Market Research)¹³⁶.

¹³⁴ Máquinas de fotolitografia são usadas para padronizar bilhões de características eletrônicas em um *wafer* de silício, usando diferentes tipos de luz que passam através de uma máscara até o *wafer* subjacente que é coberto por materiais sensíveis à luz. À medida que os recursos eletrônicos se tornam menores em “nós” de processo mais avançados, são necessários comprimentos de onda de luz menores. Em nós mais avançados (3 e 5 nm), a luz ultravioleta extrema (EUV) é usada para padronizar circuitos no *wafer*, enquanto em os nós mais maduros (acima de 7 nm) usam luz ultravioleta profunda (DUV). Para alcançar êxito nos equipamentos de fotolitografia EUV foram necessárias décadas de P&D pela empresa holandesa *Advanced Semiconductor Materials Lithography* (ASML) (Khan, Mann e Peterson, 2021:30-35).

¹³⁵ Um *cluster* de GPU consiste em uma coleção de dispositivos de computação interconectados, abarcando hardware e software. O hardware especializado foi projetado para facilitar cálculos paralelos de alto desempenho, capaz de lidar com tarefas complexas como processamento de imagem e vídeo ou treinamento de redes neurais artificiais em grande escala. Ao combinar várias GPUs em um *cluster*, os processos computacionais são significativamente acelerados.

¹³⁶ Disponível em <https://www.alliedmarketresearch.com/artificial-intelligence-chip-market>. Acesso em 25 ago. 2024.

As arquiteturas de redes neurais em que se baseiam os modelos de fronteira de IA, particularmente Transformers, exigem cálculos computacionais gigantescos para processamento e reproprocessamento em várias camadas de massivos conjuntos de dados e demandaram avanços tecnológicos concomitantes nos sistemas computacionais e chips especializados. À medida em que o tamanho dos modelos e o volume de dados de treinamento aumentam espetacularmente, sistemas mais eficientes se tornaram essenciais para resolver problemas matemáticos e computacionais.

Com isso, a IA generativa está levando especialistas a considerar a evolução em capacidade computacional como pedra angular dos modelos de fronteira. Os primeiros modelos usavam chips lógicos comerciais prontos para uso, as tradicionais unidades centrais de processamento (CPUs, sigla de *Central Processing Unit*) nas etapas de treinamento e inferência dos modelos. As CPUs são ainda satisfatórias nas operações de inferência, embora não atendam mais às operações de treinamento, que passaram a usar unidades de processamento gráfico (GPUs - *Graphic Processing Units*). Originalmente concebidos para aprimorar interface gráfica, foram reaproveitados para uso em modelos de IA (Owens et al., 2008). Enquanto as CPUs processam informações em série, sequenciais, as GPUs possibilitaram o processamento paralelo, permitindo treinar mais rapidamente os modelos, dividindo as tarefas e executando-as simultaneamente (Ford 2021; Ferràs-Hernández, Nylund e Brem, 2023). O Google foi pioneiro no desenvolvimento das *Tensor Processing Units* (TPUs), que são Circuitos Integrados Específicos de Aplicação (ASICs), ou aceleradores, poderosos processadores personalizados destinados a acelerar o treinamento de redes neurais em sua estrutura específica (TensorFlow)¹³⁷. Além de recentes unidades de processamento, como as Language Processing Units (LPUs), desenvolvidas pela Groq, chips otimizados para NLP e lidar com os padrões singulares de velocidade e memória da inferência em LLMs.

Os chips lógicos usados para aplicações de IA são cada vez mais personalizados, projetados pelas empresas de *design* para aplicações específicas, como treinamento dos modelos de forma otimizada e eficiência energética. Abrangem uma variedade de arquiteturas específicas, como GPUs, ASICs, FPGAs (*Field Programmable Gate Arrays*), etc., para processamento otimizado de operações específicas de IA, como classificação de imagens, tradução automática, detecção de objetos, entre muitas outras (Khan e Mann, 2020). O treinamento dos modelos exige conectar muitos chips lógicos em *clusters* com outros

¹³⁷ O Google desenvolveu as TPUs em 2015, hardware proprietário visando acelerar o treinamento de redes neurais, usadas internamente e, a partir de 2018, disponibilizadas para terceiros, tanto na sua infraestrutura na nuvem, quanto em chips para venda.

semicondutores (por exemplo, chips de memória e rede) em grandes data centers ou instalações de supercomputação.

A principal característica desses chips é a computação paralela exigida pelas operações de redes neurais profundas, que envolvem grande número de operações de multiplicação de matrizes que independem dos resultados dos demais cálculos. É possível aumentar a paralelização conectando vários chips de IA em paralelo – a chamada paralelização do modelo (dividido em várias partes e cálculos realizados simultaneamente) - ou otimizar usando computação de baixa precisão – que sacrifica a precisão numérica em prol da velocidade e da eficiência -, ou ainda otimização de memória. As estruturas TensorFlow e Pytorch possuem recursos que podem aproveitar as vantagens dos recursos dos chips de IA.

O avanço dos modelos de fronteira de IA está sendo alcançado a partir de dois caminhos: a construção de *clusters* gigantescos de chips e, em especial, o desenvolvimento de chips especializados. Por um lado, a corrida pela inovação da IA se converte em uma corrida pela aquisição de hardware GPUs desde 2023 e previsão de mais de 3,5 milhões de H100 (um dos principais chips Nvidia) até o final de 2024. A corrida pela construção de *clusters* de treinamento de modelos de mais de 100 mil GPUs, ao custo individual de US\$ 4 bilhões com despesas de capital de servidores e mais de 150 MW de capacidade de data center que consome 1,59 terawatt-hora/ano, ao custo de US\$ 123,9 milhões (taxa padrão de US\$ 0,078/kWh) (Patel e Nishball, 2024b). A construção desses *clusters* e infraestrutura de treinamento de IA envolve não apenas recursos financeiros, mas alta complexidade e desafios de energia, pontos de verificação, opções de topologia de rede, esquemas de paralelismo, *layouts* de *rack* e materiais.

A Meta, por exemplo, em fevereiro de 2024, explicitou sua estratégia de longo prazo em IA no documento de divulgação de resultados no último trimestre de 2023¹³⁸, onde destaca a importância de construir infraestrutura de computação de *clusters* avançados de classe mundial, anunciando a aquisição de 350 mil chips H100 e outras GPUs (um total de 600 mil equivalentes de H100 de computação) até o final do ano. Um investimento agressivo que colocará a Meta entre as maiores operadoras de GPUs do mundo, segundo o State of AI Report Compute Index v3 (2024).

Supercomputadores desenvolvidos para uso exclusivo em modelos de IA estão entre os maiores do mundo, como o já mencionado Selene, da Nvidia, o investimento de US\$ 100 bilhões no Stargate pela Microsoft (com OpenAI) (White, 2024) e *clusters* de 100 mil GPUs para sua equipe interna (cinco vezes maior que o GPT-4) (Patel e Nishball, 2024c), a plataforma

¹³⁸ Disponível em: https://s21.q4cdn.com/399680738/files/doc_financials/2023/q4/META-Q4-2023-Earnings-Call-Transcript.pdf. Acesso 03 fev.2024

Blackwell (Nvidia, 2024) e o investimento da xAI na construção de um supercomputador com 100 mil GPUs para treinar e executar as novas versões do modelo Grok.

Por outro lado, seguindo o pioneirismo do Google, que desenvolveu hardware especializado para modelos da IA, esta é a estratégia das líderes em modelos generativos¹³⁹ e se atrela ao desenvolvimento de chips por empresas incumbentes e *startups* especializadas (Nvidia, AMD, Intel, Cerebras). O AWS Inferentia e o Trainium (Amazon)¹⁴⁰, o MTIA (Meta) (ainda não em uso), o Ali-NPU (Alibaba), o Kunlun (Baidu), o Neural Engine (Apple), o Gen (ByteDance) e o Ascend (Huawei) são exemplos de chips proprietários personalizados e integrados ao negócio de IA por um número crescente de empresas. A Microsoft, além do Braga e Maia 100, anunciou o novo chip Athena, em 2024, o projeto Brainwave (Jaokar, 2022; Patel, 2024a). As parcerias surpreendem, como da *designer* americana de chips, Broadcom, e a chinesa ByteDance (que estocou US\$ 2 bilhões em chips Nvidia em 2023, além dos chips Ascend 910B da Huawei) (Baptista, 2024). As iniciativas envolvem até desenvolvimento de chips que incorporam redes neurais, para aplicativos de visão computacional em tempo real, pela Mobileye, Intel, Qualcomm e Samsung (Penn Engineering Today, 2022), além de método para projetar chips usando técnicas de IA (Lin, 2023), com o AlphaDev, da DeepMind (Google).

Burkacky et al. (2024) atribuem a produção de chips pela indústria de IA à demanda crescente provocada pela proliferação de modelos da IA generativa, levando empresas especializadas a compartilhar o mercado de chips de IA com grandes criadores de modelos e algoritmos, que passaram também a inovar em hardware, investir na expansão de data centers e servidores dedicados e *design* de chips próprios, materiais e arquiteturas para atender necessidades específicas da IA. De fato, a TSMC só consegue atender 80% da demanda por chip das GPUs A100 e H100 da Nvidia (Robinson, 2023), com pressão permanente sobre preços. O investimento em chips próprios reduziu substancialmente os custos de computação do Google, Meta e Amazon, o que explica a corrida frenética da OpenAI para desenvolver seus próprios chips e, com isso, deixar de gastar US\$ 12 bilhões com chips Nvidia, previsto para 2024 (Vipra e West, 2023).

¹³⁹ Google e Amazon foram precursoras em chips personalizados (e computação na nuvem), além de esforços contínuos de CPUs de servidores internos, inicialmente voltados para redução de custos e melhoria de desempenho interno, mas permitiram preços mais baixos e margens mais altas que concorrentes. As arquiteturas de hardware dos principais *players* é tratada em vários estudos da semianalysis.com (Patel, Cozma e Wong, 2023; Patel, 2023).

¹⁴⁰ Chips proprietários são projetados para “tornar mais fácil para os desenvolvedores integrar o aprendizado de máquina em seus aplicativos de negócios” (AWS, 2022a), aprimorando serviços/produtos da empresa e disponibilizados para devs, mais uma vez demonstrando a natureza dupla da produção de valor para as empresas líderes em IA, que empregam recursos computacionais internamente e os monetizam externamente (Ferrari, 2023).

A movimentação recente da OpenAI na busca de parceiros e financiamentos ao redor do mundo para desenvolver chips personalizados evidencia o gap entre o investimento exigido e a capacidade financeira das empresas emergentes. Uma dessas negociações foi com a G42, empresa de IA dos Emirados Árabes com atuação em serviços financeiros, energia, saúde e serviços públicos, visando desenvolvimento de chips de IA no país árabe e que mobilizaria financiamentos entre US\$ 10 e US\$ 100 bilhões (do governo emiradense, do SoftBank Group Corp japonês¹⁴¹ e outros parceiros), além de criar mercado potencial para a OpenAI. A parceria, entretanto, enfrentou resistências do governo americano pelos vínculos da G42 com a China, tendo recebido investimentos (US\$ 1,5 bilhão) da Microsoft, levando deputados republicanos a solicitar informações sobre essas transações e vínculos ao governo Biden¹⁴².

As negociações da OpenAI com a empresa americana RainAI (Rain Neuromorphics), criada em 2017 visando desenvolver uma neuromorphic processing unit (NPU)¹⁴³, que replicaria características do cérebro humano com a promessa de cem vezes mais poder de computação e eficiência energética do que os processadores atuais (Dave, 2023), também esbarrou em restrições do governo americano¹⁴⁴. Um acordo não vinculativo foi firmado em 2019, com compromisso de investimento de US\$ 51 milhões da OpenAI e US\$ 1 milhão diretamente por Sam Altman).

A busca da OpenAI por recursos para investimentos em chips inclui também fundos soberanos saudita e emiradense (Ludlow e Vance, 2023)¹⁴⁵, como o Fundo de Investimento Público (PIF) da Arábia Saudita, Mubadala Investment Company entre outros, de forma a superar esse importante gargalo importante na cadeia de suprimentos de IA e reduzir sua atual dependência de terceiros (incluindo Microsoft) (Vipra e West, 2023).

¹⁴¹ Notícias recentes indicam investimento do SoftBank será de US\$ 500 milhões (Clark e Weinberg, 2024)

¹⁴² A G42 teria cooperado no desenvolvimento de testes e vacinas para Covid com a empresa chinesa de biotecnologia BGI Genomics (acusada de tentar coletar dados genéticos e biométricos de cidadãos americanos por meio de testes COVID doados ao estado de Nevada), além de deter US\$ 100 milhões em ações da ByteDance, proprietária do TikTok (acusado de espionagem de cidadãos americanos) (CPI, 2024). Disponível em: <https://www.pymnts.com/cpi-posts/republican-lawmakers-demand-intelligence-briefing-on-microsofts-ai-investments/>. Acesso 20 ago. 2024

¹⁴³ A computação neuromórfica e a IA tradicional diferem em termos de abordagem ao processamento de informações, pois enquanto essa última depende de algoritmos e regras para processar informações, a neuromórfica usa hardware que imita a estrutura e funcionamento do cérebro.

¹⁴⁴ A RainAI não enfrentou problemas com o governo americano pelo aporte de recursos (modestos) da Baidu, mas expressivo investimento da Prosperity7 Ventures (unidade da Aramco Ventures, parte da estatal Saudi Aramco), não foi aprovado pelo *Committee on Foreign Investment in the United States* (CFIUS) e obrigou a venda de sua participação na empresa sob alegação de riscos à segurança nacional pelo potencial acesso da China a semicondutores dos EUA por meio do uso de intermediários no Oriente Médio (Reuters, 2023).

¹⁴⁵ São expressivos os investimentos em IA pelos chamados fundos soberanos (SWFs, sigla de *Sovereign Wealth Funds*), formados por receitas e reservas de petróleo, com destaque para o maior deles, o fundo norueguês NBIM, com posições em grandes players (Apple, Google e Microsoft), e do fundo saudita PIF (*joint-venture* entre a saudita SCAI e a chinesa SenseTime, que é parcialmente de propriedade do Fundo de Investimento na Internet, estatal chinesa) com vistas à construção de um ecossistema local de IA.

3.6.3 Os principais chips usados pelos modelos de IA

O treinamento dos modelos de IA se beneficia da melhoria de dois parâmetros técnicos para chips de IA: maior poder de processamento (mensurada pelo número de operações matemáticas que pode realizar em um segundo) e mais rápida velocidade de transferência de chip para chip (rapidez com que um chip pode enviar informações, em *bytes*, para dentro e para fora do chip em *gigabytes* por segundo).

A Nvidia foi precursora na venda de GPUs no início dos anos 2000 e atualmente é o principal *designer* de chips de IA, com vários modelos, em ordem crescente de poder computacional: V100 (lançado em 2017), A100 (2020) e H100 (2022) (Harris, 2023), além dos chips de arquitetura Blackwell (B200, B100 e GB200, e os correspondentes GB200-NVL72 e GB200-SuperPod, lançados em 2024). As TPUs do Google, fabricados pela TSMC, têm também modelos variados, como o TPUv1 (2015), o TPUv2 (2017), o TPUv3 (2018), o TPUv4 (2021), TPUv5e e TPUv5p (2023) e uma sexta geração de TPU (Trillium) (2024) com maior desempenho e eficiência energética para atender aos crescentes requisitos de maior transferência de dados, comunicação mais rápida entre grandes clusters de chips, reduzindo os custos de treinamento dos modelos¹⁴⁶.

Além da qualidade dos chips Nvidia em si, o pacote que fornece aos clientes inclui o software para aprendizado de máquina, chamado CUDA, que dispensa a programação em nível de GPU pelo usuário e facilita o uso de hardware pelos “devs”. Esta barreira à entrada no mercado de chips de IA começou a ser desafiada por concorrentes desde o Tensorflow (Google) e PyTorch 2.0 (Meta), o Triton, solução de código aberto da OpenAI, o Composer, biblioteca de treinamento de aprendizado profundo de código aberto e o Foundry repositório de código de treinamento, ajuste fino e implantação de LLMs da MosaicML, facilitam o uso de chips concorrentes, como da AMD (Patel e Kostovic, 2023).

Levantamento do *State of AI report compute index v3 (2024)* (junto com a Zeta Alpha), calcula o uso dos principais chips a partir da sua menção em publicações de pesquisa de IA (*proxy* do seu uso pela indústria), mostrando esmagadora predominância dos chips Nvidia, mencionados em 95% dos artigos publicados em 2023. As TPUs do Google foram citadas em apenas 1% dos artigos, apresentando crescimento em 2024 (4% dos artigos) (Gráfico 3.11). As startups de chips, Cerebras, Graphcore, Cambricon, Habana e SambaNova, mais voltadas para a comunidade de “devs.” de código aberto, dependem de serviços de produção contratada da TSMC e Samsung para produzir e embalar seus *designs*, com desafios de custo e acesso

¹⁴⁶ As principais diferenças entre os tipos de TPU são preço, desempenho, capacidade de memória e disponibilidade zonal. Ver <https://cloud.google.com/tpu/docs/regions-zones?hl=pt-br>. Acesso 06-06-2024.

limitado a serviços de fabricação por contrato¹⁴⁷.

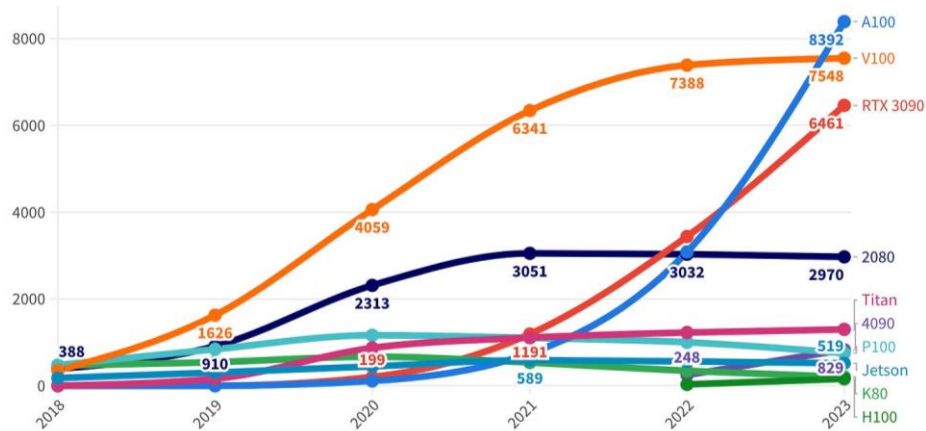
Gráfico 3.11 - Uso de tipos de Chips relatados em artigos publicados, ente 2018 e 2023 (e estimativa 2024)



Fonte: State of AI report compute index v3 (2024) e Zeta Alpha

O Gráfico 3.12 mostra os modelos de chips Nvidia mais populares, também com base na citação em artigos publicados, destacando o A100, que superou o modelo V100 em 2023, seguidos por RTX 3090, mas o recente H100 de melhor desempenho é ainda pouco utilizado.

Gráfico 3.12 - Diferentes tipos de Chips mencionados nos artigos sobre IA publicados



Fonte: State of AI report compute index v3 (2024) e Zeta Alpha

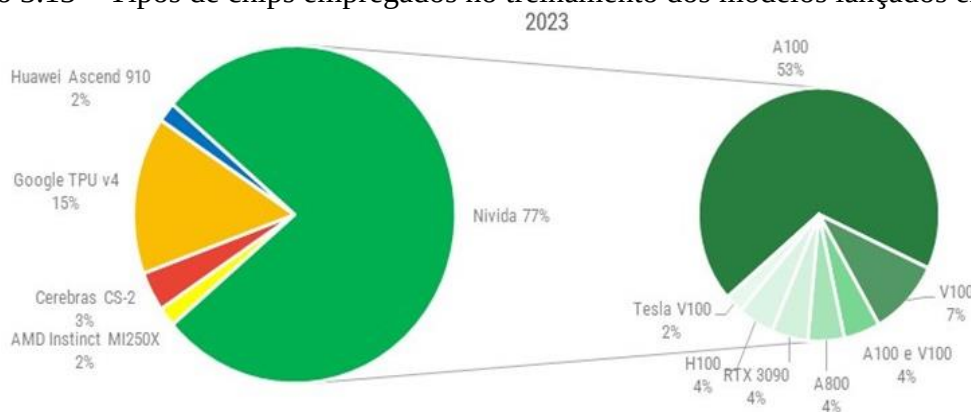
De fato, a escassez de chips Nvidia, cujos preços alcançam até US\$ 40 mil/cada, abriu espaço para rivais no uso em IA. A AMD (cujo novo chip MI300x, tem menor preço, mesmo com custo de fabricação duas vezes superior ao H100) e a Intel (desenvolveu chips para IA na nuvem a partir da aquisição das empresas Nervana Systems, Movidius e Mobileye, mas somente viu aumentadas as chances de *catching up* com Nvidia desde a política americana de

¹⁴⁷ The Information compilou uma lista surpreendente de 18 startups que trabalham com chips de IA, que juntas levantaram mais de US\$ 6 bilhões de investidores e são avaliadas em US\$ 25 bilhões.

incentivos ao desenvolvimento de chips¹⁴⁸). Versões mais lentas de processadores (A800 e H800 Nvidia), embora supostamente maior velocidade de transferência de chip para chip, começaram a ser vendidas para empresas chinesas desde às restrições à exportação de chips de ponta para a China em 2022, com preços que podem alcançar US\$ 70 mil/cada (H800) (Shilov, 2023a).

A base de dados da Epoch permite identificar os tipos de chips empregados no treinamento de uma amostra de modelos intensivos em computação, lançados em 2023 e 2024, uma vez que poucos discriminam o tipo de chips empregado (em somente 39% do total dos modelos lançados em 2023 e em 44% dos modelos de 2024). O Gráfico 3.13 evidencia a inquestionável preponderância dos chips Nvidia no treinamento dos modelos, (77% e 73%, respectivamente, em 2023 e 2024, particularmente do tipo A100).

Gráfico 3.13 – Tipos de chips empregados no treinamento dos modelos lançados em 2023



Fonte: Elaborado a partir da base de dados Epoch.

3.6.4. A infraestrutura física por trás da IA: Data centers, satélites, cabos submarinos

A infraestrutura da IA não se limita somente a chips semicondutores, envolvendo a construção de data centers (centros de processamento de dados), infraestrutura de redes, satélites e cabos submarinos de transmissão, que em seu conjunto contribui para o fornecimento de soluções completas de gerenciamento de dados, análise, escalonamento e modelos nas nuvens corporativas, difíceis de replicar sem investimento considerável. As previsões de gastos no mercado global de infraestrutura de IA são de US\$ 422,55 bilhões até o final desta década, segundo Data Bridge Market Research¹⁴⁹. As estimativas apontam um aumento de quase três vezes na capacidade de data center dos grandes provedores de nuvem nos próximos seis anos (Synergy Research Group)¹⁵⁰, exigindo investimentos para modernização das estruturas

¹⁴⁸ A Intel lançou o acelerador Gaudi2AI incorporado às CPUs, além do Gaudi3, cujo preço será US\$ 15 mil, metade do B100 Blackwell, novo chip Nvidia. Informou na Computex 2024 que o kit com oito unidades do Gaudi 3 custará \$125 mil dólares no total, ou seja US\$15.625/cada (Corumba, 2024).

¹⁴⁹ Disponível em: <https://www.databridgemarketresearch.com/>. Acesso 10 set.2024

¹⁵⁰ Disponível em: <https://www.srgresearch.com/complimentary>. Acesso em: 20 ago. 2024

existentes com incorporação de chips avançados e construção de novos data centers, em especial com maior eficiência energética, e demandando a instalação de 440 mil km de novas linhas submarinas entre 2020 e 2025.

Data centers são grandes instalações, edifícios similares a armazéns, equipados com servidores, roteadores e outros hardwares que armazenam e processam dados e fornecem computação. Os data centers de IA possuem mais servidores rodando em chips de alto desempenho do que data centers convencionais, com elevado consumo de energia (podendo chegar a 50 KW, frente a 7 KW nos convencionais), o que demanda construção de infraestrutura adicional capaz de fornecer energia e, face à geração de calor, o emprego de métodos de resfriamento para evitar superaquecimento do equipamento (em especial, sistemas de resfriamento líquido). Existem empresas especializadas na construção de data centers (Turner Construction, Holder Construction e DPR Construction), além de operadores de data center (DataBank), incluindo startups, como CoreWeave.

Os países líderes em IA esbarram atualmente em limitações de espaço físico para conseguir expandir sua rede de data centers, recorrendo a *designs* modulares. Os obstáculos energéticos têm levado a pautar a locação de data centers pelo consumo de energia, investindo em fontes renováveis.

As BigTechs são as principais proprietárias de data centers e realizam os maiores investimentos. Em 2022, os data centers da Amazon cobriam 84 zonas de disponibilidade em 26 regiões geográficas em todo o mundo, contemplando planos de expansão para mais 24 zonas de disponibilidade em 8 regiões (AWS, 2023). A Apple está investindo na expansão de data centers, dobrados em curto período, a partir de processadores próprios, com vistas ao atendimento de clientes de IA no dispositivo e na nuvem (Patel e Xie, 2024).

São também os maiores investidores em satélites e, principalmente, cabos submarinos que realizam a transferência de dados ao redor do planeta. A rede de cabos de fibra soma quase 1,5 milhão de quilômetros, quase metade adicionada na última década, transportando dados em grande velocidade, de até 250 terabits por segundo nos cabos mais recentes, transportando quase 99% do tráfego de internet intercontinental. Originalmente construídos por consórcios de operadoras de telecomunicações (como BT e Orange, antiga France Telecom), que detinham a maior capacidade, começaram a atrair investimentos privados na década de 1990 pelas perspectivas lucrativas com a venda de capacidade. Desde 2010, o aumento do tráfego de dados levou as gigantes da internet e da computação em nuvem, como Google, Microsoft, Meta e Amazon, que representam a maior parte do tráfego de dados e voz (95%), não só a alugar

capacidade, mas financiar e construir diretamente suas próprias redes. Hoje, 99% dos cabos são de propriedade de empresas privadas (TeleGeography)¹⁵¹.

Começaram então a construir redes globais de data centers, conectados por cabos submarinos, com investimentos em infraestrutura de redes proprietárias ou coproprietárias, visando assegurar rápida e disponibilidade para seus serviços globais. O pioneirismo mais uma vez coube ao Google (Cooper, 2019), desde o consórcio de cabo Unity, em 2010, seguido pela Meta, Microsoft e, mais recentemente, Amazon, que além de grandes compradores pré-venda - como arrendamentos anuais com prazo padrão de 15 anos - despontam como principais investidores. Em 2018, Google, Meta, Amazon e Microsoft possuíam ou alugavam mais da metade da largura de banda submarina (Satariano, 2019). O Google controla atualmente quase 5% dos cabos submarinos que cruzam o planeta. Atualmente, mais de 90% do tráfego da Internet flui por cabos submarinos (TeleGeography) e a capacidade dos provedores de conteúdo passou a 71%, em 2022, frente a menos de 10% antes de 2012.

Os investimentos em cabos pelas BigTechs respondem por quase um quinto dos US\$ 12 bilhões planejados nos próximos quatro anos. O Google é proprietário direto de 12 dos 26 cabos em que participa; a Microsoft e a Amazon, de 4 redes; a Meta é dona de um sistema e investidora em outros 14. Cabos dedicados possibilitam que as BigTechs não precisem competir com outras empresas por largura de banda de terceiros, podendo responder rapidamente a qualquer mudança na demanda e problemas que obriguem ao redirecionamento para outras linhas de cabos danificados de uma rota, além de projetar rotas que atendam a necessidades específicas. Essas empresas podem conectar seus cabos diretamente aos data centers próprios, acelerando o tráfego. Também alugam capacidade para operadoras de telecomunicações, que deixam de arcar com os pesados investimentos.

Cabos submarinos se tornaram ativos econômicos estratégicos valorizados, com custos de mais de US\$ 40.000 por milha (ou US\$ 225 mil/km), exigindo investimentos expressivos nos transoceânicos - entre US\$ 250 e US\$ 300 milhões para o transatlântico e US\$ 300 a US\$ 400 milhões para o transpacífico. Parcela significativa do custo (25%) corresponde à instalação marítima, arcada por consórcio de investidores que compartilham o custo e o risco, além das estações de aterragem e dos cabos em si.

No entanto, o negócio de cabos submarino também está sendo afetado pela disputa tecnológica geopolítica (The Economist, 2023b). Por um lado, esbarra em entraves dos governos à obtenção das licenças necessárias, como ocorreu com a não aprovação americana

¹⁵¹ Disponível em: <https://www.submarinecablemap.com/>. Acesso 10 set. 2024

da etapa final do trecho final do Pacific Light Cable Network (PLCN), anunciado em 2016 com apoio do Google e Meta, que ligaria a costa oeste dos EUA a Hong Kong, paralisado desde 2022. Por outro, a não aceitação de fornecedores chineses (como hmnTech) pelo oligopólio de empresas com capacidade de colocar cabos em profundidade (a francesa Alcatel Submarine Networks, a japonesa NEC e a americana SubCom), levou a China a investir sozinha em rotas de cabos alternativos (como o Peace, que liga Quênia à França, via Paquistão, como parte da Digital Silk Road), levando a preocupações pela US Federal Communications Commission (FCC) com a expansão de investimentos de operadoras chinesas (China Telecom, China Unicom e China Mobile Limited, da ordem de US\$ 500 milhões em uma rede conectando China e França via Cingapura, Paquistão e Egito).

3.6.5. Computação na nuvem: A principal vantagem competitiva das BigTechs

A computação na nuvem engloba um conjunto de produtos/serviços integrados, permitindo o fornecimento de uma ampla gama de serviços de computação, cobrindo toda a cadeia de valor digital, desde infraestrutura física básica (*Infrastructure as a Service - IaaS*), além de serviços diversos, desde *Platform as a Service (PaaS)*¹⁵², *Software as a Service (SaaS)*¹⁵³, inovações como unidades de processamento gráfico, *GPUs as a Service (GaaS)* e até *AI as a Service (AIaaS)* e *Data as a Service (DaaS)*. A prestação dessa ampla gama de recursos como serviços corresponde ao que alguns autores definem como “*Everything as a Service*” (*XaaS*) (Graux, 2021).

Um produto/serviço só se integra a outros produtos/serviços da mesma empresa, criando aprisionamento em diversos níveis e explicando a disputa estratégica entre Amazon, Google e Microsoft "para ser a infraestrutura fundamental" para a IA (Rahman e Thelen, 2019:180). O aprisionamento (*lock-in*) é a característica chave da "infraestrutura de plataformas" (Plantin et al., 2018:295-8), compreendendo o bloqueio em nível horizontal à concepção da plataforma, funcionalidade e *design*, bem como em nível vertical, onde serviços são parte do arranjo de configurações de hardware, poder computacional e capacidade de armazenamento de dados da empresa líder ("plataformização da infraestrutura"). Isso ocorre no Amazon SageMaker, segmento de negócios da divisão de computação em nuvem AWS, na versão de livre acesso do Google TensorFlow, que contém uma ampla gama de componentes e módulos, desde pacotes

¹⁵² Capacidade para implantar na infraestrutura de nuvem aplicativos usando linguagens de programação, bibliotecas, serviços e ferramentas suportadas pelo provedor. O usuário não gerencia nem controla a infraestrutura subjacente, incluindo rede, servidores, sistemas operacionais e armazenamento, somente aplicativos e definições de configuração no ambiente onde são hospedados (Mell e Grance, 2011:2-3).

¹⁵³ Aplicativos fornecidos pelo provedor para o usuário, executados na infraestrutura em nuvem, que são acessíveis aos dispositivos do usuário por meio de uma interface, como um navegador da Web (por exemplo, e-mail baseado na Web) ou uma interface de programa (Mell e Grance, 2011:2).

para simplificar a criação de redes neurais artificiais até a integração desses modelos na infraestrutura de computação em nuvem, da mesma forma que no Microsoft Azure (Ferrari 2023).

Infraestruturas na nuvem facilitam a integração dos modelos de redes neurais profundas a aplicativos de negócios de “devs.”, por meio de interfaces de programação de aplicativos (APIs) e *kits* de desenvolvimento de software (SDKs)¹⁵⁴. São, portanto, pontes que ligam BigTechs e seus clientes, que não se restringem a venda de capacidade computacional, mas habilitam tecnologias complementares, uma vez que o modelo arquitetônico da plataforma na nuvem possibilita a programação por meio de uma interface de software (Helmond, 2015).

As empresas líderes em IA treinam modelos em infraestruturas próprias de hardware, adquirindo e desenvolvendo chips de IA, com as vantagens da integração vertical. Empresas menores e mesmo organizações de código aberto treinam seus modelos remotamente, usando computação na nuvem das gigantes do setor, mediante o pagamento de taxas para acessar o hardware de que precisam, por meio de créditos de computação, compartilhamento de receita ou outros acordos contratuais. Esta parece ser a principal vantagem competitiva atual das BigTechs, o que lhes possibilita alavancar ativos e manter a efetiva liderança, comparado às empresas menores.

As potentes infraestruturas de hardware como as TPUs do Google, os chips proprietários da Amazon e os supercomputadores da Microsoft, são importante fator de competição, tanto para treinamento de modelos próprios de IA, como para acesso a modelos de terceiros por meio do fornecimento de serviços em nuvem, seja para subsidiárias integrais (do Google na DeepMind), minoritárias (da Microsoft na OpenAI) ou mesmo empresas sem quaisquer vínculos acionários.

No caso das subsidiárias, contemplam subsídios cruzados intrafirma do mesmo grupo em termos de recursos financeiros e infraestruturais. Esse é o caso da DeepMind, cujos modelos de aprendizado por reforço demandam potentes e caros recursos computacionais, dependendo integralmente da infraestrutura técnica de serviços na nuvem, TPUs e recursos financeiros do Google (Dickson, 2020). Apesar de tecnologicamente dinâmica, que desde o AlphaZero apostou quase integralmente nas técnicas de aprendizado por reforço e desenvolve modelos efetivamente disruptivos como AlphaDev, AlphaFold e o recente Gemini, a empresa é

¹⁵⁴ API (sigla de *Application Programming Interface*) é uma interface para um serviço (*upload* de dados para a nuvem), enquanto SDK é um conjunto mais abrangente de ferramentas e fragmentos de código para construir software (por exemplo, reconhecimento facial). Facilitam a criação de aplicativos de IA por devs, uma vez que têm preço reduzido, entre US\$ 0,228 a US\$ 4/hora, comparado a custo de desenvolver um modelo de IA (Ferrari, 2023).

estruturalmente deficitária (Smith, 2023). Sua principal fonte de receita é a aplicação de tecnologias de aprendizagem por reforço no resfriamento dos servidores e redução do custo energético do Google que, por seu turno, fornece serviços de computação na nuvem “com desconto”. O prejuízo, em 2019, foi de £477 milhões, mesmo com a renúncia pelo Google dos juros e amortização de £ 1,1 bilhão de “*intercompany loans*” (Ferrari, 2023). Dickson (2020) ressalta que sem o fornecimento da “infraestrutura técnica” subsidiada do Google, seu prejuízo seria maior, decorrente das despesas com computação.

Estimativas de Rubiera (2020) sobre o custo para executar o modelo revolucionário AlphaFold2 usando serviços na nuvem seria de US\$ 25 a 200 mil, que acrescido de outros recursos técnicos poderia chegar a milhões de dólares, de forma que o sucesso do modelo decorre, em grande medida, dos recursos computacionais ilimitados e subsídios do Google. Em 2023, a DeepMind, a verdadeira joia da coroa do Google, foi fundida aos extintos Google Brain e Google Research, tornando-se Google DeepMind (GoogleDeepMind, 2023), dificultando futuras análises financeiras.

Fernandez et al. (2020) e Klinge et al. (2022) lembram que subsídios cruzados são quase uma regra na economia de plataformas, com lucros de segmentos lucrativos (como o caso notório da publicidade *online*) usados para financiar a entrada/expansão em outros segmentos (saúde, por exemplo, uma das apostas do Google).

A parceria da Microsoft e OpenAI envolve uma relação intrincada, incluindo investimento de US\$ 13 bilhões na OpenAI, entre 2018 e 2023, que resultou em participação de 49% no capital da empresa (e participação no Conselho, ainda que sem direito a voto, da qual se afastou recentemente devido à investigação antitruste), em troca do acesso exclusivo aos modelos OpenAI por meio de acordo de licenciamento e recursos computacionais da Microsoft (Microsoft, 2023; Kak, West e Whittaker, 2023). Todos os exitosos modelos OpenAI, desde o GPT-3,5, base do ChatGPT, o modelo de imagem DALL-E e o multimodal GPT-4, entre outros, foram treinados e funcionam nos supercomputadores da Microsoft (Langston, 2020). Isso inclui os processos de inferência diária do ChatGPT, cálculos necessários para gerar saídas de texto, imagem etc. que de outro modo estariam além da capacidade financeira de uma empresa emergente, o que explica a OpenAI se referir à Microsoft como seu “parceiro comercial e de computação”¹⁵⁵. A conexão entre as empresas se expressa em inextricáveis vínculos computacionais e financeiros, direitos da Microsoft sobre 75% dos lucros da OpenAI

¹⁵⁵ Disponível em: <https://openai.com/our-structure/>. Acesso em: 11 ago. 2024

desde o investimento em 2023 (Zitron, 2024a), gastos muito além da capacidade da empresa emergente e perspectivas de uma perda operacional de US\$ 5 bilhões, em 2024 (Evanson, 2024).

Além dos 49% seu capital, o acordo prevê que 75% de seus lucros futuros eventuais serão destinados a amortizar os US\$ 10 bilhões aportados pela Microsoft em 2023 – que parecem ser liberados em parcelas, na forma de créditos de computação na plataforma Azure, dinheiro para uso carimbado apenas em produtos (nuvem) da Microsoft (Albergotti, 2023). A OpenAI não pode deixar de usar o Azure, seja por acordos, seja porque parte do financiamento (ou aporte de recursos) da Microsoft foi feita em créditos de computação. Zitron (2024a) acredita que, como essas contas “não fecham”, a Microsoft seja efetivamente proprietária da OpenAI – ou pelo menos de toda sua propriedade intelectual não-AGI.

Os pagamentos das versões premium disponibilizadas por empresas emergentes não parecem compatíveis com o tamanho da aposta computacional dos seus modelos e com capacidade de atrair clientes que possibilitem receitas e ampliação de suas margens brutas. A OpenAI vem buscando um modelo de negócios que remunere sua inovação, copiando produtos existentes consagrados, incluindo sua não exitosa GPT Store, inspirada nas lojas Apple e Google, e o recente lançamento do Search GPT, inspirado no Search Engine (Google) (OpenAI, 2024c). Na falta de um modelo de negócios exitoso e de condições de apropriabilidade, a própria existência da OpenAI depende visceralmente da Microsoft que é, ao mesmo tempo, provedora e concorrente direta, seguindo a mesma lógica usual das plataformas digitais.

A relação entre as duas empresas, contudo, parece de mão dupla. Os modelos da OpenAI talvez não tivessem alcançado êxito sem os recursos de hardware da Microsoft, embora a OpenAI também pareça ter sido basal para evitar que a Microsoft repetisse o erro estratégico que a levou a perder a corrida dos smartphones (Yoffie, Gawer e Cusumano, 2019; Vergara, 2024).

A Microsoft não destina integralmente sua capacidade de GPU a OpenAI, com a maior parcela do US\$ 50 bilhões planejados pela Microsoft em data centers de IA voltada para cargas de trabalho internas. E procura alternativas para o risco de ruptura do acordo que lhe dá acesso aos modelos OpenAI – e que não abarca eventuais modelos de AGI, principal meta da OpenAI (Patel e Nishball, 2024a). Os autores argumentam que “a Microsoft está tentando mover a maioria de seus volumes de inferência dos modelos da OpenAI para seus próprios produtos onde detém a propriedade intelectual”, como Copilot e Bing, concluindo que: “*Embora a Microsoft não seja competitiva com a OpenAI [...], está rapidamente tentando desenvolver essas habilidades o mais rápido possível. A complexa relação entre as duas empresas parece explicar a busca da OpenAI por parcerias (e *funding*) para desenvolvimento de chips*

proprietários, da mesma forma que as apostas da Microsoft com outros criadores de modelos de IA generativa. A pseudo-aquisição (*acqui and hiring*) da Inflection lhe dá um salto rápido para um modelo decente e uma sólida equipe de pré-treinamento e infraestrutura, mas é necessário mais para fechar o alvo sempre em movimento da OpenAI” (grifos da autora).

A infraestrutura de hardware (além da capacidade financeira) parece também central em parcerias das BigTechs, que atuam como provedoras de computação na nuvem para *players* com os quais não mantêm vínculos acionários. O CEO do Google, Sandar Pichai, afirma que as startups da IA generativa dependem das capacidades do Google Cloud (Moreno, 2023). De fato, as empresas com os modelos de IA mais promissores estão formando parcerias com as provedoras na nuvem que, se por um lado, são recursos essenciais para treinamento e ajuste fino, por outro, têm suas plataformas potencializadas pelos novos recursos da IA generativa, como ocorreu com o Azure desde a parceria com a OpenAI (Roach, 2023). Os custos de construir capacidade computacional do zero são expressivos, além de requisitos de interoperabilidade em pontos-chaves da computação, estão sendo enfrentados por meio da garantia de créditos de computação ou outros acordos contratuais entre empresas menores e as BigTechs infraestruturais (West e Vipra, 2023).

Empresas emergentes com modelos promissores estão cada vez mais vinculadas às líderes do setor. A Mistral AI, recebeu investimento de €\$ 15 milhões da Microsoft - ainda sujeito à aprovação pela US Federal Trade Commission e reguladores europeus. A Inflection AI (empresa avaliada em US\$ 4 bilhões, em 2023), que teve aporte conjunto da Microsoft e Nvidia (US\$ 1,3 bilhão) para construir um cluster de 22.000 GPUs, acabou perdendo setenta especialistas e até o cofundador (ex-DeepMind/Google), contratados pela Microsoft para chefiar sua recém-criada unidade de IA de consumo e um acordo de licenciamento dos modelos por US\$ 650 milhões (Coldewey, 2024; Hu e Varghese, 2024). Somam-se investimentos da Microsoft em infraestrutura na nuvem no Reino Unido (US\$ 3,2 bilhões), mais que dobrando sua presença em data centers no país – que Kak, West e Whittaker (2023) veem como tentativa de atenuar a investigação sobre o mercado de nuvem pela CMA, por comportamento anticoncorrencial (lock-in) -, na Malásia (US\$ 2,2 bilhões) e na Suécia (US\$ 3,2 bilhões) (Reuters, 2024a; Lopes, 2024) e até no Brasil (US\$ 2,7 bilhões) (Reuters, 2024b).

Isso é considerado pela literatura como início do processo chamado “*acquisition and hiring*” de empresas emergentes, que difere de sua tradicional aquisição, contemplando apenas a propriedade intelectual sobre os modelos e a contratação de parte de seus talentos, como vem ocorrendo com Inflection (pela Microsoft), Character AI (cujos cofundadores, Noam Shazeer e Daniel de Freitas, retornam ao Google e firmam acordo de licenciamento não exclusivo de suas tecnologias) (Mehta,

2024), Adept AI (contratação dos ex-fundadores e equipe técnica e licenciamento da tecnologia pela Amazon) (Wiggers, 2024) e potencial de alcançar outras empresas emergentes da IA generativa.

A Anthropic, outra estrela emergente, ainda resiste, mas avançou graças ao aporte de recursos do Google, em parceria desde 2023 (Google, 2023), e Amazon, com parceria de longo prazo que prevê investimentos de até US\$ 4 bilhões em troca de participação minoritária e disponibilização dos modelos gratuitos (Claude 3 Sonnet) e por assinatura paga (Claude 3 Opus) no AWS, bem como modelos de outras empresas, tornando-se provedor da nuvem preferencial para desenvolvimento/treinamento de modelos e compromisso de uso dos chips proprietários (Trainium) da Amazon, otimizados para construir IA em larga escala (Amazon, 2023). Outras negociações recentes se mostraram infrutíferas, envolvendo a Ideogram e a Reka (ambas ex-Google).

CAPÍTULO 4 – INTERVENÇÃO GOVERNAMENTAL E EFEITOS POTENCIAIS SOBRE A ARQUITETURA DA INDÚSTRIA, O PADRÃO DE COMPETIÇÃO E O REGIME DE APROPRIABILIDADE DA IA

A literatura mostrou que a intervenção governamental, em particular por meio de políticas antitruste, teve papel fundamental na arquitetura da indústria de TICs nos EUA, com sua modularização, desverticalização das empresas e um fluxo permanente de novas empresas especializadas em módulos da indústria, como hardware, software e a própria IA.

O momento atual em meio à grande propagação de modelos da IA generativa é de renascimento de amplas dimensões regulatórias, acompanhadas por inquietações quase generalizadas sobre os possíveis riscos e danos da IA para a sociedade e da concentração de mercados e assimetrias, a despeito dos amplos benefícios potenciais desta poderosa tecnologia. Em termos econômicos, correspondem ao enfrentamento das típicas externalidades, positivas e negativas, das inovações tecnológicas, que levam a várias formas de intervenção do Estado na economia com medidas de natureza regulatória, políticas da concorrência/antitruste e legislação de propriedade intelectual.

Os modelos de redes neurais profundas geram grandes inquietações em virtude da sua complexidade, hermetismo e baixa “explicabilidade”, além de problemas relacionados aos dados e ao modelo de negócios de plataformas digitais. Às profundas inquietações relacionadas aos danos potenciais à privacidade, proteção e vieses nos dados, preconceitos de diversas naturezas, concentração de mercado e práticas anticompetitivas das empresas orquestradoras de plataformas digitais, somaram-se riscos no uso de sistemas de IA, que evoluíram para mudanças normativas e legais variadas.

Os riscos dos modelos da IA generativa geraram agitação adicional pela capacidade de criar conteúdo “artificial”, difíceis de distinguir do original, potencializando a desinformação, a criação de *deep fakes* e os chamados riscos à humanidade pelo mau uso dessas poderosas tecnologias. O chamado risco existencial (risco X) (Bengio et al., 2023), que não seria causado por uma AGI ou computadores dotados de inteligência sobre-humana, que possam destruir a vida no planeta, mas riscos decorrentes do mau uso da tecnologia, envolvendo segurança cibernética, desenvolvimento de armas químicas e biológicas, armas autônomas letais (enxames de drones) e sistemas que adquiram tendência à autopreservação, tentando se replicar ou evitar serem desligados, por meio de agentes generativos autônomos com *plug-ins* que os ligam a aplicações de terceiros, permitindo que os modelos operem com dados em tempo real e tornando-se ainda mais autônomos (Anderljung et al., 2023). Passou a haver uma convergência global sobre a necessidade de regulação da IA, pelos riscos e baixa confiabilidade dos sistemas,

o que, por outro lado, levanta obstáculos adicionais para sua implementação em setores da economia e uma crise de credibilidade generalizada como alerta a futurista Amy Webb (Gil, 2023).

Inúmeras formas de intervenção pública sobre o ecossistema da IA envolvem regulação, políticas antitruste, tratados internacionais, políticas industriais e comerciais relacionadas a cadeia de valor da IA. A política antitruste e da concorrência, passou a estar voltada não apenas para a concentração de mercado da IA, mas também a concentração da infraestrutura computacional. As ações regulatórias estenderam sua abordagem dos usos da IA e dados, alcançando os ativos tangíveis para a produção da IA, ao estabelecer requisitos para os modelos com base na quantidade da computação empregada. Esses temas são abordados neste capítulo sem a profundidade e pretensão de esgotar assuntos tão amplos e complexos, somente visando identificar seus impactos potenciais sobre a inovação, competição e apropriabilidade, além da (re) definição de vencedores e perdedores.

4.1. Avanços recentes na regulação com abordagens sobre modelos de fronteira

Inúmeras iniciativas normativas estão surgindo ao redor do mundo desde o surgimento da IA generativa, por diferentes nações, blocos regionais e abordagens de natureza global nos fóruns internacionais. As principais estão resumidas no Quadro 4.1, destacando sua reorientação recente dirigida aos modelos de fronteira, de fundação intensivos em computação.

4.1.1. A abordagem europeia

O primeiro projeto de regulação abrangente da IA, ainda na fase dos modelos preditivos, proposto pela União Europeia em abril de 2021, mas cuja versão definitiva do Artificial Intelligence Act, foi aprovada somente em dezembro de 2023, passando por mudanças importantes, incluindo a própria definição de sistemas de IA, a inclusão de uma nova categoria de sistemas e foco particular em grandes modelos da IA generativa¹⁵⁶. A lei visa assegurar sistemas de IA seguros, transparentes, rastreáveis e não discriminatórios, como parte da ampla estratégia digital em construção desde 2018, com a lei de proteção de dados pessoais e privacidade (GDPR – *General Data Protection Regulation*)¹⁵⁷, hoje *benchmark* global¹⁵⁸, relacionadas à proteção, governança e compartilhamento (portabilidade) de dados, analisado no

¹⁵⁶ Disponível em: <https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf>. Acesso em: 03 jun. 2024.

¹⁵⁷ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN>. Acesso em 10 set.2024

¹⁵⁸ A importância do mercado europeu abriu espaço para perspectivas de impactação global do seu amplo aparato regulatório, por meio do *Brussels' effect*, expressão cunhada por Bradford (2012, 2020).

próximo capítulo, com destaque para o ambicioso EU Data Act 2022/0047¹⁵⁹ que visa criação de um mercado único europeu de dados.

Quadro 4.1 - Principais Desenvolvimentos Regulatórios da IA²

Iniciativas regulatórias	Observações
Principais iniciativas nacionais	
China <u>Regulamento de mecanismos de recomendação algorítmica</u> (27/08/2021), <u>Disposições sobre a Administração de Síntese Profunda de Informações Baseadas na Internet Serviços</u> (11/12/2022) e <u>Medidas Administrativas para Serviços de Inteligência Artificial Generativa</u> (11/04/2023)	Rápida divulgação de medidas de regulação dos mecanismos de recomendação algorítmica da IA preditiva, da IA generativa e das “deepfakes” produzidas com tecnologias de síntese profunda, pela <i>Cyberspace Administration of China</i>
Reino Unido <u>Pro-innovation approach to AI regulation</u> , (Livro Branco) (março de 2023) e criação do <u>AI Safety Institute</u> (novembro de 2023)	O Governo do Reino Unido adotou uma abordagem <i>soft-law</i> de regulação e a criação de uma agência reguladora cuja proposta foi submetida ao Parlamento em novembro de 2023
Estados Unidos <u>US Executive Order 14.110</u> (EUA (30/10/2023))	A Ordem Executiva orienta as agências federais nos EUA a tomarem medidas para impulsionar o desenvolvimento seguro, protegido e confiável da IA, abrangendo áreas como proteção e segurança, privacidade, direitos civis e do consumidor, ao mesmo tempo em que promovem a inovação e a competição. Contempla medidas de controle dos sistemas de IA de uso dual e desenvolvidos com modelos intensivos em computação, ou seja, quantidade de computação superior 10 ²⁶ FLOPS, e ainda 10 ²³ FLOPS no caso dos modelos que usem dados de sequência biológica. É implícita a aspiração americana de liderar a inovação e a regulação global da IA
Iniciativas Regionais e Globais	
<u>Declaração do Processo de Hiroshima G7</u> (30/10/2023)	Conjunto voluntário de princípios orientadores internacionais para promover a segurança e a confiabilidade em sistemas avançados de IA. Os princípios se aplicam a todos os atores de IA no ciclo de vida da IA, incluindo design, desenvolvimento, implantação e uso de sistemas avançados de IA
<u>UN AI Advisory Board</u> (27/10/2023)	Proposta para fortalecer a governança internacional da IA, realizando sete funções críticas, como varredura de horizontes para riscos e apoio à colaboração internacional em dados, e capacidade computacional e talento para alcançar os Objetivos de Desenvolvimento Sustentável (ODS), incluindo recomendações para melhorar a responsabilização e garantir voz equitativa para todos os países
<u>Declaração de Bletchley</u> (01/11/2023)	Documento de política assinado por representantes de mais de 25 territórios (incluindo os EUA, a UE e a China) com estabelecimento de objetivos compartilhados de que a IA deve ser “concebida, desenvolvida, implementada e utilizada de uma forma segura, centrada no ser humano, fiável e responsável
<u>Declaração de Seul</u> (22/05/2024)	A Declaração Ministerial assinada por 27 países e União Europeia confirma a compreensão compartilhada das oportunidades e riscos da IA, dando seguimento às discussões de <i>Bletchley</i> .
<u>EU AI Act</u> (21/05/2024) ¹	Amplas regras harmonizadas para garantir que os sistemas de IA na UE “sejam seguros e respeitem os direitos fundamentais e os valores da UE”, seguindo uma abordagem baseada em categorias de risco no uso da IA nos países do bloco europeu, bem como sistemas GPAI quando a quantidade cumulativa de computação usada no seu treinamento é superior a 10 ²⁵ FLOPS

Fonte: Adaptado de Barcellos e Bastos (2024) e CMA (2024:35)

Notas: ¹ Data da aprovação pela Comissão Europeia

² O Quadro não é exaustivo, não contemplando iniciativas regulatórias como Canadá, Austrália, Brasil, entre outros.

¹⁵⁹ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0068&from=EN>. Acesso em 11 set. 2024

diretivas antitruste (*Digital Services Act*¹⁶⁰ e *Digital Market Act*¹⁶¹) para lidar com o padrão diferenciado de competição das plataformas digitais, além de um amplo conjunto de medidas

A abordagem regulatória da IA se baseou originalmente em uma hierarquia de quatro categorias de riscos potenciais nos usos da IA - riscos mínimos, riscos limitados, alto risco e risco inaceitável -, com maiores obrigações à medida em que os riscos se ampliam, como os sistemas de “alto risco”, com requisitos severos de transparência, segurança, supervisão, mitigação de riscos e, principalmente, nos usos de risco inaceitável, com compromissos rigorosos de testes e prestação de contas, com alguns usos proibidos, como pontuação social, manipulação e algumas instâncias de vigilância biométrica.

No proposta original, de 2021¹⁶² a definição de sistema de IA era de um:

“software desenvolvido com uma ou mais das técnicas e abordagens enumeradas no anexo I e que pode, para um determinado conjunto de objetivos definidos pelo homem, gerar resultados como conteúdos, previsões, recomendações ou decisões que influenciam os ambientes com os quais interagem”.

Na versão aprovada, a definição passou a ser:

“sistema baseado em máquina concebido para funcionar *com vários níveis de autonomia e que pode apresentar capacidade de adaptabilidade* após a implantação e que, para objetivos explícitos ou implícitos, infere, a partir dos dados que recebe, como gerar resultados como previsões, conteúdo, recomendações ou decisões que podem influenciar ambientes físicos ou virtuais” (AI Act, Artigo 3º (1)) (grifos da autora).

A IA generativa levou à inclusão de uma nova categoria, não contemplada na abordagem original de usos, baseada em modelos que, pelo uso geral e risco sistêmico, estarão sob maior escrutínio quando superarem determinado uso de capacidade computacional no treinamento (Volpicelli, 2023). Os modelos generativos, que não se encaixam na abordagem de usos específicos, foram incorporados à lei por meio do conceito de IA de Propósito Geral (GPAI, sigla de *General Purpose AI*), genérico, que:

“significa um modelo de IA, inclusive quando esse modelo de IA é *treinado com uma grande quantidade de dados usando autosupervisão em escala*, que apresenta *generalidade significativa* e é *capaz de executar com competência uma ampla gama de tarefas distintas*, independentemente da forma como o modelo é colocado no mercado e que podem ser integrado numa variedade de sistemas ou aplicações a jusante, exceto os modelos de IA que são utilizados para atividades de investigação, desenvolvimento ou prototipagem antes de serem colocados no mercado “ (AI Act, Artigo 3º (63)) (grifos da autora).

Com isso, além da regulação baseada no uso dos sistemas de IA, passou a contemplar também a regulação das tecnologias da IA, ou seja, a produção da IA, para os modelos de maior

¹⁶⁰ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en> Acesso em: 09 jun. 2024

¹⁶¹ Disponível em: eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0842&from=en. Acesso em: 09 jun. 2024

¹⁶² Disponível em: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF. Acesso em: 10 mai. 2022

escala e computacionalmente intensivos incluídos como GPAI com “risco sistêmico”, definido como:

“um risco específico das capacidades de elevado impacto¹⁶³ dos modelos de IA de uso geral, com um impacto significativo no mercado da União [Europeia] devido ao seu alcance ou devido a efeitos negativos reais ou razoavelmente previsíveis na saúde pública, segurança, segurança pública, direitos fundamentais ou na sociedade como um todo, que podem ser propagados em escala ao longo da cadeia de valor” (AI Act, Artigo 3º (65)).

O maior escrutínio regulatório recairá sobre os modelos GPAI com risco sistêmico, que são aqueles que usam maior capacidade computacional, em “operação de ponto flutuante” (FLOP). A mensuração do desempenho computacional em FLOPS (ou operação de ponto flutuante por segundo) indica quantas operações matemáticas de ponto flutuante o modelo é capaz de resolver em um segundo¹⁶⁴.

Em síntese, a lei considera que um modelo GPAI tem capacidade de risco sistêmico e elevado impacto quando a quantidade cumulativa de computação utilizada para o seu treinamento for superior a 10^{25} FLOPS (Artigo 51 (2)). Este limite será alterado pela Comissão Europeia por meio de atos delegados em virtude da evolução tecnológica, melhorias algorítmicas ou maior eficiência do hardware, que afetem o estado de a arte. Para esses modelos os provedores deverão elaborar e manter documentação técnica (treinamento, testes e resultados de avaliação), fornecida também para fornecedores, sem prejuízo de proteger propriedade intelectual e informações comerciais confidenciais e *trade secrets*, de forma a permitir que tenham compreensão das capacidades e limitações do modelo e possam cumprir obrigações da regulação. Essas obrigações não se aplicam a prestadores de modelos GPAI disponibilizados ao público sob licença de código aberto e gratuita, que permite o acesso, uso, modificação e distribuição do modelo, e cujos parâmetros, incluindo os pesos, as informações sobre a arquitetura do modelo e as informações sobre seu uso, são tornados públicos, a menos que o modelo apresente riscos sistêmicos.

4.1.2. A regulação pragmática chinesa

A China foi a efetiva pioneira na aprovação da regulação da IA preditiva, por meio do Regulamento de mecanismos de recomendação algorítmica¹⁶⁵, em agosto de 2021, e na regulação da IA generativa, por meio das Medidas Administrativas para Serviços de

¹⁶³ Capacidades de alto impacto, são “capacidades que correspondem ou excedem as capacidades registradas nos modelos de IA de uso geral mais avançados” (AI Act, Artigo 3 (64)).

¹⁶⁴ O desempenho em FLOPS pode ser grafado com os multiplicadores que estamos acostumados. Por exemplo, 1 kiloflops corresponde a 10^3 , 1 megaflop corresponde a 1 milhão de flops (10^6 , ou seja, 10 com seis zeros), 1 gigaflops corresponde a 1 bilhão de flops (10^9), 1 teraflop corresponde a 1 trilhão de flops (10^{12}), 1 petaflop corresponde a 1 quatrilhão de flops (10^{15}), 1 exaflop (10^{18}), etc. A representação de ponto flutuante é semelhante à notação científica, exceto que tudo é realizado na base dois, em vez da base dez.

¹⁶⁵ <https://www.cnbc.com/2021/09/03/chinas-tech-regulation-turns-to-algorithms.html>. Acesso em 10 out. 2023

Inteligência Artificial Generativa, em abril de 2023 (Luo, Dan e Shepherd, 2023), bem como medidas relacionadas a “*deepfakes*” com as Disposições sobre a Administração de Síntese Profunda de Informações Baseadas na Internet Serviços, em dezembro de 2022 (Interesse, 2022), que se somaram às tradicionais medidas de segurança cibernética. Tais iniciativas parecem ser o embrião de uma lei de IA chinesa, ainda em fase de elaboração (Webster et al., 2023). A China também lançou, em 18 de outubro de 2023, a “Iniciativa Global de Governança de IA”, com três princípios a serem cumpridos: bem-estar de toda a humanidade; segurança, sob controle humano, com melhoria constante na interpretabilidade e previsibilidade; e garanta a justiça e igualdade de condições para os países no desenvolvimento e benefícios da IA.

O país estaria desenvolvendo há dois anos uma arquitetura básica de dados do país, contemplando sistema de direitos de propriedade sobre dados, sistema de circulação e negociação, sistema de distribuição de receitas e sistema de governança de segurança, conforme as Diretrizes aprovadas em dezembro de 2022. O reconhecimento legal dos conjuntos de dados como ativo intangível, envolve seu registro em demonstrações contábeis e seu retorno compartilhado com a sociedade. A National Data Administration (NDA), agência criada em 2023 para supervisão da economia digital e que terá a responsabilidade de coordenar a construção da estrutura de dados do país, referiu-se aos dados (públicos, privados e empresariais) como um tipo diferente de fator de produção e *driver* do desenvolvimento da nova era digital (Xu, 2023). Essas ações definem a estratégia de dados e sua governança na China (Ma, 2024).

4.1.3. A regulação americana marcada pela disputa tecnológica geopolítica

Nos Estados Unidos, a regulação fragmentada e setorial, originalmente baseada no mercado e autorregulação, complementada por legislação dos estados e agências reguladoras federais em setores sensíveis, assumiu uma dimensão federal centralizada com a edição da Ordem Executiva 14.110, de 30 de outubro de 2023¹⁶⁶ (Barcellos e Bastos, 2024). Documento, similar ao nosso decreto presidencial, transformou as salvaguardas não vinculantes do *Blueprint for an AI Bill of Rights*, de 2022¹⁶⁷, em política efetiva de governo, com incomum mobilização da máquina federal para o desenvolvimento e uso de uma IA responsável e segura, que resguarde privacidade, liberdades civis e trabalhadores americanos. O documento menciona entre seus princípios o objetivo de “liderar os rumos do progresso social, econômico e tecnológico global como fez em épocas anteriores de mudança disruptivas”, expressando o

¹⁶⁶ Disponível em: <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>. Acesso em: 02 nov. 2023

¹⁶⁷ Compromissos voluntários firmados em códigos de conduta por 15 desenvolvedores de modelos. Disponível em: <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>. Acesso em: 23 nov.2022

objetivo de liderar a inovação tecnológica e a regulação global da IA, com a promoção do uso seguro das tecnologias da IA, ainda que viesada pela segurança cibernética, sem destruir seu profícuo ecossistema de inovação.

Prevê “*AI red-teaming*” - esforço de teste estruturado para encontrar falhas e vulnerabilidades em um sistema de IA, muitas vezes em um ambiente controlado e em colaboração com devs de IA -, cujos resultados deverão ser informados ao governo, especialmente na IA generativa com “uso dual” (civil e militar) a partir de orientações técnicas, padrões e testes de segurança dos modelos a serem desenvolvidos pelo NIST (*National Institute of Standards and Technology*) e, no caso daqueles a serem usados no setor público federal, testes coordenados pela *Cybersecurity and Infrastructure Security Agency*.

O maior destaque são os requisitos rigorosos para os modelos de fundação com uso dual, exigindo que as empresas informem ao Departamento de Comércio americano o treinamento de algoritmos de IA de alta potência, a quantidade de computação usada, informações sobre propriedade dos dados e detalhes dos testes de segurança, ou seja, informações empresariais normalmente resguardadas por sigilo.

A preocupação central são os modelos de fundação de uso dual treinados usando pelo menos 10^{26} operações de ponto flutuante (superior ao limite do IA Act) ou pelo menos 10^{23} operações de ponto flutuante de modelos que usem dados de sequência biológica, exigindo registro, testes amplos, apresentação de informações e relatórios de monitoramento e supervisão organizacional contínuos, incluindo cibersegurança.

Muitos dos novos modelos, sobretudo proprietários, deixaram de divulgar detalhes técnicos relevantes, incluindo a capacidade computacional empregada. OpenAI e Google não divulgam o poder de computação usado para treinar seus modelos mais poderosos, respectivamente, GPT-4 e Gemini. Relatório do serviço de pesquisa do Congresso sobre a Ordem Executiva (Harris e Jaikaran, 2024) sugere que 10^{26} FLOPS é um pouco mais do que foi usado para treinar GPT-4, de forma que atingirá muito poucas empresas.

Bommasani et al. (2023b) questionam o uso de capacidade computacional como métrica para regulação dos modelos de fundação, uma vez que modelos menos intensivos computacionalmente podem gerar danos maiores, citando o AlphaFold, da DeepMind. Ainda que o uso da computação seja a forma mais viável para rastrear os modelos, há questões sobre o potencial de abuso ao dar amparo legal ao que na prática já pode ser feito pelos provedores na nuvem (potencialmente, inspecionar dados e modelos de clientes), que de outro modo seria uma prática anticompetitiva, desleal e ilegal em termos de proteção de dados, privacidade e confidencialidade.

A *accountability* também recai sobre serviços de computação na nuvem prestados para qualquer empresa, em qualquer lugar do mundo, acima de determinada capacidade computacional, usados para treinar modelos de IA, pelo potencial de risco por usos (inclusive cibernéticos) maliciosos. Isso evidencia que os tentáculos da regulação americana vão além do território americano e, de fato, se confundem com instrumentos da disputa geopolítica da IA. Os maiores provedores de computação na nuvem são Amazon, Microsoft e Google (65% do mercado global), dos quais dependem muitas organizações ao redor do mundo para desenvolver modelos de IA. As gigantes chinesas dispõem de suas próprias infraestruturas na nuvem, mas diante de restrições à exportação de chips, software e equipamentos, poderiam ter que recorrer à computação na nuvem das BigTechs americanas¹⁶⁸.

4.1.4. As novas iniciativas regulatórias em nível global

As preocupações com os modelos generativos e seus riscos diferenciados levou a regulação da IA em termos globais aos fóruns internacionais. A Declaração do *Hiroshima AI Process* pelos países do G7, em outubro de 2023¹⁶⁹, definiu princípios orientadores internacionais para modelos mais avançados de IA (modelos de fundação). O tema dos riscos da IA foi incluído na agenda do Conselho de Segurança das Nações Unidas, em julho passado¹⁷⁰, criando um órgão consultivo de 39 especialistas para a governança internacional da IA ética, segura e responsável para a humanidade, que poderá se tornar uma organização de identificação e monitoramento dos danos e riscos da IA. A Declaração de *Bletchley*¹⁷¹ foi assinada pelos países participantes da *AI Safety Summit 2023*¹⁷², promovida pelo governo britânico nos dois primeiros dias de novembro, complementada pela Declaração de Seul¹⁷³, de maio de 2024. Por fim, a Declaração *Golden Gate*, dos líderes da APEC¹⁷⁴ (*Asia-Pacific Economic Cooperation*), em 17 de novembro de 2023, que entre outros temas do comércio internacional contempla um

¹⁶⁸ Tanto o Google quanto a Microsoft oferecem serviços de nuvem com tecnologia Nvidia para empresas chinesas, por meio de data centers fora da China, enfraquecendo os controles de exportação (Osawa e Holmes, 2024).

¹⁶⁹ Disponível em: <https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process>. Acesso em: 03 nov. 2023

¹⁷⁰ Disponível em: <https://www.un.org/en/ai-advisory-body>. Acesso em: 23 dez. 2023

¹⁷¹ Disponível em: <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>. Acesso em: 03 nov. 2023

¹⁷² Disponível em: <https://www.gov.uk/government/publications/ai-safety-summit-programme/ai-safety-summit-day-1-and-2-programme>. Acesso em: 03 nov. 2023

¹⁷³ Disponível em: <https://www.industry.gov.au/publications/seoul-declaration-countries-attending-ai-seoul-summit-21-22-may-2024>. Acesso em: 12 jun. 2024

¹⁷⁴ Disponível em: <https://www.apec.org/meeting-papers/leaders-declarations/2023/2023-leaders-declaration>. Acesso em 18 nov. 2023. Os países-Membros da APEC são Austrália, Brunei, Canadá, Indonésia, Japão, Malásia, Nova Zelândia, Filipinas, Cingapura, Coreia do Sul, Tailândia e Estados Unidos (desde 1989), China, Hong Kong, Taiwan (desde 1991), México, Papua Nova Guiné (desde 1993), Chile (desde 1994), Peru, Rússia e Vietnã (desde 1998).

posicionamento regional sobre a IA, entre outros arranjos jurídicos inéditos do direito internacional.

4.1.5. O posicionamento divergente das empresas sobre a regulação da IA

As conjecturas sobre “risco x” foram feitas por renomados especialistas em IA, desde Geoffrey Hinton, quando deixou o Google Brain, ou Youshua Bengio, dois precursores dos modelos de redes neurais profundas (The Economist 2023c). Acarretaram duas cartas abertas, uma da Future of Life,¹⁷⁵ de março de 2023, que pede um “waiver” nas pesquisas em IA até sua regulação e conta com quase 34 mil signatários, outra do Center for AI Safety¹⁷⁶, dois meses depois, que faz referência ao riscoX e tem como signatários os principais cientistas e executivos das empresas líderes – Hinton, Bengio, Bill Gates (Microsoft) e Sandar Pichai (Alphabet/Google) – e de empresas emergentes, como Sam Altman (OpenAI), Demis Hassabis (DeepMind), Dario Amodei (Anthropic), lembrando a Carta aberta contra armas autônomas, também da Future of Life, de 2016¹⁷⁷.

No extremo oposto, a carta aberta da Mozilla¹⁷⁸, sob o título *Joint Statement on AI Safety and Openness*, defende que a prioridade global deva ser a abertura, a transparência e o acesso amplo dos sistemas de IA como forma de mitigar seus danos atuais e futuros em escala, mesmo reconhecendo o atual momento crítico para a governança da IA e a necessidade de leis para garantir aplicações seguras, justas e transparentes. A carta, divulgada um dia depois da Ordem Executiva e um dia antes do AI Safety Summit 2023, conta com 1821 signatários, entre os quais Yann LeCun, cientista chefe da Meta AI que dividiu com Hinton e Bengio o Turing Award 2018, Andrew Ng, da Coursera, fundador do Google Brain e ex-CEO da Baidu. Ambos têm feito declarações polêmicas, na mídia e nas redes sociais, criticando abertamente os exageros sobre riscos da IA e atribuem à “captura regulatória” por grupos de interesses.

As correntes divergentes sobre a regulação da IA estão hoje em planos opostos também no que tange à abertura dos modelos da IA. Microsoft, Google e OpenAI, defendem maior regulamentação para a IA, apoiando-se na ideia de riscos existenciais, embora sejam as mesmas que desenvolvem poderosos e herméticos modelos e que progressivamente divulgam menos informações técnicas. A Microsoft passou a incluir a exigência de registro de terceiros no seu regime de licenciamento em modelos acima de determinado limite de desempenho. A Meta, por outro lado, e a comunidade de devs de código aberto, contestam argumentos do risco X e

¹⁷⁵ Disponível em: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>. Acesso em: 25 mai. 2023

¹⁷⁶ Disponível em: <https://www.safe.ai/statement-on-ai-risk#open-letter>. Acesso em: 30 jun.2023

¹⁷⁷ Disponível em: <https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/>. Acesso em: 24 jul.2023

¹⁷⁸ Disponível em: <https://open.mozilla.org/letter/>. Acesso em: 03 nov. 2023

atribuem o cenário atual à intensa concorrência na área. LeCun, que se transformou em uma espécie de porta-voz do movimento, classifica a noção de risco X como culto apocalíptico. Andrew Ng postou na rede social X (@AndrewYNg) que a ideia de que a inteligência artificial levará à extinção da humanidade é uma mentira promulgada pelas grandes tecnologias para desencadear uma regulamentação pesada que encerraria a concorrência no mercado de IA. Para os críticos da regulação, a abertura seria o antídoto, não o veneno.

De fato, o discurso atual em torno da IA de código aberto foi incorporado e instrumentalizado por grandes corporações e pode visar a compreensão do público e de *policy makers* sobre a IA com implicações sobre o debate sobre sua regulamentação. A Meta vem sendo duplamente criticada, pelas rivais por disponibilizar seus sistemas em *open source* cujo descontrole da disseminação amplia os riscos da IA e, por parte da comunidade em virtude da baixa abertura (*open for business*) dos modelos.

Em resumo, a proliferação de modelos da IA generativa acelerou, em 2023 e 2024, o avanço das medidas regulatórias e uma convergência de regulação do uso de capacidade computacional, exigindo informações sigilosas das empresas, parece migrar da preocupação com o uso para o desenvolvimento dos modelos de IA. O distinto posicionamento das empresas sobre a regulação reflete diferentes estratégias empresariais e mostra que a regulação se transformou em um instrumento de competição. De fato, nenhuma regra é neutra, até porque as empresas por elas afetadas não são iguais.

4.2. O escrutínio antitruste se transfere para a computação na nuvem

O poder de mercado, os efeitos de rede, os dados pessoais em troca de serviços gratuitos financiados pela monetização de vigilância extensiva, o aprisionamento de clientes e as práticas anticoncorrenciais pelos orquestradores de plataformas digitais foram temas da regulação antitruste das plataformas digitais amplamente debatidos nos últimos anos da IA preditiva.

O controle sobre dados, o poder de mercado das plataformas online e práticas anticompetitivas das BigTechs em vários segmentos de mercado, bloqueando concorrentes, em virtude da sua usual dupla função como provedor de serviços e concorrente direto, geraram investigações da regulação. Este é o caso de “devs.” de aplicativos que para ter acesso à Apple Store são obrigados a aceitar cláusulas contratuais que dão liberdade a Apple para “...usar qualquer informações, sugestão, recomendação fornecida pelo desenvolvedor”, evidenciando seu duplo papel, ora provedor de serviço de venda de aplicativo, ora concorrente no negócio de aplicativos (House Digital Markets Report, 2020:365). A Amazon hospeda vendedores terceirizados, mediante pagamento de taxas, que são também concorrentes de produtos da

marca própria AmazonBasics, com persuasão e auto preferência, além de aprisionamento. O pagamento de taxas para o Google Search ser o aplicativo padrão na maioria dos dispositivos e navegadores do mundo, sofre acusação de assegurar um quase-monopólio ao produto atrelado à propaganda *online*, sua principal fonte de receita, explorando assimetrias de informação e rastreando dados em tempo real entre mercados através de programas e monitorando concorrentes reais e potenciais (projeto Android Lockbox) (House Digital Markets Report, 2020:14-15). A Meta também usaria a “vantagem de dados” para identificar ameaças competitivas nascentes, “... a partir das quais adquiriu, copiou ou eliminou as empresas rivais” (House Digital Markets Report, 2020:14).

A despeito das investigações do Congresso americano sobre as plataformas digitais (House Digital Markets Report, 2020), a importância das BigTechs para a economia americana e para a disputa geopolítica pela liderança em IA¹⁷⁹, amparada em *lobby* ativo, têm feito com que as investigações de práticas anticompetitivas e abuso de poder de mercado tenham impactos reduzidos sobre as empresas americanas, tal como a política antitruste chinesa em relação às empresas domésticas, comparativamente aos reguladores europeus.

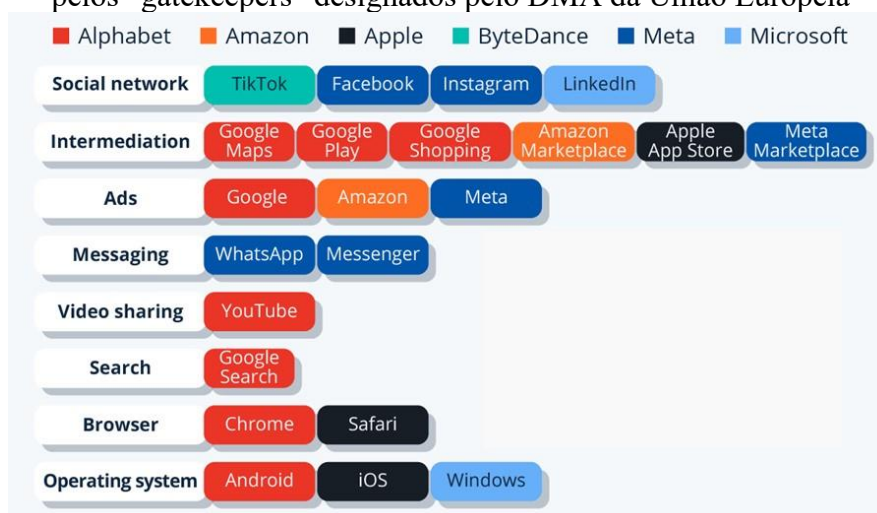
Desde 2016, as plataformas digitais estão sob forte escrutínio na Europa (Coyle, 2020; Vezzoso, 2022), com investigações diversas que resultaram na imposição de multas e recomendações diversas para os orquestradores – incluindo o Reino Unido pré-Brexit (CMA, 2020) – que culminou em uma nova estrutura legal para regulação da concorrência nos mercados e serviços digitais por meio dos mencionados *Digital Services Act* (DSA) e *Digital Market Act* (DMA). O impacto desses complexos corpos jurídicos atingem qualquer empresa que atue no bloco europeu, independente do seu país sede e terá efeitos desproporcionais sobre as plataformas maiores de origem americana (Gráfico 4.1).

Nos Estados Unidos, o avanço da abordagem neo-brandesiana em contraposição ao enfoque dominante da escola de Chicago (Khan, 2018; Wu, 2018) culminou em uma renovada regulação do poder de mercado das plataformas digitais por meio da ação *ad hoc* das agências federais, principalmente na administração Biden - a US Federal Trade Commission (FTC), sob comando de Lina Khan, a Divisão Antitruste do US Department of Justice, comandada pelo

¹⁷⁹ A competição com a China tem sido preocupação frequente e fonte de resistência da área de Defesa dos EUA a mudanças na política antitruste, como expressa a Carta assinada por 12 ex-dirigentes de agências de defesa dos EUA enviada ao Congresso em 15 de setembro de 2021 alertando sobre possíveis consequências dos projetos de lei antitruste sobre plataformas digitais e BigTechs, pela importância dessas empresas para a área de defesa. A nomeação de Lina Khan para a agência antitruste americana, Federal Trade Commission (FTC), e de Jonathan Kanter para a Divisão Antitruste do US Department of Justice, não ocorreu sem reação das empresas. Facebook e Amazon apresentaram petições pressionando o Congresso, cuja resposta foi objeto da Carta, de 12 de outubro de 2021. A atual titular da FTC produziu artigos acadêmicos antológicos, base da nova abordagem antitruste (Khan, 2017).

Procurador-geral assistente Jonathan Kanter, e Tim Wu, ex-assessor de Biden e um dos responsáveis pelo desenho da nova abordagem. Nenhuma nova legislação antitruste foi editada, com o tema se arrastando no Congresso americano, como os projetos da Senadora Amy Klobuchar (Klobuchar, 2024). A emergência dos modelos e das novas ferramentas da IA generativa trouxeram novo impulso ao tema (Khan, 2023) e novas questões, relacionadas à computação na nuvem e aos vastos estoques de dados de treinamento.

Gráfico 4.1- BigTechs reguladas pelo Digital Market Act - Serviços de plataforma fornecidos pelos “gatekeepers” designados pelo DMA da União Europeia



Fonte: European Commission apud Statista. Disponível em: <https://www.statista.com/chart/31876/companies-and-services-regulated-under-the-digital-markets-act/>. Acesso em: 28 jun. 2024

O avanço da IA generativa trouxe esses novos temas para a política da concorrência/antitruste, em especial a concentração em poucas empresas do controle sobre os ativos necessários para o treinamento dos modelos, como os conjuntos de dados excepcionalmente grandes e a capacidade computacional extrema. Efeitos potenciais de *lock-in*, conluio, monopolização, fusões, discriminação de preços e práticas de concorrência injustas relacionadas à IA generativa começam a emergir.

O trecho a seguir extraído do blog da FTC evidencia a preocupação sobre o controle sobre esses ativos:

“Generative AI depends on a set of necessary inputs. If a single company or a handful of firms control one or several of these essential inputs, they may be able to leverage their control to dampen or distort competition in generative AI markets. And if generative AI itself becomes an increasingly critical tool, then those who control its essential inputs could wield outsized influence over a significant swath of economic activity.” (FTC, 2023a).

Mesmo sem aprovação de nova legislação antitruste nos EUA, a FTC acredita dispor de instrumentos de controle com base na legislação vigente - incluindo o FTC Act¹⁸⁰ e

¹⁸⁰ Disponível em: <https://www.ftc.gov/legal-library/browse/statutes/federal-trade-commission-act>. Acesso em 22 ago. 2024

leis vigentes, como o *Equal Credit Opportunity Act*¹⁸¹, que proíbem discriminação, coleta ou uso exploratório de dados pessoais (Khan, 2023) – e investigações foram abertas visando aprofundar o entendimento das relações entre empresas inovadoras de menor porte e provedores de computação na nuvem nas parcerias da Microsoft e OpenAI, Amazon e Anthropic e Google e Anthropic (FTC, 2024).

A computação na nuvem, segunda geração de plataformas e embrião do modelo de negócios para a implementação da IA, é um ativo sob forte controle do chamado subconjunto “infraestrutural” das BigTechs, liderado por Amazon e Microsoft, com 31% e 25% do mercado, respectivamente, seguidas pelo Google Cloud (11%).

O subconjunto ou “núcleo infraestrutural” engloba as cinco BigTechs americanas e as chinesas Alibaba e Tencent (Fernandez et al., 2020)¹⁸², empresas verticalmente integradas que internalizaram os ativos complementares da IA (Jacobides, Brusoni e Candelon, 2021). Ainda que a maior parcela do mercado de infraestrutura na nuvem pertença a três BigTechs infraestruturais americanas, além das chinesas, está ocorrendo a incorporação de novos *players*, empresas incumbentes de alta tecnologia e mesmo a Nvidia, o que amplia seu papel de provedor chave de chips de IA (The Economist 2023d; The Economist 2023e).

A infraestrutura de computação está se tornando a principal barreira à entrada para as startups com modelos promissores da IA e a concentração infraestrutural tem implicações potenciais sobre a concorrência. Primeiro, a mudança de paradigma da inovação no sentido de modelos cada vez maiores – impulsionados pelas próprias BigTechs – significa a dependência estrutural das empresas menores. Sua falta de capacidade financeira (e técnica) para construir infraestrutura proprietária, torna inexorável que se convertam em clientes cativos das poderosas plataformas na nuvem das BigTechs, cujos preços são aparentemente reduzidos (Mickle, Metz e Grant, 2023), mas cujo custo marginal dos serviços e algoritmos fornecidos na nuvem é próximo de zero para os provedores, uma vez que linhas de códigos podem ser usadas simultaneamente por vários *players* (Rikap 2022), o que resulta em lucros expressivos (APNews, 2023).

Segundo, empresas menores se submetem ao aprisionamento (*lock-in*), inclusive

¹⁸¹ Disponível em: <https://www.ftc.gov/legal-library/browse/statutes/equal-credit-opportunity-act>. Acesso em: 20 ago. 2024

¹⁸² A noção aqui empregada de subconjunto infraestrutural difere um pouco do que Fernandez et al. (2020). Para esses autores, o “núcleo infraestrutural”, ou seja “as plataformas infraestruturais” são aquelas “...sobre as quais muitas outras plataformas e aplicações podem ser construídas” (Van Dijck, Poell e De Waal, 2018:12-13). Isso inclui os ativos de hardware que são aqui privilegiados, mas também outros mecanismos de toda plataforma digital, como os exemplos mencionados pelo autor, como a dependência do Uber à solução do Google Maps da Alphabet. A situação que focamos nesta seção é mais relacionada a casos como da dependência do Netflix à nuvem da AWS.

aceitando recursos complementares fornecidos pelo provedor como condição para prestação da computação na nuvem e/ou porque os códigos fornecidos na nuvem não são inteiramente disponibilizados aos clientes (Rikap, 2023c), agregando serviços à infraestrutura de nuvem (Kak, West e Whittaker, 2023). Isso vem sendo considerado potencialmente nocivo à concorrência no campo da IA, por meio de agrupamento (*bundling*, quando uma empresa oferece vários produtos juntos em um pacote único) e/ou vinculação (*tying*, quando uma empresa condiciona a venda de um produto à compra de outro) (FTC, 2023b). À medida em que as tecnologias de IA passem a constituir a principal fonte da inovação em diversos setores *downstream*, criarão forte dependência das BigTechs e suas estratégias expansionistas setoriais (Rikap, 2023d).

Terceiro, as vantagens infraestruturais permitem os subsídios cruzados entre líderes incumbentes e empresas emergentes “eleitas”, desde subsidiárias a empresas com participações minoritárias ou parcerias estratégicas, por meio do fornecimento subsidiado de recursos computacionais a preço de custo, em detrimento de outras empresas emergentes que não desejem ou não despertem interesse para vínculos acionários ou estratégicos, ou investimentos por meio de fundos de capital de risco proprietários das BigTechs – um investimento sem risco, que retorna por meio dos pagamentos pelos serviços de computação na nuvem.

Outra barreira à entrada da IA generativa é o acesso a gigantescos conjuntos de dados, que as empresas emergentes enfrentaram “raspando” toda a Internet, ao custo e risco das ações judiciais que estão enfrentando por violação de direitos autorais, frente à vantagem do pioneiro das gigantes do setor, que construíram amplos conjuntos de dados proprietários, mantidos em sigilo, a partir, principalmente, de ferramentas de coletas de dados nas plataformas digitais (FTC, 2023b). Nesse sentido, os reguladores têm igualmente abordado questões do controle sobre dados, não somente por questões relacionadas à privacidade, persuasão e outras, mas também pela assimetria entre os *players* da IA.

A FTC (2023b) também menciona os crescentes requisitos de recursos humanos nos modelos maiores e demandam conhecimento especializado, tácito ou incorporado em softwares e estruturas próprias, como TensorFlow e Pytorch, para tirar o máximo proveito do hardware, passando a constituir barreira à entrada adicional, aprofundando o poder infraestrutural e econômico de poucas empresas, líderes do mercado de computação na nuvem, que se beneficiaram das vantagens do pioneirismo da IA de grande escala pelos efeitos de rede e aos elevados custos de mudança, como sublinhado na investigação do Congresso americano (House Digital Markets Report, 2020).

Entre as recentes ações do DoJ e FTC, além de estados americanos, contra práticas

abusivas dessas empresas, se destaca a investigação aberta em 2023 contra o Google sobre a limitação da concorrência pelo mecanismo de busca e negócios de publicidade digital, envolvendo pagamentos a Apple para inclusão do Google Search como serviço padrão em computadores e celulares desde o início dos anos 2000¹⁸³. A Alphabet atribui o sucesso do produto (com quase 90% do mercado) à sua qualidade, não à conduta anticompetitiva, superado rivais porque não inovaram e investiram adequadamente. A Microsoft, por seu turno, que incorporou IA generativa no Bing (Mehdi, 2023), parece apostar no apoio da regulação para enfrentar a concorrência. A recente decisão de um tribunal americano reconhecendo monopólio do Google no mercado de busca online pressagia um futuro incerto para um dos principais produtos do modelo de negócios das BigTechs (Robins-Early, 2024).

O investimento/participação da Microsoft na OpenAI enfrenta escrutínio antitruste no Reino Unido, nos Estados Unidos (Muvija, Mehta e Soni, 2023) e na União Europeia (Chee, 2024). Em depoimento, em março de 2024, a Microsoft tenta justificar suas parcerias como forma de competir com o Google, não para superar seu próprio atraso em IA¹⁸⁴. A incomparável vantagem competitiva da rival seria controlar amplos conjuntos de dados proprietários (Google Search Index, e YouTube que hospeda cerca de 14 bilhões de vídeos, que outros “devs.” de IA não tem acesso) e chips otimizados para IA. Além disso, mencionou a posição privilegiada do Google e também da Apple, com Google Assistant e Siri – esquecendo de mencionar a tentativa frustrada da Microsoft de competir nesse mercado por meio da Cortana (Ferreira, 2024). Além disso, argumentou que Google e Amazon investiram na Anthropic, rival da OpenAI, e da Salesforce e Nvidia na Cohere.

O Google, por outro lado, em carta à agência reguladora antitruste da Grã-Bretanha *Competition and Markets Authority* (CMA), pediu investigação da participação da Microsoft no mercado de nuvem do Reino Unido, alegando que suas práticas de licenciamento desencorajam os clientes a utilizar serviços concorrentes (Ferreira, 2023). Desde 2022, a Microsoft permite a interoperabilidade, mas mediante cobrança de preços mais altos para clientes de licenças do Windows ou de software que usem o Google Cloud ou AWS, em lugar do Azure¹⁸⁵. Há também uma investigação em curso sobre o mercado de computação, desde 2022, pelo órgão de regulação francês (*Autorité de la concurrence*, 2023).

¹⁸³ Argumenta-se que o Google paga mais de US\$ 10 bilhões por ano por acordos para garantir que seja o mecanismo de busca padrão em navegadores da web e dispositivos móveis. Também avaliará se os atrasos do Google em permitir que a Microsoft tenha acesso à sua ferramenta de publicidade de pesquisa Search Ads 360 equivalem a comportamento anticompetitivo ilegal.

¹⁸⁴ Cabe lembrar que a Microsoft é a BigTech que mais adquire empresas emergentes (Affeldt e Kesler, 2021).

¹⁸⁵ Em 2022, a AWS e o Azure detinham uma participação conjunta de 70 a 80% do mercado britânico de serviços de infraestrutura em nuvem, enquanto o Google, apenas 5 a 10%.

Dessa forma, o foco dos reguladores da concorrência no mundo hoje converge em investigar os serviços e infraestrutura de computação na nuvem. Segundo a Federal Trade Commission dos EUA:

“New entrants typically access compute by turning to cloud computing services, which provide high-performance compute resources on demand. However, cloud services can be expensive and are currently provided by only a handful of firms, raising the risk of anticompetitive practices” (FTC, 2023a)¹⁸⁶.

Em 23 de junho de 2024, a abordagem convergente das políticas antitruste/concorrência dos modelos de fundação ficou explícita com a divulgação da Declaração Conjunta¹⁸⁷ da Comissão Europeia, do FTC/DoJ e da CMA. A Declaração alerta sobre práticas prejudiciais à concorrência, como os riscos das empresas restringirem insumos essenciais (chips, computação, dados em escala e conhecimento técnico), inclusive por meio do controle dos canais de distribuição de IA ou serviços habilitados para IA (computação na nuvem), além de aproveitarem feedback e efeitos de rede para aumentar as barreiras à entrada e prejudicar a concorrência, estendendo seu poder nos mercados digitais para ecossistemas ou mercados de IA adjacentes, entre outros temas comuns da política antitruste recente.

Os signatários da Declaração se comprometem a usar os poderes que dispõem “para lidar com esses riscos antes que eles se tornem danos arraigados ou irreversíveis”, com base na experiência nos mercados digitais. Exigirão negociações justas, privilegiando a interoperabilidade e opções de escolha pelos usuários, empreenderão o monitoramento de mecanismos de *lock-in*, monopsonios e táticas de exclusão decorrentes das parcerias, fusões ou outras formas de controle das empresas incumbentes sobre algumas emergentes, além do combate à fixação algorítmica de preços, discriminação e compartilhamento de informações sensíveis e uso de dados de usuários para treinamento dos modelos – proibição que já levou a Meta a suspender a IA generativa na UE, enfrentar dificuldades no Brasil (em função de decisão da agência reguladora de dados, ANPD) e dos questionamentos recentes do LinkedIn (Microsoft) pelo uso de dados de usuários. A depender de como avance a política antitruste,

¹⁸⁶ West e Vipra (2023) assinalam que a concentração no mercado da computação em nuvem e a força no desenvolvimento da IA se reforçam mutuamente, em virtude da escalabilidade dos modelos com dependência de grandes quantidades de dados e poder computacional, conferindo significativo poder de mercado ao pequeno grupo de empresas que controlam esses recursos, o núcleo infraestrutural das BigTechs e que operam ecossistemas de plataformas, usufruindo da vantagem de serem pioneiros na IA em grande escala, pelos efeitos de rede e elevados custos de mudança. Em virtude de significativas barreiras à entrada representadas pelos elevados custos iniciais, pela falta de interoperabilidade em pontos da pilha de computação e gargalos na cadeia de fornecimento de componentes-chaves da infraestrutura de computação, as pequenas empresas que atuam em mercados de IA precisam garantir créditos de computação ou fazer outros acordos contratuais com grandes empresas tecnológicas. As autoras mencionam ainda os enormes requisitos em talentos à medida que aumenta a computação pela necessidade de conhecimento especializado e tácito para tirar o máximo partido do hardware escasso (Musser et al., 2023).

¹⁸⁷ Disponível em: <https://www.gov.uk/government/publications/joint-statement-on-competition-in-generative-ai-foundation-models-and-ai-products/joint-statement-on-competition-in-generative-ai-foundation-models-and-ai-products>. Acesso 25 jun. 2024

poderá haver consequências sobre a inovação da IA generativa, sobre a arquitetura da indústria e ecossistemas das BigTechs, com potencial de redefinir vencedores e perdedores da corrida tecnológica.

4.3. Políticas industriais e comerciais de chips e IA generativa

A grande maioria das economias avançadas parece começar a enfrentar a era da IA com políticas públicas ativas. A política americana visando manter a superioridade tecnológica em semicondutores, chips e IA, principalmente com respeito à segurança nacional e defesa, parece uma preocupação central diante da competição chinesa, conforme as recomendações do relatório da *National Security Commission on Artificial Intelligence*¹⁸⁸. A introdução do relatório resume os objetivos estratégicos relacionados ao tema:

“We are at the beginning of this new era of competition. We now know the uses of AI in all aspects of life will grow and the pace of innovation will accelerate. *We know adversaries are determined to turn AI capabilities against us. We know a competitor is determined to surpass us in AI leadership.* We know AI is accelerating breakthroughs in a wide array of fields. We know that whoever translates AI developments into applications first will have the advantage. Now we must act. The principles we establish, the federal investments we make, the national security applications we field, the organizations we redesign, the partnerships we forge, the coalitions we build, and the talent we cultivate will set America's strategic course. The United States should invest what it takes to maintain its innovation leadership, to responsibly use to defend free people and free societies, and to advance the frontiers of science for the benefit of all humanity. AI is going to reorganize the world. America must lead the charge” (grifos da autora) (NSCAI, 2021:28).

O principal instrumento da política do governo americano para enfrentar a rivalidade chinesa e manter a liderança tecnológica em IA são as medidas de política industrial expressas no US Chips and Science Act (Public Law 117–167/2022)¹⁸⁹, de agosto de 2022, (precedido pela Lei de Criação de Incentivos Úteis para a Produção de Semicondutores (CHIPS) para a América, (PL 116-283), complementadas por políticas comerciais de impacto imediato, com restrições à exportação de chips para a China. Essas políticas resultaram de forte lobby da

¹⁸⁸ O relatório da NSCAI, coordenado por Eric Schmidt, ex-CEO do Google, e Robert Work, ex-vice-secretário de defesa dos governos Obama e Trump, foi solicitado por Comissão bipartidária do Congresso americano, criada em 2018, visando recomendações relacionadas a IA, aprendizado de máquina e tecnologias associadas, cessando suas atividades em outubro de 2021. O relatório apresentou recomendações que demandavam investimentos governamentais substantivos para impedir o atraso comparado a China. Outros estudos, como SCSP (2023) e do *Center for Security and Emerging Technologies* (CSET) da Georgetown University, também compartilham a preocupação com a rivalidade chinesa. Sobre a abordagem da indústria sobre o tema dos semicondutores, ver SIA (2021)

¹⁸⁹ Disponível em: https://science.house.gov/imo/media/doc/the_chips_and_science_act.pdf. Acesso em: 18 ago. 2023. A política industrial americana sobre chips deve ser consideradas em conjunto com a *United States Innovation and Competition Act - USICA*), legislação ampla de competitividade, lançada pela ala democrata do Congresso objetivando a retomada da liderança econômica americana e o enfrentamento da China por meio de forte apoio à ciência, mas cuja versão fina, após amplo debate terminou como uma lei de apoio à indústria de semicondutores. A proposta original foi inspirada no relatório “*Science – the Endless Frontier*” de Vannevar Bush, diretor do *Office of Scientific Research and Development*, ao Presidente Roosevelt, no imediato pós-segunda guerra.

indústria, conforme pode ser observado em Carta da SIA ao governo Biden, na qual solicita subsídios para competitividade e produção de semicondutores no país (SIA, 2021).

O *US Chips and Science Act* contempla medidas de médio e longo prazo por meio de incentivos a P&D e aumento da produção de chips nos EUA (e países aliados) e fabricação de equipamentos especializados. A lei estabeleceu investimentos e incentivos de US\$ 52 bilhões para fabricação, P&D e segurança na cadeia de semicondutores, deixando de fora a dependência de minerais estratégicos da China, como será visto.

A preocupação com o desequilíbrio entre indústria e pesquisa acadêmica em IA também estimulou iniciativas como o programa piloto *National AI Research Resource* (NAIRR)¹⁹⁰, coordenado pela National Science Foundation, visando construção/compartilhamento de infraestrutura computacional acadêmica e de pesquisa. Em 2023, a NSF anunciou investimento de US\$ 140 milhões no projeto, objetivando lançar sete Institutos Nacionais de Pesquisa de IA, liderados por universidades, para uso de IA na mitigação de efeitos da mudança climática e melhorias da educação. No entanto, a iniciativa também conta com a contribuição de recursos de computação de empresas de tecnologia, como a doação de US\$ 20 milhões em créditos de computação Azure da Microsoft e US\$ 30 milhões da Nvidia (US\$ 24 milhões para acesso de computação em sua plataforma DGX) (Madison, 2024).

Além disso, com impactos a curto prazo, o Departamento de Comércio americano passou a exigir, desde outubro de 2022, licenças de exportação para China e Macau de chips mais sofisticados, fundamentais para a inovação de modelos da IA generativa, em usos civis e militares (2022 IFR)¹⁹¹. A política comercial visou conter a exportação de chips de alto desempenho, ou seja, chips lógicos com velocidade de transferência chip a chip superior a 600 gigabytes por segundo e poder computacional superior a 600 TOPS, atingindo diretamente os modelos A100 e H100 da Nvidia.

A China é o maior mercado de semicondutores do mundo (cerca de 37% do mercado) e apresenta as maiores taxas de crescimento, respondendo pela maior parcela da receita das empresas especializadas. O mercado chinês representou um quarto da receita anual total da Nvidia nos últimos anos e 25% e 20% da Intel e Micron Technologies, além de metade da receita da Qorvo, Texas Instruments e Broadcom e cerca de 60% da Qualcomm. Esses números levantam dúvidas se interesses corporativos se subordinariam às preocupações geopolíticas, o que parece explicar os planos de apoio do governo americano a essas empresas.

¹⁹⁰ Disponível em: <https://nairrpilot.org/>, Acesso em 16 jul. 2023

¹⁹¹ Disponível em: <https://www.govinfo.gov/content/pkg/FR-2022-10-13/pdf/2022-21658.pdf>. Acesso em 25 out. 2023

Chips de menor velocidade/menor poder de processamento recém desenvolvidos pela Nvidia escapavam ao controle de exportações e começaram a ser vendidos para a China, como o A800, cuja velocidade de transferência chip a chip é mais baixa (400 gigabytes por segundo), a partir de novembro de 2022, e do H800, desde março de 2023. Serviços em nuvem dentro ou fora do país também podiam ser usados para treinar modelos de empresas chinesas (Harris, 2023). Com isso, era crescente a pressão do Congresso americano por controles mais rigorosos, principalmente depois do lançamento do novo celular da Huawei, em setembro de 2023, com chips que desejavam manter fora do alcance chinês (Hawkins, 2023).

Uma nova rodada de restrições à exportação veio no ano seguinte, incluindo esses chips menos sofisticados e partes da cadeia de valor de semicondutores de IA, além de estender os requisitos de licença para outros países. Regras mais abrangentes foram divulgadas em outubro de 2023 pelo *Bureau of Industry and Security* (BIS) do Departamento de Comércio americano, estendendo as restrições aos modelos Nvidia A800, H800, L40, L40S e RTX 4090, bem como das rivais AMD e Intel, além de softwares especializados e equipamentos de fabricação de semicondutores que utilizem tecnologia americana (das empresas Applied Materials, Lam e KLA), podendo afetar empresas não americanas (a holandesa ASML), bem como serviços na nuvem prestados pelos CSP americanos (Powell et al., 2023) usados dentro ou fora do país para treinamento de modelos. Vale lembrar da obrigatoriedade de informações sobre venda de serviços na nuvem impostas pela Ordem Executiva 14.110/2023. Os controles também exigem licença para instalação de fabricação de chips avançados por empresas de *design* de chips sediadas na China se o fabricante usar tecnologia ou software de origem americana (*Advanced Computing Foreign Direct Product Rule*), atingindo empresas estrangeiras, como a taiwanesa TSMC (que produzia chips avançados para empresas chinesas de *design* de chips de IA, como Biren) (Pike, 2024; Gibson Dunn, 2023).

As restrições afetam empresas de nações “aliadas”, como Holanda – além de Japão e Coreia do Sul –, que desde 2018 sofre pressão dos EUA para adotar medidas restritivas (Amaro e Kharpal, 2023) à exportação de equipamentos de litografia ultravioleta extremo da ASLM (com preços de cerca de US\$ 200 milhões), usados por fabricantes de chips 5nm ou menor, litografia de imersão e deposição avançada, que passaram a depender de licença do governo holandês (Kharpal, 2022; Flaherty, 2023).

Não obstante, os efeitos das políticas protecionistas americanas podem ser contrários ao pretendido. Na década de 1990, as restrições americanas ao acesso de tecnologias de satélite e GPS foram enfrentadas com políticas públicas ativas para produção de satélites chineses e para construção do sistema chinês BeiDou Global Navigation Satellite System (GNSS), atualmente

o maior do mundo. As restrições aos componentes da ZTE e Huawei, acabaram estimulando a construção de um amplo ecossistema doméstico e avanços importantes nas tecnologias digitais críticas, como comunicação sem fio, semicondutores e IA.

No caso dos semicondutores, a resposta chinesa ao protecionismo americano contemplou políticas industriais e incentivos para a produção local. O *National Integrated Circuit Industry Investment Fund*, conhecido como “Grande Fundo” estatal¹⁹², criado em 2014, estabeleceu em 2024 sua terceira etapa de financiamento ao desenvolvimento de chips de última geração com vistas à reduzir a dependência de tecnologia estrangeira diante das barreiras comerciais dos EUA e aliados (Holanda, Alemanha, Coreia do Sul e Japão). Conta com recursos dos governos provinciais e empresas estatais (a firma de investimentos Chengtong, a geradora de energia *State Development*, etc.) e uma pequena parcela do governo central. O fundo já liberou US\$ 45 bilhões, tornando-se acionista de startups e 48 empresas locais de chips, com uma rodada adicional de US\$ 70 bilhões destinados a incentivar empresas como SMIC, *Yangtze Memory Technologies*, *Changxin Xinqiao Memory Technologies*.

O governo chinês resgiu impondo barreiras comerciais à exportação de minerais críticos para produção de chips, dos quais os EUA é grande importador, também sob o argumento de salvaguardar segurança e interesses nacionais. O Anúncio nº 23¹⁹³, de julho de 2023, do Ministério do Comércio e da Administração Geral das Alfândegas, estabeleceu a exigência de licença de exportação para oito itens de gálio e seis de germânio (Cytera, 2023). A China controla a oferta mundial desses minerais, com um quase monopólio em gálio¹⁹⁴ - após minar concorrentes na década de 2010 e tornando inviável a produção fora da China - e quase 70% do germânio (Force Distance Times, 2022).

Ainda que no curto prazo os fabricantes de semicondutores afirmem ter estoques suficientes, no médio e longo prazo será necessário retomar a produção em instalações

¹⁹² Disponível em: <https://oglobo.globo.com/economia/tecnologia/noticia/2024/03/08/fundo-secreto-de-us-70-bi-e-a-estrategia-chinesa-para-financiar-chips-de-ultima-geracao-entenda.ghtml>. Acesso em 08 mar.2024

¹⁹³ Disponível em: <http://www.mofcom.gov.cn/article/zwgk/gkzcfb/202307/20230703419666.shtml>. Acesso em: 12 nov. 2023

¹⁹⁴ Os produtos que passaram a exigir licença de exportação são nitreto de gálio, óxido de gálio, fosfeto de gálio, arseneto de gálio, arsênio de índio gálio, seleneto de gálio, antimônio de gálio, bem como lingotes de germânio fundido distrital, fosfato de germânio de zinco, substratos de crescimento epitaxial de germânio, dióxido de germânio, tetracloreto de germânio e entre outros itens. O gálio é um insumo cada vez mais crítico, com o arseneto de gálio usado em chips de radiofrequência para celulares e comunicação via satélite (radares e dispositivos de comunicação de rádio, satélites e LED), por suportar frequências e temperaturas mais altas que o silício. O germânio é usado em aplicações críticas como câmeras de imagem térmica, painéis solares, fibra ótica para telecomunicações. Esses elementos são obtidos como subprodutos da mineração de outros elementos, como zinco e cobre – onde a China é o primeiro e segundo produtor mundial, respectivamente - em quantidades pequenas e alto custo. Existem materiais substitutos em certas aplicações, mas cada um tem propriedades físicas, desempenho elétrico, escalabilidade, eficiência e custo, de modo que a escolha depende das características desejadas para cada chip específico. Nenhum gálio doméstico dos EUA foi recuperado desde 1987.

ocidentais, como EUA, Austrália e países produtores de zinco e cobre. O gálio e o germânio são componentes críticos para uma variedade de dispositivos eletrônicos avançados e as restrições chinesas poderão gerar aumento de custos, escassez da oferta e investimento em substitutos e reciclagem.

É grande a vulnerabilidade americana, dependendo de importações de minerais estratégicos de diversos países, principalmente adversários geopolíticos, como no caso do gálio, 100% do consumo aparente vindo de importações (Quadro 4.2). O impacto pode alcançar outros países e indústrias, como o Japão, que importa 60% do gálio que consome, 70% proveniente da China (Shilov, 2023b), além da União Europeia, criando vulnerabilidade às flutuações de preços e a riscos geopolíticos (Draghi, 2024A:52).

A preocupação do governo americano já era clara em Comunicado Casa Branca, de fevereiro de 2022: *“The U.S. is increasingly dependent on foreign sources for many of the processed versions of these minerals. Globally, China controls most of the market for processing and refining for cobalt, lithium, rare earths and other critical minerals.”* (The White House, 2022)

Essa dependência tem levado a busca incessante de fontes alternativas e diversificação de fornecedores para países como Austrália, Canadá, países europeus como Alemanha, França, Bélgica e Holanda, além de latino-americanos, como Chile e Peru, embora seja necessário de três a cinco anos para construir infraestrutura de extração/processamento, recrutamento de mão-de-obra e licenças operacionais e ambientais.

No caso da União Europeia, a política relacionada a IA, ao contrário do que muitos supõem, não se restringe a iniciativas regulatórias, contemplando políticas públicas ativas dentro da noção de soberania digital (EU Commission EPSC, 2019; Madiega, 2020) que avançaram no período recente com o EU Chips Act (2022)¹⁹⁵, com concessão de incentivos ao desenvolvimento/produção de chips nos países do bloco. A estratégia de inovação em IA divulgada pela Comissão Europeia após a aprovação do AI Act (Cooper et al., 2024) menciona a criação do Gabinete Europeu para a IA¹⁹⁶; do espaço comum europeu de dados; investimentos de €\$ 4 bilhões, até 2027; medidas de apoio a pequenas e médias empresas, criação de Fábricas de IA com infraestrutura de supercomputação e recursos humanos locais; e a Iniciativa GenAI4EU, relacionada ao desenvolvimento de casos de utilização de IA em diversos setores.

¹⁹⁵ Disponível em: https://eur-lex.europa.eu/resource.html?uri=cellar:ca05000a-89d4-11ec-8c40-01aa75ed71a1.0005.02/DOC_1&format=PDF. Acesso 03 mar.2022

¹⁹⁶ Disponível em: <https://digital-strategy.ec.europa.eu/en/library/commission-decision-establishing-european-ai-office>. Acesso em: 30 jan.2024

Quadro 4.2 -Dependência Líquida de Importação dos EUA em 2023

Commodity	Net import reliance as a percentage of apparent consumption in 2023	Leading import sources (2019–22) ²
ARSENIC, all forms	100	China, ³ Morocco, Malaysia, Belgium
ASBESTOS	100	Brazil, Russia
CESIUM	100	Germany
FLUORSPAR	100	Mexico, Vietnam, China, South Africa
GALLIUM	100	Japan, China, Germany, Canada
GRAPHITE (NATURAL)	100	China, ³ Mexico, Canada, Madagascar
INDIUM	100	Republic of Korea, Canada, Belgium
MANGANESE	100	Gabon, South Africa, Australia, Georgia
MICA (NATURAL), sheet	100	China, Brazil, India, Belgium
NIOBIUM (COLUMBIUM)	100	Brazil, Canada
RUBIDIUM	100	China, Germany, Russia
SCANDIUM	100	Japan, China, Germany, Philippines
STRONTIUM	100	Mexico, Germany, China
TANTALUM	100	China, ³ Germany, Australia, Indonesia
YTTRIUM	100	China, ³ Germany, France, Republic of Korea
GEMSTONES	99	India, Israel, Belgium, South Africa
ABRASIVES, fused aluminum oxide	>95	China, ³ Canada, Brazil, Austria
NEPHELINE SYENITE	>95	Canada
RARE EARTHS, ⁴ compounds and metals	>95	China, ³ Malaysia, Japan, Estonia
TITANIUM, sponge metal	>95	Japan, Kazakhstan, Saudi Arabia, Ukraine
BISMUTH	94	China, ³ Republic of Korea, Belgium, Mexico
POTASH	91	Canada, Russia, Belarus
STONE (DIMENSION)	87	Brazil, China, ³ Italy, Turkey
DIAMOND (INDUSTRIAL), stones	84	India, South Africa, Russia, Congo (Kinshasa)
PLATINUM	83	South Africa, Switzerland, Germany, Belgium
ANTIMONY, metal and oxide	82	China, ³ Belgium, India, Bolivia
ZINC, refined	77	Canada, Mexico, Peru, Republic of Korea
BARITE	>75	India, China, ³ Morocco, Mexico
BAUXITE	>75	Jamaica, Turkey, Guyana, Australia
IRON OXIDE PIGMENTS, natural and synthetic	75	China, ³ Germany, Brazil, Canada
TITANIUM MINERAL CONCENTRATES	75	South Africa, Madagascar, Australia, Canada
CHROMIUM, all forms	74	South Africa, Kazakhstan, Russia, Canada
PEAT	74	Canada
TIN, refined	74	Peru, Bolivia, Indonesia, Malaysia
ABRASIVES, silicon carbide	73	China, ³ Brazil, Canada, Netherlands
SILVER	69	Mexico, Canada, Poland, Switzerland
COBALT	67	Norway, Canada, Finland, Japan
GARNET (INDUSTRIAL)	67	South Africa, Australia, China, ³ India
RHENIUM	60	Chile, Canada, Germany, Kazakhstan
ALUMINA	59	Brazil, Australia, Jamaica, Canada
VANADIUM	58	Canada, Brazil, Austria, Russia
NICKEL	57	Canada, Norway, Finland, Russia
DIAMOND (INDUSTRIAL), bort, grit, and dust and powder	56	China, ³ Republic of Korea, Ireland, Russia
MAGNESIUM COMPOUNDS	52	China, ³ Israel, Canada, Brazil
GERMANIUM	>50	Belgium, China, Canada
IODINE	>50	Chile, Japan
MAGNESIUM METAL	>50	Canada, China, ³ Israel, Taiwan
SELENIUM	>50	Philippines, Mexico, Germany, Canada
TUNGSTEN	>50	China, ³ Germany, Bolivia, Vietnam
SILICON, metal and ferrosilicon	<50	Brazil, Russia, Canada, Norway
COPPER, refined	46	Chile, Canada, Mexico
ALUMINIUM	44	Canada, United Arab Emirates, Bahrain, Russia
PALLADIUM	37	Russia, South Africa, Italy, Canada
LEAD, refined	35	Canada, Mexico, Republic of Korea, Australia
MICA (NATURAL), scrap and flake	28	China, Canada, India, Finland
PERLITE	26	Greece, China, Mexico
LITHIUM	>25	Argentina, Chile, China, Russia
TELLURIUM	>25	Canada, Germany, Philippines, Japan
SALT	25	Canada, Chile, Mexico, Egypt
BROMINE	<25	Israel, Jordan, China ³
ZIRCONIUM, ores and concentrates	<25	South Africa, Australia, Senegal, Russia
CEMENT	22	Turkey, Canada, Greece, Mexico
VERMICULITE	20	South Africa, Brazil, Zimbabwe

Fonte: US Geological Survey (2024)

A proatividade europeia deverá se ampliar diante do diagnóstico contundente do relatório Draghi (2024A; 2024B) sobre a competitividade do bloco, cuja indústria se apoia em tecnologias maduras e atraso em tecnologias emergentes. Apenas quatro das cinquenta principais empresas de tecnologia são da União Europeia, nenhuma com capitalização de mercado superior a € 100 bilhões (US\$ 100 bilhões) e que no seu conjunto representam somente 7% do investimento em P&D das principais empresas de software e internet do mundo (frente a 71% por empresas americanas e 15% por chinesas) - em hardware e equipamento eletrônico, as empresas europeias respondem por 12% da P&D global (40% empresas americanas e 19% chinesas). Só quatro das 50 maiores plataformas digitais são da UE, com um mercado de computação na nuvem de apenas € 87 bilhões em 2022 (Draghi, 2024B:77).

O relatório, encomendado pela Comissão Europeia, prioriza três áreas: IA, descarbonização e redução da dependência de insumos críticos, propondo políticas setoriais e horizontais requeridas para alcançar as ambiciosas metas de competitividade europeia. Por conseguinte, deverá acelerar as políticas industriais digitais do bloco econômico, que requer investimentos mínimos anuais de até € 800 bilhões (4,7% do PIB da UE em 2023, um esforço muito superior ao 1-2% do PIB da UE no âmbito do Plano Marshall), que não prescindirão de recursos públicos.

Somam-se a isso iniciativas nacionais, com destaque na Alemanha, França e Itália. O presidente Macron reafirmou que a IA é uma prioridade estratégica para a França, dedicando quase € 2,5 bilhões do France 2030 que, acrescido de outros programas, deverá alcançar €7 bilhões, além de investimentos de grandes grupos locais¹⁹⁷. O programa Italia Digital, por seu turno, está destinando € 1 bilhão para IA (Bichicchi, 2024), enquanto o apoio governamental alemão de € 1,75 bilhão a startups de IA, além de articular recursos públicos e privados (Digital Watch Observatory, 2024), por meio do *German Future Fund*¹⁹⁸ (KfW), do *European Recovery Program*¹⁹⁹ e investimentos das BigTechs, como os €3 bilhões da Microsoft para IA e infraestrutura (Bundesregierung, 2024). Os instrumentos financeiros da Política Digital da União Europeia são amplos e dispersos, com prioridade para investimentos em IA, com destaque para Itália, Espanha e Dinamarca (Cingari, 2024; European Commission, 2023).

4.4. Tratados internacionais de comércio como instrumento de proteção das tecnologias da IA

Outra forma de intervenção com possíveis implicações sobre a arquitetura da indústria e da apropriabilidade da IA se relaciona à inclusão de temas da economia digital em acordos internacionais de comércio pelo governo americano, seguindo a tradição de negociar acordos que atendam aos interesses de grandes corporações americanas, diante do avanço de regulações nacionais restritivas sobre os usos da IA, privacidade, transferência de dados transfronteiriços, antitruste e propriedade intelectual. Embora acordos comerciais não impeçam que seus países membros aprovem legislação que contradiga seus termos, podem levar a disputas comerciais, imposição de tarifas ou mesmo a defesa da saída de um país-membro.

Dispositivos relacionados ao comércio eletrônico e à transferência transfronteiriça de dados constaram de acordos firmados no governo Trump, buscando uma harmonização

¹⁹⁷ Disponível em: <https://www.culture.gouv.fr/en/press/press-releases/france-2030-ai-as-an-accelerator-and-differentiator-of-innovation>. Acesso em: 10 ago. 2024

¹⁹⁸ Disponível em: https://www.eif.org/what_we_do/resources/gff-eif-growth-facility/index.htm. Acesso em: 02 ago. 2024

¹⁹⁹ Fundo criado para recuperação econômica da Europa pós-Covid. Disponível em: <https://www.bmwk.de/Redaktion/DE/Artikel/Mittelstand/erp-sondervermoegen.html>. Acesso em: 27 jul.2024

internacional no sentido de garantir sigilo algorítmico, do código fonte e de informações técnicas dos modelos de IA (Rangel e Wallach, 2023; Burri, 2021). Isso ocorreu com o acordo de livre comércio entre EUA, Canadá e México, o chamado USMCA (*US-Mexico-Canada Agreement*, ou Nafta 2.0), de 2018, por meio de regras de "não discriminação" e a legislação e regulação nascente dos países como combater "barreiras comerciais digitais" e efeitos discriminatórios sobre empresas americanas (USTR, 2017). Disposições similares foram negociadas em tratados comerciais regionais e bilaterais, adotadas por 18 países, e havia a intenção de inserir o tema do comércio digital na agenda da Organização Mundial do Comércio (OMC).

A política da concorrência e a regulação dos usos da IA da União Europeia, os requisitos governamentais chineses de acesso ao código-fonte da IA foram interpretadas como discriminatórias, controle político, espionagem e restrições para favorecer a construção de uma indústria doméstica (McCabe e Swanson, 2019). A proposta de criação de um imposto sobre serviços digitais da França, em 2019, levou o *U.S. Trade Representative* (USTR) a ameaçar investigação sobre discriminação de empresas americanas, como ocorreu em outros dez países. As regras tributárias, de segurança cibernética e privacidade de dados do Vietnã (Hoang, 2021), a lei que obriga a instalação de escritórios das BigTechs na Rússia (Reuters, 2021), a política da concorrência da Índia (PWOnlyIAS, 2023) e a lei de AppStore da Coreia - que obriga AppStore e Google Play a aceitar outros sistemas de pagamento que não os próprios (Park, 2023) - foram consideradas contrárias ao princípio da não discriminação.

A intenção de estender tais dispositivos a outros acordos de comércio, incluindo o *Indo-Pacific Economic Framework* (IPEF), já no governo Biden, criando regras supranacionais (Gao, 2021), esbarrou na reação da sociedade e do Congresso americano, que apontavam o *lobby* das empresas digitais líderes e de associações de classe. O vazamento de emails de funcionários do USTR pedindo orientações a ex-colegas das BigTechs sobre as novas legislações nacionais foi citado em relatório do Congresso (Warren, 2023) e parecia ser um caso de captura regulatória (Li, 2023).

Minuta do IPEF, divulgada pelo USTR, que já contava com apoio do Japão, Singapura e Austrália, liderando uma iniciativa conjunta na OMC, incluía disposições contra barreiras ao comércio digital, impedindo que qualquer país membro conferisse tratamento diferenciado a produtos/serviços digitais de outro país, incluindo fluxos de dados, acesso ao código-fonte e

dados das empresas, nem a imposição de tarifas adicionais ao comércio digital²⁰⁰. Em Carta²⁰¹ ao USTR, de abril de 2023, um grupo de Senadores (incluindo a senadora Warren) critica a tentativa de elaborar um quadro comercial sobre um tema controvertido como comércio digital, limitando a atuação do Congresso.

O governo Biden acabou recuando e as negociações tomaram outro rumo, com os Estados Unidos anunciando, em novembro de 2023, a retirada da proposta para o comércio digital do IPEF e OMC - inclusive pela posição minoritária na OMC, com apenas 34% dos países apoiando a livre transferência de dados, com a maioria (57%) condicionando o fluxo transfronteiriço de dados às leis nacionais de privacidade/proteção de dados pessoais, e os 9% restantes (incluindo Rússia e China) defendendo políticas restritivas à transferência de dados (Fefer, 2021). Com isso, a nova abordagem americana para a economia digital, incluindo a regulamentação da IA (Ordem Executiva 14.110), pareceu reorientada para a construção de um novo consenso global (Stockman, 2023). Aparentemente, buscando conciliar as duas agendas conflitantes da política americana apontadas por Rodrik (2023): “*One agenda is inward-looking, focusing on the creation of an inclusive, resilient, prosperous, and sustainable American economy. The other focuses on geopolitical and on maintaining US primacy over China.*” No entanto, nada garante que futuros governos americanos voltem a recorrer a esse tipo de instrumento.

No lugar de um acordo de livre comércio abrangente com a União Europeia, foi estabelecido o *EU-US Trade and Technology Council (TCC)*²⁰², em 2021, cooperação em questões tecnológicas, econômicas e comerciais, apoiada em “valores democráticos orientados para o mercado”. A iniciativa se converteu em instrumento de monitoramento da cadeia de valor de semicondutores, colaboração em wireless 6G, tecnologias quânticas e roteiro conjunto de abordagem regulatória da IA – mesmo com divergências sobre o DMA e DSA, discriminatórios das empresas americanas - a partir de 2023, em função da necessidade de respostas coordenadas à guerra Rússia-Ucrânia e preocupações com a ascensão chinesa. Relacionado ao tema da IA e semicondutores, estão em curso negociações entre EUA e União Europeia sobre minerais críticos (*Critical Minerals Agreement*).

²⁰⁰ Disponível em: <https://ustr.gov/sites/default/files/files/agreements/FTA/USMCA/Text/19-Digital-Trade.pdf>. Acesso em: 22 jan. 2024

²⁰¹ Disponível em: <https://www.warren.senate.gov/imo/media/doc/FINAL%20Letter%20to%20Biden%20in%20Support%20of%20USTR%20Digital%20Trade%20Work.pdf>. Acesso em 09 jan.2024

²⁰² Disponível em: <https://ustr.gov/useuttc>. Acesso em: 30 set. 2024. O TTC reuniu-se cinco vezes (setembro de 2021, maio de 2022, dezembro de 2022, maio de 2023 e janeiro de 2024) com declarações conjuntas sobre cooperação em questões específicas (Akhtar, Schwarzenberg e Trachtenberg, 2024).

CAPÍTULO 5 – DIREITOS DE PROPRIEDADE INTELECTUAL DA INTELIGÊNCIA ARTIFICIAL

O sistema de PI corresponde ao arcabouço legal visando proteger criações do intelecto humano, conferindo a autores e inventores um direito exclusivo sobre sua obra ou invenção. Contempla a propriedade industrial, que abarca marcas, patentes de invenção e de modelos de utilidade, desenho industrial e indicações geográficas, e o direito autoral (ou *Copyright*) que protege obras artísticas, literárias e científicas, incluindo programas de computador e bases de dados subjacentes. Casos em que a singularidade e peculiaridade de uma invenção não são adequadamente contemplados pelos instrumentos existentes, tiveram sua proteção efetuada por meio de direitos *sui generis* - expressão em latim que significa “de seu próprio gênero” ou “de espécie única”. Os DPI costumam ser complementares, apesar de oferecerem ferramentas alternativas, como a proteção de software parece tipificar.

Cada um desses mecanismos tem condições e requisitos formais que devem ser cumpridos para fazer jus aos direitos. No direito autoral, cujo registro não é obrigatório para a obtenção de direitos, ao contrário das patentes, o que evidentemente tem implicações sobre sua divulgação para a sociedade. As criações devem atender a requisitos de criatividade, novidade e originalidade, não estabelecidos em TRIPS, mas definidos pela Convenção de Berna e detalhados nas respectivas legislações nacionais. Existe a previsão de limitações ou exceções aos direitos exclusivos, em casos especiais e desde que não conflitem com a exploração normal da obra e não prejudiquem injustificadamente os interesses legítimos do titular do direito (TRIPS, artigo 13).

As patentes de invenção contemplam invenções de produtos ou de processos, em todos os campos da tecnologia, mas precisam do depósito para fazer jus a direitos de propriedade, devendo atender aos requisitos de: novidade; envolver uma etapa inventiva (inventividade ou não obviedade, em algumas jurisdições²⁰³); e suscetível de aplicação industrial (ou utilidade, em algumas jurisdições) (TRIPS, artigo 27(1), sobre matéria patenteável). Além disso, a patente deve divulgar a invenção de maneira suficientemente clara e completa para que a invenção possa ser realizada por uma pessoa qualificada na arte (TRIPS, artigo 29(1)). Na prática, essas condições são detalhadas

As condições e requisitos de patenteabilidade, harmonizados internacionalmente em TRIPS, tiveram assegurada flexibilidade e são detalhados nas respectivas legislações nacionais, o que pode resultar em diferenças entre jurisdições. O requisito de novidade significa que a

²⁰³ As noções de utilidade e não-obviedade são usadas na lei de patentes do Canadá e dos Estados Unidos.

invenção não está incorporada ao estado da técnica, que é definido como o conhecimento existente antes da data do depósito/prioridade de um pedido de patente, por meio de divulgação escrita ou oral. Atividade inventiva (ou não-obviedade/inventividade, como é chamado em algumas jurisdições) é apresentar distância do estado da técnica para um “perito na técnica” - *person having ordinary skill in the art* – PHOSITA, no sentido da invenção acrescentar algo ao estado da técnica. A aplicação industrial significa estar sujeita a uma aplicação prática e produção demonstrável. É obrigatória a divulgação da invenção através de sua descrição detalhada, suficientemente clara e completa para que a invenção seja realizada pelo PHOSITA (Suficiência Descritiva).

Relatório European Commission (2020:97-98) identificou na Convenção de Paris²⁰⁴, de 1883 (revisada em 1967) e no Acordo TRIPS, de 1994, disposições relevantes para a discussão das patentes da IA, em particular relacionadas à figura do inventor, o tratamento nacional e a independência das patentes concedidas em cada jurisdição, o direito do inventor ser nomeado na patente, tratamento de nação mais favorecida e patentes devendo ser disponíveis para invenções em quaisquer campos da tecnologia quando atendam aos requisitos para patenteamento²⁰⁵.

Inovações da IA, em particular em redes neurais profundas, possuem muitas idiossincrasias frente a qualquer outra tecnologia existente, com implicações sobre o regime legal de PI. O presente Capítulo busca identificar os DPI referentes a esse tipo particular de invenção da IA (redes neurais profundas) e a BigData (conjuntos de dados de treinamento dos modelos). Observações metodológicas prévias são necessárias. Em primeiro lugar, embora o foco da tese recaia primordialmente sobre patentes de invenção, algumas reflexões sobre direito autoral são necessárias, tanto porque a IA é usualmente expressa por meio do software, quanto pela importância dos conjuntos de dados de treinamento dos modelos que são usualmente protegidos junto com o software.

Em segundo lugar, ainda que fosse pretendida uma abordagem mais geral e válida a ser seguida pelos diferentes países focando apenas em acordos internacionais, o caráter territorial da PI e as especificidades da IA acabaram exigindo um olhar mais atento sobre as legislações nacionais, tendo sido selecionadas apenas as nações líderes na corrida tecnológica da IA, como

²⁰⁴ Disponível em: https://www.wipo.int/treaties/en/ip/paris/summary_paris.html. Acesso em: 20 jan. 2021

²⁰⁵ O Tratado de Direito de Patentes (*Patent Law Treaty - PLT*), de 2000, ao qual poucos países aderiram, estabelece que o que constitui invenção será determinado de acordo com a lei aplicável (Regra 16(9) PLT). Disponível em: https://www.wipo.int/treaties/en/ip/plt/summary_plt.html. Acesso em: 12 mar. 2023. O Tratado de Cooperação de Patentes (*Patent Cooperation Treaty - PCT*), de 1970, permite (mas não exige) que os países solicitem dados sobre o inventor (Artigos 4 (1)(v), 22(1) e 27(3) PCT). Disponível em: <https://www.wipo.int/pct/en/texts/articles/atoc.html>. Acesso em: 15 mai. 2022

Estados Unidos e China, ou onde uma reflexão aprofundada sobre esses direitos tem procurado desenvolver *corpus* legislativos específicos, como no caso da União Europeia. O caso da China apresentará limitações óbvias, decorrentes da dificuldade de obtenção de informações completas.

Terceiro, a sobreposição de distintos DPI exige uma análise em duas etapas e tomou emprestada a metodologia empregada por Schirru (2020) na análise do direito autoral das criações da IA, baseada na abordagem da União Europeia que separa insumos e produtos da IA. Mesmo que nos Estados Unidos não há essa distinção de produtos e insumos da IA, com invenções e bancos de dados tratados em seu conjunto, esta abordagem tem a vantagem de facilitar o entendimento dos diferentes DPI da IA em toda a sua idiossincrasia. Mesmo que o foco da tese sejam as invenções, não as criações, foram necessárias algumas observações sobre esse outro instrumento de PI que protege o software, além de distinguir entre, de um lado, os insumos da IA (sistemas de IA e BigData) e, de outro, os produtos da IA (as invenções propriamente ditas)²⁰⁶.

O legado de um aparente duplo mecanismo de proteção do software, que teve os EUA como precursor, englobando direito autoral e, ao longo do tempo, também patentes, entrelaçou os DPI da IA e do software, mas contemplando diferentes direitos que protegem coisas díspares. Adicionalmente, a despeito das regras harmonizadas por acordos internacionais, o detalhamento das condições e dos requisitos do direito autoral e, principalmente, das patentes de invenção, os quais são essenciais para a concessão ou mesmo para o depósito, mostra diferenças importantes entre os países. Apesar das diferenças entre os sistemas jurídicos, em questões centrais para a racionalidade do sistema, parece haver uma certa convergência global no caso da IA, que, com isso, parece prescindir de harmonização formal, como sugere a resposta orquestrada em distintos escritórios nacionais e regionais de PI (ENPIs) aos depósitos dos pedidos de patente do sistema do DABUS, descrito no final deste capítulo.

5.1. O legado da proteção dupla do software

Como a IA depende da codificação de software, é importante rastrear inicialmente seu mecanismo original de PI, do direito autoral que, mais adiante, passou a contemplar também patentes de invenção.

²⁰⁶ Poucas jurisdições preveem direitos autorais para criações geradas por computador, como Hong Kong ([section 11\(3\), Copyright Ordinance](#)), Índia ([section 2\(d\)\(vi\), Copyright Act, 1957](#)), Irlanda ([Art. 21, Copyright and Related Rights Act 2000](#)), Nova Zelândia ([section 5\(2\)\(a\), Copyright Act 1994](#)), África do Sul ([section 2\(h\), Copyright Act 1978](#)) e Reino Unido ([Copyright, Designs and Patents Act 1988, section 9\(3\)](#)) (Hervey, 2023).

Os programas de computador, ou software, foram definidos pelo Acordo TRIPS, seguindo a tradição da Convenção de Berna, como uma forma de expressão, tornando-se, assim, elegível para proteção por direito autoral. No entanto, conforme observa por Reichman (1994), se antes do acordo TRIPS o status das patentes relacionadas a programas de computador na lei internacional de PI era indefinido, depois o status internacional de direitos autorais de software se tornou controverso.

A relação entre o sistema de patentes e o software nem sempre foi clara. O software é uma construção única, que incorpora aspectos criativos e funcionais. Os primeiros, na sua expressão, são protegidos pela lei de direitos autorais, enquanto os aspectos funcionais e as ideias subjacentes às invenções que origina podem ser protegidos por patentes.

A primeira patente de software concedida pelo USPTO data de 1968 (Patente N. 3.380.029) para um pedido depositado em 1965 (Quinn, 2014). No entanto, o processo de patenteamento de software não foi linear e ocorreram avanços e retrocessos ao longo da história. Originalmente, o software comercial era criado, em grande parte, pelas grandes firmas integradas, inserido como software embarcado e comercializado associado aos equipamentos físicos. Isso mudou quando software e hardware foram separados, particularmente desde que a IBM, à época a principal empresa integrada de tecnologia, foi confrontada por uma longa ação antitruste iniciada em 1969 e que se arrastou até a década de 1980, resultando na separação desses dois elementos e dando início à progressiva modularização que caracteriza a indústria de TICs, demarcando o nascimento formal da indústria de software (Rosenberg 2010:44).

Por conseguinte, programas de computador, ao gerarem soluções técnicas que não são cobertas por direitos autorais, passaram a ter sua funcionalidade protegida por patentes, no contexto das chamadas “invenções implementadas por computador” (CII, da sigla em inglês *computer-implemented inventions*). Essas invenções são definidas como aquelas “*involving the use of a computer, computer network or other programmable apparatus, or one or more features are realized wholly or partially by means of a computer program*” (Baruffaldi et al., 2020:37).

Embora a patenteabilidade de tecnologias baseadas em software tenha recebido tratamento diverso em diferentes países, atualmente há uma certa convergência em aceitar patentes de software desvinculado de hardware. Enquanto os EUA permitem que o software seja patenteado como tal, se torna patenteável na União Europeia por meio do conceito de CII (Baruffaldi et al., 2020).

O reconhecimento de patentes não diz respeito à expressão do software, mas às invenções que gera devido ao seu efeito técnico. Historicamente, os países, incluindo os EUA,

relutaram em patentear invenções relacionadas ao software desde que o Acordo TRIPS²⁰⁷ o levou para dentro do sistema internacional de direitos autorais, ainda que sem esclarecer adequadamente o escopo da proteção, deixado às leis domésticas.

No entanto, há trinta anos Reichman (1994:768) argumentava que embora muitos países relutassem em patentear invenções relacionadas a software, na prática a maioria dos ENPIs concedia patentes para aplicações de programas de computador que produzissem efeitos técnicos específicos. Com TRIPS e a consolidação do direito autoral como forma de proteção, a função econômica do software, fornecendo soluções técnicas, foi negligenciada, ignorando a realidade de inúmeras patentes concedidas ao redor do mundo.

Esses dois marcos legais para proteger invenções e criações do intelecto humano, concedem a autores e inventores direitos exclusivos sobre criações ou invenções por determinado período, com direitos vitalícios para o autor e direitos patrimoniais que se estendem aos herdeiros por prazo não inferior a 50 anos no direito autoral, e de até 20 anos a partir do depósito nas patentes de invenção. TRIPS menciona a proteção do software por meio de direitos autorais (Artigos 9 a 14, em especial Artigo 10), não havendo menção a patentes de software. O direito autoral esteve associado à natureza distinta do software, que executa funções técnicas por meio de algoritmos. Contudo, os algoritmos em si, centrais no software, enquanto conceito abstrato, não têm qualquer proteção, tal como descobertas científicas, atos mentais e ideias abstratas (Artigo 9 (2)), excluídos do patenteamento e do direito autoral em todas as jurisdições.

Com efeito, patentes não são concedidas ao software em si, protegido alternativamnte pelo direito autoral, mas às soluções/invenções técnicas que fornece, parecendo que diferentes direitos se sobreponham sobre um mesmo objeto físico, como apontado por Medeiros (2017) e Medeiros e Wachowicz (2019). De fato, o software incorpora soluções técnicas exigidas pelas leis de patentes, sem incluir algoritmos (Schirru, 2020:131)²⁰⁸.

A proteção de software difere do padrão de proteção usual das tecnologias ao abranger um mecanismo duplo, cuja evolução atípica é atrelada à modularização da indústria de TICs. O direito autoral protege a forma como o software é criado ou escrito (código-fonte ou código-objeto), não sua funcionalidade, que é onde reside seu valor econômico (Tigre e Marques, 2009:548), mesmo que a forma como o código é escrito seja ditada pela função que se destina a desempenhar (Schirru, 2020:130). As patentes, por seu turno, protegem a funcionalidade do

²⁰⁷ Disponível em: https://www.wto.org/english/docs_e/legal_e/trips_e.htm#art1. Acesso em: 20 jan. 2021

²⁰⁸ Schirru (2020:135) alerta para os efeitos negativos que a proteção exclusiva a um algoritmo pode ter no caso da IA, reforçando sua opacidade ou hermetismo.

software (ou seja, as invenções por ele implementadas, que no passado era contemplada no software “embarcado”, com patentes que cobriam todo o sistema integrado de hardware e software), não sua expressão.

Além da expressão e das invenções do software, há também proteção dos bancos de dados subjacentes. Dados (brutos e individualizados) não se qualificam em si a qualquer direito de propriedade, mas seu conjunto organizado em bancos de dados empregados pelo software convencional (e pela IA), são protegidos por direito autoral e, em algumas jurisdições, também por direitos *sui generis*²⁰⁹.

Por um lado, nos termos do Acordo TRIPS (Artigo 10 (2)):

“Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected as such. Such protection, which shall not extend to the data or material itself, shall be without prejudice to any copyright subsisting in the data or material itself.”

Com isso, as compilações de dados ou outros materiais do software, legíveis por máquina ou não, que em razão da seleção ou disposição de seu conteúdo constituem criações intelectuais (ou parte do trabalho criativo do programador), devem ser protegidos como tal por direito autoral²¹⁰.

Por outro, o investimento na coleta, organização, armazenamento e tratamento dos dados passou a admitir proteção por direitos *sui generis*. A Diretiva da Comissão Europeia de bancos de dados (Directive 96/9/CE)²¹¹ conferiu direitos jurídicos específicos, estendendo a proteção às situações não previstas no direito autoral (ou seja, quando não resultem de trabalho “criativo”, nem atendam ao requisito de originalidade), mas envolvam investimento “substancial” na base de dados, em sentido “qualitativo” e “quantitativo” (Artigo 7)²¹².

²⁰⁹ Além da União Europeia, tratada no análise do artigo, investimentos em bancos de dados também são protegidos por direitos *sui generis* em Cingapura (Singapore Copyright Act 22/2021, Artigo 244) e Japão (Copyright Law of Japan, Artigos 2, 12 e 15). O Singapore Copyright Act 22/2021 está disponível em: <https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/>. Acesso em: 09 jan. 2023. A Copyright Law of Japan está disponível em: <https://www.cric.or.jp/english/clj/cl2.html>. Acesso em: 09 jan. 2023. Gotzen (2007:1) atribui a origem desses direitos a uma Conferência Diplomática da OMPI, em 1996, que discutiu a adoção de um regime internacional sobre banco de dados para permitir recuperar o investimento e garantir uma proteção mais eficaz e uniforme. Contudo, como lembra o autor *“Fears had been expressed, especially in the United States of America, that this new type of right might lead to a monopolization of the data contained in the database”*.

²¹⁰ Revisão da Convenção de Berna (Artigo 2 (5)) (Gervais, 2007).

²¹¹ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009>. Acesso em: 20 jan. 2021

²¹² O investimento qualitativo pode resultar do emprego da experiência de um profissional (programador de IA ou cientista de dados), enquanto o investimento quantitativo envolve “a implantação de recursos financeiros e/ou a ampliação de tempo, esforço e energia” (European Commission, 2020:93). Isso pode incluir gastos em sistemas de IA ou tecnologia usada para produzir o banco de dados, como será visto.

As bases de dados são definidas (Artigo 1(2) Directive 96/9/CE)²¹³ como “*a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means*”. O titular dos direitos é o “criador da base de dados” (Artigo 7 (1)), ou seja, a pessoa que toma a iniciativa e assume o risco do investimento, beneficiando nacionais, residentes dos Estados-Membros, sociedades constituídas conforme as respectivas legislações nacionais e com sede, administração central ou estabelecimento principal na UE (Artigo 11). Por fim, requer que investimento substancial seja feito “*in either the obtaining, verification or presentation of the contents*” do banco de dados (Artigo 7 (1)). Não é exigido registo e a proteção é por 15 anos, mas admite novos direitos cada vez que a base de dados é substancialmente modificada²¹⁴.

Alternativamente, bancos de dados podem ser mantidos em sigilo e, em alguns países, proteção legal equivalente a DPIs por *trade secrets*, que confere aos titulares direitos contra a aquisição, uso e divulgação do objeto obtido por meios ilegais²¹⁵, desde que mantidos secretos, tenham valor comercial e sujeitos a medidas razoáveis para mantê-los confidenciais, que são as exigências da lei.

Em síntese, a proteção de banco de dados na União Europeia inclui direitos autorais (proteção da expressão original da ideia contra cópia), direitos *sui generis* (proteção do seu conteúdo, quando investimento substancial tenha sido feito na obtenção, verificação ou apresentação desse conteúdo) e proteção por *trade secrets* (proteção contra aquisição, uso e divulgação do objeto por meios ilegais) (Drexler et al., 2016).

5.2. Propriedade intelectual dos insumos e produtos da IA

A despeito das tecnologias da IA compartilharem os mesmos instrumentos de proteção, têm uma dimensão peculiar em relação ao software convencional, mesmo que nele se expresse, englobando seus insumos – técnicas de aprendizado, algoritmos de autoaprendizagem e BigData – e os produtos da IA – invenções (ou criações artísticas/literárias) – alguns dos quais dão origem a DPI específicos, conforme resumidos na Figura 5.1.

²¹³ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009>. Acesso em: 20 jan. 2021

²¹⁴ Nardin (2021) lembra que o licenciamento compulsório para bancos de dados de fonte única que foi originalmente proposto pela Comissão Europeia acabou excluído na versão aprovada da Diretiva europeia.

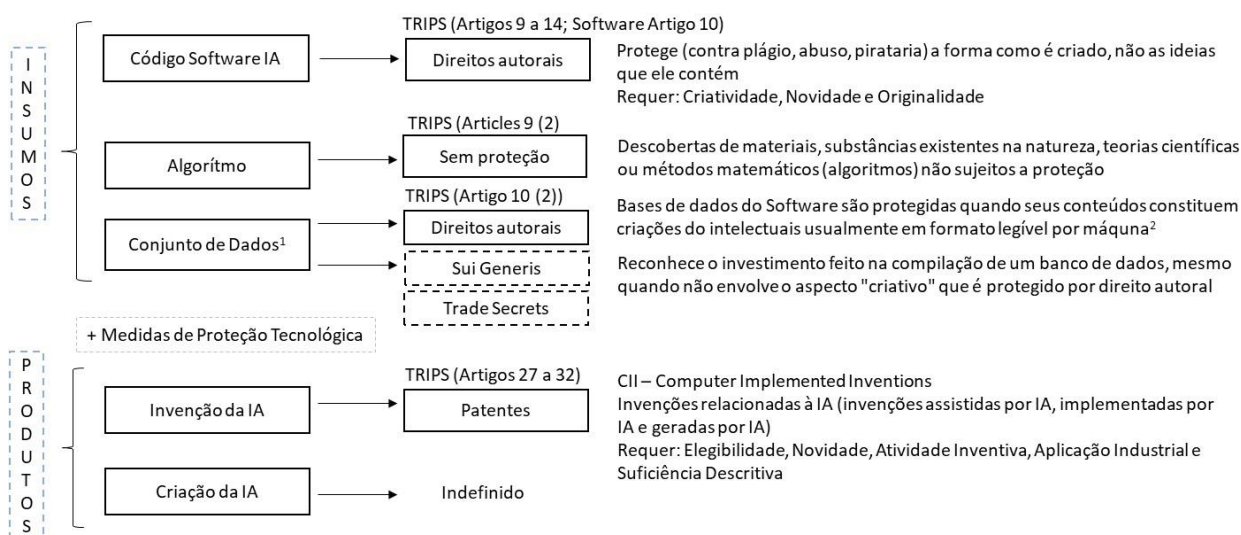
²¹⁵ A legislação europeia define as circunstâncias em que a aquisição, uso e divulgação do *trade secret* é considerada lícita, com a proteção jurídica só podendo ser invocada quando informações foram obtidas por meios legítimos nos termos da lei.

5.2.1. Insumos da IA

5.2.1.1. Modelos de IA

O código-fonte dos modelos de IA - que pode ser entendidos como um insumo para invenções relacionadas à IA – é passível de proteção por direito autoral em termos de sua expressão (a forma como foi escrito pelo programador) como o software convencional. Não obstante, foi visto que durante o processo de treinamento dos modelos de redes neurais profundas, com grandes conjuntos de dados, de diversos domínios, e o emprego de modernas técnicas de aprendizado autosupervisionado, os modelos, o próprio código e os algoritmos específicos de IA (“autodidatas”) se modificam (evoluem) em relação à programação original, o que não ocorre no software convencional.

Figura 5.1 – Regime Jurídico dos Direitos de Propriedade Intelectual da IA



Fonte: Elaborado pela autora a partir dos acordos internacionais e alguma legislação vigente. Os retângulos tracejados representam formas de proteção que existem apenas em algumas jurisdições, mas não foram estabelecidas em TRIPS, por isso diferenciadas na figura. Correspondem aos direitos sui generis sobre conjuntos de dados da União Europeia, Japão e Cingapura; além da possibilidade de mantidos em sigilo, mas que, do ponto de vista de direitos legais de PI, devem seguir as condições estabelecidas na legislação de *trade secrets* nos Estados Unidos e União Europeia.

Notas: ¹ Proteção da PI de bases de dados contém diferenças importantes entre as legislações nacionais.

² Direitos exclusivos sobre cópia, distribuição e tradução de bases de dados

O ritmo acelerado do progresso técnico da IA e o processo de aprendizado contínuo com os dados que parece somente interrompido quando alcança um alto nível de precisão e desempenho em determinada aplicação funcional (por exemplo, visão computacional, processamento de linguagem natural etc.), parecem desabonar a importância do direito autoral como instrumento de proteção.

Sistemas e modelos de IA concebidos para funcionar “*com vários níveis de autonomia e que pode apresentar capacidade de adaptabilidade* após a implantação...” (conforme a própria definição da regulação europeia de usos da IA), esbarram inexoravelmente no requisito

de originalidade do autor humano. Fenômenos como “emergência” dos modelos parecem consolidar o entendimento de que se adaptam aos crescentes e variados volumes dados, desenvolvendo habilidades imprevistas. Por um lado, isso parece tornar inviável a proteção; por outro, parece tornar duvidosa a utilidade (e o valor) da proteção legal desses insumos, que são sistemas dinâmicos que passam por modificações internas, além de rapidamente se tornarem obsoletos com a valocidade em que avançam as técnicas, os conjuntos de dados e a capacidade computacional. Por conseguinte, reforça a noção de que, mais que o software convencional, o real valor dos sistemas de IA parece determinado pela sua funcionalidade técnica, não pela sua expressão, explicando, em parte, a tendência crescente do *open-source da IA*.

O programador original pode obter a autoria legal do modelo de IA que escreve, mas os resultados alcançados pelo modelo não foram integralmente programados (no sentido de um modelo teórico cientificamente demonstrável) e sua funcionalidade depende das mudanças nos parâmetros e algoritmos, que o autor pode não explicar completamente. Por fim, os modelos da IA dependem cada vez mais de muitos outros detalhes técnicos, muito além do seu código, tal como arquitetura, ajuste fino, parâmetros, dados de treinamento, tipos de hardware e capacidade computacional empregada, que progressivamente se revestem de grande sigilo, comprometendo até mesmo a real abertura dos modelos disponibilizados em regime *open-source*.

5.2.1.2. Bases de dados

As invenções em redes neurais profundas derivam de modelos matemáticos e estatísticos, cujas capacidades e desempenho são em boa medida resultado de treinamento extensivo com amplos e variados conjuntos dos dados e de análise de BigData, que revelam padrões, identificam correlações, fazem previsões e criam conteúdo. A BigData é composta cada vez mais por dados não estruturados, cuja rotulagem e análise é crescentemetne realizada pelos próprios sistemas/algoritmos de aprendizado, tornando sua proteção mais complexa que o software.

Os tratados internacionais permitem proteger bases de dados pelo direito autoral quando constituam criações intelectuais originais e uma expressão do autor. No caso da BigData, os grandes conjuntos de dados de treinamento dos modelos de aprendizado parecem longe de constituir resultado “exclusivo” do trabalho do programador e curadoria humana. Por um lado, resultam de um fluxo de dados quase contínuo em tempo real, coletados e analisados por BigData e sistemas de IA com vistas a treinar e melhorar a precisão dos modelos, que levanta

dúvidas sobre qual “base de dados” estaria protegida²¹⁶. Por outro, ainda que a coleta e rotulagem inicial dos dados envolvam um trabalho de programação humana, há um incremento quase “autônomo” para alcançar o desempenho, a eficiência e as capacidades dos sistemas.

Essas observações são convergentes com Gervais (2019:7), que lembra que, embora não mencionado expressamente no texto da Convenção de Berna, um comitê de peritos definiu que o único requisito obrigatório para que uma obra seja protegida é ser “original”²¹⁷. Por conseguinte, a BigData, com bases de dados cada vez mais geradas automaticamente por sistemas de IA, teriam dificuldades para se qualificar aos direitos autorais. É improvável que a Big Data “selecione ou organize” os dados de uma maneira (pelo tipo de ferramenta que usa) capaz de atender ao critério de originalidade autoral humana como os bancos de dados do software (Gervais, 2019; 2020)²¹⁸. Na ausência de tal originalidade, nenhum direito autoral se aplicaria às bases de dados da IA (Graux, 2021).

Com relação a direitos *sui generis*, relatório da European Commission (2020:93) concluiu que como esse direito não requer autoria humana, é possível a proteção de bancos de dados desde que derivada de um investimento substancial. Assim, a Diretiva de Bancos de Dados²¹⁹ pareceria, à primeira vista, melhor acomodar o BigData.

No entanto, Gervais (2019:9-10) vê problemas na proteção desses conjuntos de dados gerados por máquina mesmo por direito *sui generis*, pelo fato dos resultados na análise de BigData serem por ela criados, comprometendo a lógica de recompensar o investimento. Este autor lembra a dificuldade de ampliar o escopo das reivindicações, os desafios relacionados ao inventor e a justificativa para conceder direitos a uma invenção feita por uma máquina, questionando se poderia ser descrita de forma suficiente para a concessão de uma patente e para a definição do estado da arte. Ao definir BigData como a “transformação de dados brutos em soluções técnicas”, é natural a conclusão do autor de que se qualificaria à proteção por patentes,

²¹⁶ Existem até propostas de criar e proteger repositórios de algoritmos da IA.

²¹⁷ Vale lembrar que os requisitos obrigatórios de proteção pelo direito autoral, quais sejam criatividade, novidade e originalidade, apenas este último pareça mais difíceis de serem atendidos pelos conjuntos de dados da IA.

²¹⁸ Gervais (2019:8) justifica a crescente adesão das firmas, inclusive as BigTechs, ao OSS para BigData pelo fato das ferramentas que utiliza não se qualificarem para proteção de direitos autorais. Ao contrário, o peso crescente do OSS, talvez decorra do fato do direito autoral poder ser considerado um regime fraco e incidir sobre bases de dados, em constante mutação, que são cruciais para a invenção da IA, o que explica a preferência por patentes.

²¹⁹ No caso de bancos de dados, há um amplo e complexo emaranhado de normas em constante evolução que vem desenhando a estratégia de dados da União Europeia, compreendidas em várias Diretivas, que demandam uma análise mais aprofundada do que é possível neste trabalho, inclusive como inspiração para propostas legislativas, que fica como tema sugerido para novas pesquisas. São elas a Diretiva 2009/24/EC, que trata especificamente da proteção legal dos programas de computador, a Diretiva 2001/29/EC (Artigos 2 e 3), conhecida como *Information Society Directive*, alterada pela Diretiva 2017/1564/EC, relacionada à harmonização de aspectos de direitos autorais e direitos conexos da sociedade da informação, alterada pela Diretiva 2019/790/EC, que trata direitos autorais e direitos conexos no “*Digital Single Mercado*”.

ainda que a BigData, sozinha, não gere invenções, mas apenas contribua para o treinamento dos modelos de IA em redes neurais profundas²²⁰.

Por fim, cabe lembrar que os DPI sobre bases de dados começam a ser afetados pelo amplo contexto normativo recente da União Europeia sobre regulação dos mercados e serviços de plataformas digitais, privacidade e proteção de dados pessoais, governança e compartilhamento de dados e o uso da IA. Em particular, as novas abordagens sobre compartilhamento e portabilidade dos dados afetarão diretamente a PI sobre bases de dados, estando em curso a revisão da Diretiva de Bancos de Dados, em especial desde a Directive EU 2019/1024²²¹ (ODD), chamada *Open data and the re-use of public sector information*²²².

5.2.1.3. A proteção dos chips de computador

A partir da escalada recente dos modelos da IA generativa e da crescente importância dos ativos tangíveis infraestruturais, cabem comentários sobre os chips para IA, que traduzem uma dimensão importante do ecossistema de IA. O padrão de alta fragmentação global da indústria de semicondutores, com empresas especializadas em tarefas específicas, dispersas pelo mundo, claramente confere um incentivo a mais para a PI, enfatizada pela literatura para proteção contra pirataria e *trojans* de hardware (Knechtel, Patnaik e Sinanoglu, 2019), sobretudo no *design* de chips. Direitos de PI sobre *design* e arquitetura complexos de GPUs e TPUs constituem uma importante barreira à entrada, que se somam às barreiras de capital para sustentar o expressivo investimento em P&D exigido, com custos crescentes para instalação de

²²⁰ Essa leitura um pouco imprecisa da BigData justificaria patentear tanto invenções, quanto bases de dados da IA, o que parece estar ocorrendo na prática. No entanto, como o direito autoral é considerado um regime fraco, o tema da proteção de bases de dados por patentes, segredos comerciais ou meios contratuais vem sendo discutido (Debussche e César, 2019; Koutroumpis et al., 2020), mesmo considerando que criam mais direitos sobre o que para muitos talvez devesse estar em domínio público. Isso não significa que muitos pedidos de patentes analisados na pesquisa desta tese, principalmente nos EUA, não contemplem exclusivamente bases de dados.

²²¹ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1024&from=EN>. Acesso em: 20 jan. 2022

²²² Uma das inovações da *Open Data Directive* (ODD) é um acesso dinâmico aos dados pelo uso obrigatório de interfaces de programas de aplicativos (APIs) para os chamados “conjuntos de dados de alto valor”, que abrangem aqueles que não se qualificam para DPIs por conterem dados exclusivamente objetivos e factuais (que desqualificam para direitos autorais) e gerados como subproduto de uma principal atividade e não um fim em si mesmos (o que desqualificaria para direitos *sui generis*) (Graux 2021:7). A Diretiva (UE) 2019/1024, em vigor desde julho de 2021, sobre dados abertos e reutilização de informações do setor público, reformulou a “*Public Sector Information (PSI) Directive*” (Diretiva (UE) 2003/98, alterada pela Diretiva (UE) 2013/37), e terá repercussão sobre DPIs sobre bases de dados pois seu alcance vai além de dados diretos do setor público, atingindo também bases de dados de firmas privadas construídas com apoio financeiro público. A ODD se junta a General Data Protection Regulation (GDPR), o Data Governance Act (DGA), a regulação da concorrência nos mercados e serviços digitais pelo Digital Services Act (DSA) e Digital Market Act (DMA) e à regulação dos usos da IA (Artificial Intelligence Act), além da recente proposta de EU Directive 2022/0303, de 28/09/2022, sobre responsabilidade civil por eventuais danos da IA. Em seu conjunto, essas leis podem afetar DPIs sobre dados na União Europeia e, pela importância do mercado europeu, replicar pelo mundo e abrir espaço para que o amplo aparato regulatório tenha impacto global (*Brussels effect*). Como lembram Graef e Prüfer (2021), a proteção da PI não é absoluta e deve ser equilibrada com esses outros direitos.

produção de chips (quase triplicaram entre 2017 e 2023, quando alcançaram US\$ 20 bilhões) (Azhar et al., 2024).

Rubinger (2020) considera existir uma corrida paralela de patentes de chips de IA e computação na nuvem – que talvez represente uma migração de patentes sobre invenções de IA para patentes de chips – refletindo não somente a tradicional alta taxa de patenteamento em semicondutores, como a importância estratégica de proteção de chips para IA em nuvem. Este é o caso de empresas como Intel, Nvidia, IBM, Qualcomm, Graphcore, Intellifusion, Huawei, além de Baidu e Google – as TPUs implantadas nos data centers do Google eram protegidas por 34 famílias de patentes até 2019. Isso poderia evidenciar uma mudança de foco da inovação, que se voltaria para hardware especializado de IA para data centers, nuvem, smartphones etc. Casos de litígio começam a surgir nesse cenário, como a ação movida pelo fundador da Singular Computing contra o Google por violação de suas patentes de chips de IA compartilhadas com o Google entre 2010 e 2014 e que alega terem sido copiadas para o desenvolvimento das versões v2 e v3 TPUs (Burt, 2024).

5.2.2. Produtos da IA: invenções relacionadas à IA

Os sistemas de IA cada vez mais dão origem a invenções - da mesma forma que criações artísticas/literárias - aqui tratados como produtos da IA, resultado de sua aplicação técnica e de sua funcionalidade. Essas invenções são protegidas, da mesma forma que o software convencional, por patentes de invenções implementadas por computador (CII) – as criações artísticas/literárias geradas por IA²²³, contudo, ainda parecem ter DPI indefinidos.

As invenções relacionadas à IA sempre foram protegidas por patentes. Inicialmente, embarcada em conjuntos abrangendo IA, software e hardware; mais tarde, com a desagregação de software e dispositivos de hardware e a modularização da indústria, começaram a ser protegidas separadamente por patentes de invenção. A Patente US N. 5.852.815, de 15 de maio de 1998, se tornou a primeira patente conhecida a ser emitida para uma invenção gerada por IA (WEF, 2018).

A categoria de CII, criada pela União Europeia, foi rapidamente incorporada por vários outros países, para identificar as patentes de software e, mais recentemente, também invenções relacionadas à IA. A Figura 5.2 é uma reprodução da figura apresentada em Frietsch (2015:9), à qual adicionei as “invenções relacionadas à IA”, correspondendo ao retângulo de cor laranja

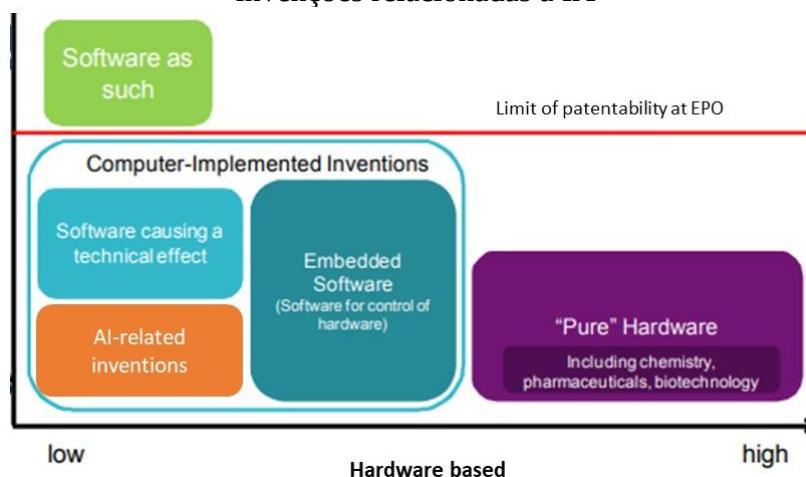
²²³ Como o musical *Beyond the Fence*, composições do DeepMind/Google e AIVA (*Artificial Intelligence Virtual Artist*), além da conclusão da 8ª Sinfonia de Schubert pelo system de IA da Huawei e o projeto Next Rembrandt que cria obras inéditas de pintor famoso. Muitas dessas criações usam a mesma técnica de *Generative Adversarial Networks (GAN)* usada em invenções e soluções técnicas (Mazzone e Elgammal, 2019).

inserido entre os demais tipos de CII.

Drexler et al. (2020) identificaram uma tipologia de invenções relacionadas à IA, que parece relacionada ao grau de autonomia dos sistemas de IA. São elas as invenções assistidas por IA, quando a IA é usada como ferramenta para inventar; as invenções implementadas por IA, quando a IA faz parte da invenção; e as invenções geradas por IA, quando a IA age de forma autônoma sem intervenção humana.

Enquanto a primeira categoria poderia ser protegida como o software tradicional, as duas últimas são efetivamente fenômenos novos e envolvem dimensões mais complexas de DPI. Isso porque, ao contrário do papel passivo do software, alguns sistemas de redes neurais profundas podem desempenhar um papel parcial (em invenções implementadas por IA) ou totalmente ativo (em invenções geradas por IA). A diferença das invenções relacionadas a IA frente ao software está no seu grau de autonomia, complexidade e opacidade (efeito “caixa preta”). As questões mais desafiadoras ao regime de DPI surgem quando, segundo Gervais (2020), o “limiar da autonomia” (*‘autonomy threshold’*) é ultrapassado.

Figura 5.2 - Representação gráfica das *Computer Implemented Inventions*, incluindo invenções relacionadas a IA



Fonte: Adaptado de Frietsch (2015:9)

A noção de autonomia, que alguns resistiam em admitir e pareceu definitivamente aceita desde o surgimento da IA generativa e a definição do AI Act, bem como a “não explicabilidade” dos modelos é que parecem desafiar o atual regime de patentes, conforme abordado por um conjunto relevante da literatura (WEF, 2018; Abbott, 2016a; Abbott, 2016b; Verstinsky, 2017; Ravid e Liu, 2018; Hilty et al., 2020; entre outros). Há vozes dissonantes que resistem à autonomia dos sistemas de IA (Kim 2020; Kim et al. 2021)²²⁴ das atuais tecnologias de IA

²²⁴ Uma resenha extensa da literatura é apresentada em Kim (2020), aparentemente destinada a negar a alegada autonomia do sistema DABUS. A autora admite que os modelos de aprendizado em redes neurais envolvem transformações, rotulagem, verificação, modelagem de dados por meio de funções matemáticas e estatísticas sofisticadas, mas não considera haver qualquer autonomia.

(ANI), preferindo empregar a expressão “imprevisibilidade de soluções implementadas por computador”. Crawford (2021), outra autora que busca desmistificar a ideia de inteligência (e com isso, autonomia) dos modelos de IA, ressalta o trabalho humano na rotulagem dos dados e no desempenho dos modelos. Nem toda a IA é uma solução técnica para um problema técnico identificado; muitas vezes ela cria respostas para problemas sequer formulados.

O emprego do termo “autonomia” não significa que sistemas de IA prescindem de programação; apenas que o funcionamento interno dos modelos resulta de soluções computacionais não inteiramente previstas e explicadas. Vale lembrar a argumentação do artigo de Samuelson (1985:1208) sobre direitos autorais em situações em que o software gera automaticamente uma saída que não é idêntica ao seu próprio texto, que pode ser válida para as atuais das patentes de invenção da IA, o que corresponderia a conceder direitos que “*over-reward . . . the programmer, particularly in light of the fact that the programmer is no more able to anticipate the output than anyone else*”.

O tratamento de TRIPS é genérico e têm levado leis nacionais, principalmente por meio de reedições das diretrizes de exame dos principais ENPIs, a buscar suas próprias respostas, ainda que muitas vezes convergentes, às lacunas e indefinições sobre invenções da IA. Os desafios dessas invenções ao sistema de patentes dizem respeito à elegibilidade e aos mencionados requisitos para patenteamento, inclusive sobre a figura do inventor. Por outro lado, os sistemas de IA baseados em modelos matemáticos e estatísticos que, junto aos algoritmos, são matérias de natureza abstrata, reproduzem as incertezas do debate do software convencional. A avaliação desses pontos polêmicos exigiu, portanto, um olhar mais atento sobre as legislações e diretrizes de exame dos principais países.

5.3. Patentes de Invenção relacionadas à IA na prática dos ENPIs

Os principais aspectos relacionados às invenções da IA estão elencados nas leis nacionais e, sobretudo, nas diretrizes de exame dos ENPIs. De fato, as condições e requisitos substantivos exigidos para o depósito e concessão de patentes de IA estão contidos nesses documentos. O website da OMPI disponibiliza os links dessas diretrizes²²⁵, incompleto e desatualizado, de forma que o melhor caminho para a pesquisa pareceu indicar a consulta diretamente nos websites do USPTO, EPO e, ainda que mais superficialmente, do CNIPA.

Na União Europeia os pedidos de patentes são depositados no European Patent Office (EPO)²²⁶ e examinados com base na European Patent Convention (EPC)²²⁷, embora após a

²²⁵ Disponível em: <https://wipo.int/patents/en/guidelines.html>. Acesso em: 12 out. 2024

²²⁶ Disponível em: <https://www.epo.org/>. Acesso em: 20 jan. 2021

²²⁷ Disponível em: <https://www.epo.org/law-practice/legal-texts/epc.html>. Acesso em: 20 jan. 2021

concessão da patente a competência seja transferida para os estados designados na patente e tribunais nacionais, que determinam sua validade e alcance. As leis de patentes dos Estados-Membros variam consideravelmente, alguns contemplando sistema de exame e outros apenas sistema de registo.

Uma das primeiras tentativas de definição formal da IA para fins de PI veio com o Grupo de Peritos de Alto Nível em Inteligência Artificial, criado pela Comissão Europeia (AI HLEG 2018:1)²²⁸. Tal definição foi incluída na Comunicação da Comissão Europeia sobre IA (EU COM (2018) 237)²²⁹:

“Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be *purely software-based*, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech, and face recognition systems) or AI can be *embedded in hardware* devices (e.g., advanced robots, autonomous cars, drones, or Internet of Things applications).”
(grifos da autora)

No website do EPO, IA é referida como a capacidade de programas de computador e máquinas realizarem tarefas mentais comumente associadas a humanos, como aprendizado, raciocínio e resolução de problemas. Na prática, contudo, as invenções relacionadas à IA são tratadas pelo EPO por meio do conceito de CII²³⁰, que abrange invenções envolvendo computadores, redes de computadores ou outros dispositivos programáveis, onde pelo menos um recurso é realizado por meio de um programa, nos termos das Diretrizes de Exame de Patentes (EPO Guidelines for Examination, F-IV, 3.9)²³¹.

Nessa última versão dessas diretrizes, em vigor a partir de 1º de março de 2024, os sistemas de IA são:

“[...] software systems (and possibly also hardware) [...] that act in the physical or digital dimension perceiving their environment through data acquisition, interpreting the collected data, [...] reasoning about the knowledge or processing the information, derived from that data, and deciding the best action(s) to be taken to achieve a certain objective. AI systems can also [...] adapt their behavior by analyzing how the environment is affected by their previous actions”.

O software é excluído das invenções patenteáveis pela EPC por não serem considerados invenções (Artigo 52 (2) (c) EPC), exceto em invenções que envolvam software e sejam de natureza técnica (CII) e atendam aos requisitos de patenteabilidade: novidade, atividade inventiva e aplicação industrial (Artigo 52 (1) EPC).

²²⁸ Disponível em: https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf. Acesso em: 20 jan. 2021

²²⁹ Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN>. Acesso em: 20 jan. 2021

²³⁰ Disponível em: <https://www.epo.org/news-events/in-focus/ict/hardware-and-software.html>. Acesso em: 10 jul. 2022

²³¹ Disponível em: <https://www.epo.org/law-practice/legal-texts/guidelines.html>. Acesso em: 10 jul. 2022

Nas diretrizes de exame os requisitos básicos para patenteabilidade são detalhados da seguinte forma: (a) deve haver uma invenção pertencente a qualquer campo da tecnologia (Guidelines G-II); (b) a invenção deve ser passível de aplicação industrial (Artigo 57 EPC) (G-III); (c) a invenção deve ser nova (Artigo 54 EPC) (G-IV, V e VI); e (d) a invenção deve envolver atividade inventiva (Artigo 56 EPC) (G-VII). Por fim, a invenção deve ser descrita de forma suficiente (EPC Implementation, Rule 42).

No tópico que trata das invenções implementadas por computador, há um item específico sobre **métodos matemáticos** (G-II, 3.3.1), que cita IA e aprendizado de máquina, baseados em modelos computacionais e algoritmos de classificação, agrupamento, regressão e redução de dimensionalidade, como redes neurais, algoritmos genéticos, máquinas de vetores de suporte, k-means, regressão de kernel e análise discriminante. Tais modelos computacionais, de natureza matemática abstrata *per se*, independentemente de poderem ser "treinados" com dados, com vários exemplos listados, podem ter patentes concedidas quando a IA sai do reino abstrato e é aplicada para resolver um problema técnico em um campo da tecnologia. Métodos matemáticos e programas de computador podem, assim, ser protegidos por patentes quando usados como parte de um sistema de IA que contribua para a produção de um efeito técnico adicional. A "abstração" da IA é desconsiderada se a invenção acrescentar um efeito técnico. O caráter técnico da invenção passa a ser, assim, a qualidade central para a elegibilidade.

Em relação à divulgação da invenção (Artigo 83 EPC), as orientações das Diretrizes são genéricas, com poucos exemplos, comprometendo a clareza. Novamente é feita referência à funcionalidade da invenção, exigindo "*...that the invention is described not only in terms of its structure but also in terms of its function...*". Observa que em campos técnicos como de computadores, "*...a clear description of function may be much more appropriate than an over-detailed description of structure*" [*of the innovation*] (F-III, 1). Essa descrição não pode ser feita apenas em linguagem de programação, mas escrita em linguagem natural, ainda que aceitos pequenos trechos em linguagem de programação (FII, 4.12).

É na lista das exclusões que é detalhado o que deve ser considerado efeito técnico para fins de patenteabilidade. Para ter um caráter técnico, um programa de computador deve produzir um "efeito técnico adicional" quando executado em um computador (G II, 3.6). Nesse caso, a eficiência computacional de um algoritmo que afeta o efeito técnico estabelecido contribui para o caráter técnico da invenção e, portanto, para a atividade inventiva. Um programa de computador não pode derivar um caráter técnico do simples fato de ter sido projetado de modo a ser executado automaticamente por um computador, exigindo considerações técnicas adicionais além de um algoritmo de computador para realizar uma tarefa. Em síntese, mesmo

nas diretrizes o tema ainda é hermético e não deixa claro o que diferencia de uma matéria (puramente) abstrata que, pela legislação, não seria patenteável.

Embora o EPO mencione em alguns documentos as três categorias de invenções relacionadas à IA de Drexel et al. (2020)²³², não há qualquer referência a elas na EPC, nas regras de implementação e nas diretrizes de exame, levando a supor que o exame não contempla qualquer avaliação sobre o grau de autonomia da invenção da IA, ou seja, se a invenção foi implementada, gerada ou apenas assistida pelo sistema de IA. Na prática, a IA parece uma simples ferramenta para o inventor humano (Botha, 2019). Este ponto guarda relação com o requisito de etapa inventiva (inventividade ou não-obviedade) e, por conseguinte, quem é o inventor, que será abordada adiante.

No caso da BigData, a omissão é quase completa nas diretrizes de exame (como nos demais ENPIs comentados), uma vez não protegidos por patentes. Mesmo assim pedidos de patentes têm sido depositados abrangendo exclusivamente bases de dados.

Nos Estados Unidos, por seu turno, qualquer processo, máquina, fabricação ou composição de matéria são elegíveis para patentes, exceto ideias abstratas, leis da natureza e fenômenos naturais, como algoritmos (35 US Code §101)²³³. Mesmo assim, no país precursor do patenteamento do software, ocorreram muitas idas e vindas nas decisões do USPTO e tribunais relacionadas ao software e, depois à IA, com debates sobre patenteamento de modelos de negócios, fórmulas matemáticas, algoritmos e, mais recentemente, ideias abstratas, até a decisão da Suprema Corte sobre o caso *Alice Corp v. CLS Bank*, em 2014²³⁴, que definiu um teste em duas etapas que passou a ser exigido para patentes de software e IA.

²³² O *Max Planck Institute for Innovation and Competition* vem contribuindo para o debate sobre DPIs em invenções da IA, inclusive com uma categoria com autonomia na taxonomia proposta para as invenções geradas pela IA (Drexel et al. 2019; Drexel et al. 2020; e Hilty et al. 2020), parece reorientar sua abordagem desconsiderando a autonomia dos sistemas da IA (ver Kim 2020; Kim et al. 2021), inclusive com referências às tentativas de patenteamento das invenções do sistema DABUS.

²³³ Disponível em: <https://www.uspto.gov/web/offices/pac/mpep/mpep-9015-appx-l.html#d0e302376>. Acesso em: 10 jul. 2022

²³⁴ Ideias abstratas, leis da natureza e fenômenos naturais são excluídos de proteção intelectual (cujas racionalidade é que a proteção de verdades científicas fundamentais significaria retirar da sociedade as ferramentas necessárias para inovar). Foram muitas disputas judiciais envolvendo patentes de software até a decisão do caso *Alice Corp. v. CLS Bank International* que acabou se tornando um paradigma para avaliar pedidos de patentes de invenção relacionadas ao software e a inteligência artificial. A batalha envolveu a empresa australiana *Alice Corp.*, que obteve patentes de invenção de soluções de software usadas em negociações de valores mobiliários, e o *CLS Bank*, que desenvolveu softwares para uso cativo e contestou a legalidade das patentes da rival, alegando se tratar de ideias abstratas, enquanto a *Alice* processava o *CLS* por violação de suas patentes. Foram muitas idas e vindas nas várias instâncias do tribunal distrital, tribunal de apelações do Circuito Federal até chegar à Suprema Corte dos Estados Unidos, que definiu o teste em duas etapas como a regra a ser seguida, que, para muitos criou grandes dificuldades para essas patentes. Calvin e Leung (2020) mencionam que os grandes detentores de patentes, como Google, Amazon e Meta, assinaram um *amicus curiae* - expressão latina utilizada para designar um terceiro que ingressa em um processo para função fornecer subsídios às decisões dos tribunais, de modo a oferecer melhor base para questões relevantes e de grande impacto - para a Suprema Corte, em 2014, defendendo que o tribunal tornasse

Projetado para evitar o patenteamento de algoritmos, o teste Alice/Mayo como ficou conhecido, passou a ser exigido visando determinar: a) se a invenção é direcionada a conceitos não patenteáveis (ideias abstratas); e, em sendo, b) se os elementos reivindicados das invenções implementadas por computador fornecem um conceito inventivo no sentido de converter ideias abstratas em pedido elegível. A partir de então, os pedidos de patente nos EUA para software e IA passaram a ser examinados com base no teste, embora a incerteza jurídica persista (Quinn, 2014; Millien, 2020)²³⁵.

A nona edição das diretrizes de exame de patentes do USPTO²³⁶ (*Manual of Patent Examining Procedure - MPEP*)²³⁷, de janeiro de 2024, trata detalhadamente das questões do software e IA, com definições e muitos exemplos. A elegibilidade para patentes (Capítulo 2100, seção 2106) exige que uma invenção esteja incluída em uma das quatro categorias definidas pelo *US Code Title 35* (35 USC §101): processos, máquinas, manufaturas ou composições de matéria. Recomenda o uso do teste Alice na avaliação da elegibilidade. Se a reivindicação é direcionada a uma ideia abstrata, lei da natureza ou fenômeno natural ou se elementos adicionais na reivindicação integram a ideia abstrata com vistas a uma aplicação prática (MPEP § 2106.04(d)). Quando a reivindicação não mencionar ideias abstratas, a invenção é elegível (O MPEP § 2106.04(a)(1)).

O Manual esclarece que o “software não é automaticamente uma **ideia abstrata**, mesmo que a execução de uma tarefa de software envolva um cálculo matemático subjacente” e “um método de negócios é simplesmente um tipo de 'método' que é, pelo menos em algumas circunstâncias, elegível para patente”.

Para facilitar o exame, passaram a ser mostrados exemplos de casos de ideias abstratas que seriam elegíveis, que os examinadores antes buscavam na jurisprudência, agrupados em três categorias: a) Conceitos matemáticos – relações matemáticas, fórmulas ou equações matemáticas, cálculos matemáticos (MPEP § 2106.04(a)(2), subseção I); b) certos métodos de

mais desafiador patentear ideias abstratas em invenções mais teóricas de software, IA e aprendizado de máquina, lembrando que patentes de software abstratas prejudicaram a inovação ao conceder direitos exclusivos sobre ideias de alto nível, impedindo que outros empreendessem tarefas verdadeiramente inovadoras de desenvolver aplicações específicas.

²³⁵ As patentes de software nos EUA quintuplicaram entre 1991 e 2011 (US GAO, 2013), mas com críticas à sua qualidade. Usando a mesma metodologia, Millien (2022) estimou que essas patentes continuaram a crescer, representando 63% do total de patentes de utilidade concedidas em 2021 pelo USPTO – 49,4% das patentes do EPO e 40,2% do CNIPA.

²³⁶ Disponível em: <https://www.uspto.gov/initiatives/artificial-intelligence/artificial-intelligence-resources>. Acesso em: 25 out. 2024

²³⁷ Disponível em: <https://www.uspto.gov/web/offices/pac/mpep/index.html>. Acesso em: 25 out. 2024. Ver também “AI related resources” no website do USPTO, para o detalhamento de questões pertinentes às invenções da IA. Disponível em: <https://uspto.gov/initiatives/artificial-intelligence/artificial-intelligence-resources>. Acesso em: 25 out. 2024

organização da atividade humana – princípios ou práticas econômicas fundamentais (incluindo cobertura, seguro, mitigação de risco); interações comerciais ou legais (incluindo acordos na forma de contratos; obrigações legais; publicidade, marketing ou atividades ou comportamento de vendas; relações comerciais); gerenciamento de comportamento pessoal ou relacionamentos ou interações entre pessoas (incluindo atividades sociais, ensinar e seguir regras ou instruções (MPEP § 2106.04(a)(2), subseção II); e c) processos mentais - conceitos realizados na mente humana (incluindo observação, avaliação, julgamento, parecer) (MPEP § 2106.04(a)(2), inciso III).

O Manual observa ainda que a patenteabilidade deve também considerar os demais artigos da lei (35 USC §112, 35 USC §102 e 35 USC §103) com vistas a observar os requisitos relacionados à novidade, não obviedade e suficiência descritiva. A orientação quanto à suficiência descritiva é uma das mais extensas e exemplificada através de decisões de tribunais.

O pedido deve incluir uma descrição escrita da invenção, do processo de fabricação usando a invenção e sobre o modo considerado pelo inventor de realizar a invenção (MPEP 2161.01)²³⁸. O nível de detalhe necessário para satisfazer ao requisito de descrição escrita varia dependendo da natureza e escopo da reivindicação e a complexidade da tecnologia, que no caso de programas de computador onde a invenção muitas vezes não é descrita, apenas apresentada, em linguagem funcional, como o resultado pode ser alcançado²³⁹. O algoritmo²⁴⁰ ou as etapas/procedimentos para executar a função devem ser descritos em detalhes suficientes para que um especialista na técnica entenda como o inventor pretende que a função seja desempenhada (MPEP §§ 2163.02 e 2181). Nas invenções implementadas por computador, a suficiência da divulgação exige investigação sobre hardware e software devido à inter-relação e interdependência entre ambos, inclusive o algoritmo, que se não contemplar detalhes suficientes para demonstrar uma habilidade comum na técnica possuída pelo inventor, pode ser rejeitada.

A lei de patentes dos EUA não permite que ideias abstratas sejam patenteadas, uma vez que ideias abstratas, leis da natureza e fenômenos naturais são ferramentas básicas do trabalho científico e tecnológico, conforme preocupação manifestada pela Suprema Corte de que fossem monopolizadas pela concessão de direitos de patente e impedissem, ao invés de promover, a

²³⁸ O Manual diferencia os pedidos apresentados antes e depois do Leahy-Smith America Invents Act (AIA), (35 US Code 112(b) e pre-AIA 35 US Code 112)), que mudou o critério de prioridade do primeiro a inventar para o primeiro a depositar.

²³⁹ “Language that simply describes the function to be performed describes a result, not a means to achieve that result” (Microsoft Computer Dictionary, Microsoft Press, 5th ed. 2002)

²⁴⁰ Até a definição de algoritmo é apresentada - “a finite sequence of steps to solve a logical or mathematical problem or perform a task” (Microsoft Computer Dictionary, Microsoft Press, 5th ed. 2002).

invenção. Contudo, a integração de uma ideia abstrata, lei da natureza ou fenômeno natural a uma aplicação prática pode ser elegível para proteção de patente²⁴¹.

A despeito da maior orientação por meio das diretrizes de exame, mais uma vez o elemento decisivo é o “efeito técnico”, em detrimento dos demais requisitos de patenteamento, demonstrado em disputas judiciais de casos notórios. Na ausência desses exemplos, as diretrizes pouco contribuem para esclarecer as questões levantadas pela literatura. Com isso, o tema permanece polêmico e tem levado a frequente reversão de decisões do USPTO em ações judiciais, criando incerteza jurídica.

Na China, as diretrizes de exame de patentes do CNIPA para invenções da IA passaram por três mudanças implementadas desde 2017. A primeira modificação definiu que uma invenção relacionada a programa de computador, que tenha características técnicas e se destine a solução de problemas técnicos não seria excluída de patenteabilidade.

Uma segunda alteração, de dezembro de 2019, em vigor em fevereiro de 2020, modificou o critério de elegibilidade das invenções da IA. A solução que adotar um meio técnico (elemento físico, hardware ou dispositivo) para resolver um problema técnico e obter efeito técnico, para tornar a solução uma solução técnica (Artigo 2 (2) da Lei de Patentes Chinesa). Foi adicionada uma seção (6) ao Capítulo 9 da Parte II das Diretrizes de exame relacionada a pedidos que incluam recursos abstratos como algoritmos ou métodos de negócios. Tais recursos abstratos não devem ser separados dos recursos técnicos e o exame deve observar o seu todo com vistas a analisar os problemas técnicos, os meios técnicos utilizados e os efeitos técnicos alcançados. Se a reivindicação incluir apenas características abstratas, sem quaisquer características técnicas, a reivindicação deve ser rejeitada como regra ou método de atividade mental (Artigo 25 da Lei de Patentes Chinesa). Se a reivindicação incluir quaisquer recursos técnicos além de recursos abstratos, deve ser examinada para ver se constitui uma solução técnica (Artigo 2 (2) da Lei de Patentes). O algoritmo de IA em si mesmo não é considerado um meio técnico e, portanto, continuou inelegível (Zhang, 2021; Wei, 2022).

A terceira modificação nas diretrizes de exame, de novembro de 2020 e em vigor desde junho de 2021, decorreu da modificação da Lei de Patentes Chinesa em outubro de 2020 (Huang et al., 2020). A mudança na Lei estabeleceu que a elegibilidade para patentes de invenção relacionada a IA é considerada "uma etapa concluída por um computador", como meio técnico qualificado, enquadrado em duas situações: 1) a reivindicação define um hardware e o conteúdo controlado ou processado pelo hardware executa um programa de computador; e 2) mesmo que

²⁴¹ "That a mathematical equation is needed to complete the claimed method and system does not condemn the claims to abstraction", *Alice Corp.* (citando *Diamond v. Diehr* 1981).

nenhum hardware seja definido diretamente, a solução preenchida pelo programa de computador pode refletir o conteúdo controlado ou processado pelo hardware. Com isso, aplicativos de algoritmos relacionados à IA tornaram-se elegíveis, mesmo quando não vinculados a um hardware. Se um recurso de algoritmo levar a uma melhoria no desempenho interno do computador (como reduzir o armazenamento de dados, reduzir a transferência de dados ou acelerar o processamento de hardware), a contribuição do algoritmo será considerada na avaliação da inventividade (exceto para métodos de negócios envolvidos em soluções financeiras, sujeitos a tratamento mais rigoroso).

De acordo com esta nova orientação, o pedido de patente não deve separar o modelo de negócio do método algorítmico, que deve ser analisado como um todo, englobando os meios técnicos envolvidos, os problemas técnicos a serem resolvidos e os efeitos técnicos a serem obtidos. Caso o pedido não possua essa característica de problema/efeito técnico, será considerado um algoritmo abstrato, não sujeito a proteção; caso contrário, torna-se elegível²⁴².

Além das mudanças legais com vistas a acomodar patentes relacionadas a invenções algorítmicas da IA, o crescimento do patenteamento de invenções em geral é um objetivo explícito perseguido pelo governo Chinês, empregando vários tipos de incentivos financeiros. A meta de crescimento de patentes inclui desde a concessão de um valor fixo para “reembolso” das taxas do pedido (Lei, Sun e Wright, 2012), subsídios financeiros diretos²⁴³, isenções tributárias²⁴⁴ e benefícios sociais concedidos pelos governos provinciais aos depositantes de patentes (pesquisadores e empresas)²⁴⁵ e inúmeros outros benefícios conferidos pelas patentes

²⁴² No Japão e na Coreia do Sul o software convencional é protegido por direitos autorais, mas se reivindicado junto com o hardware, a combinação do método de operação e do meio legível por computador que implementa o software tornam-se elegíveis para patente. A mesma abordagem tem sido estendida às invenções de IA. O Manual de Patentes do Japão interpreta os requisitos para patenteamento com base em exemplos listados em seu Anexo. As diretrizes de exame de patentes da Coreia do Sul foram alteradas em 2018, com vistas a fornecer princípios e casos exemplares para invenções relacionadas à IA, que são elegíveis “quando software e hardware cooperam”, embora exija descrição detalhada de forma que os especialistas possam reproduzir a invenção.

²⁴³ Yang (2022) descreve os vários tipos de incentivos, desde subsídios financeiros diretos, como o Fundo de Apoio a Pedidos de Patentes de Pequim, segundo o qual uma entidade poderia receber até ¥ 5 mil (cerca de US\$ 770) para cada pedido de patente de invenção e ¥ 500 (US\$ 77) para cada modelo de utilidade ou desenho industrial e incentivos adicionais à medida que mais patentes são depositadas e concedidas – acréscimo de ¥ 1 mil (US\$ 154) a partir do 101º pedido de patente de invenção depositado no ano e ¥ 1.500 (US\$ 231) por patente concedida; e até ¥ 500 mil (US\$ 77 mil) por ano para depósito de pedidos de patentes estrangeiras, incluindo Pedidos do Tratado de Cooperação (PCT).

²⁴⁴ Os pedidos de patentes também habilitariam alíquota reduzida (15%) de imposto de renda para empresas consideradas “empresas alta tecnologia”, frente à alíquota normal de 25% (artigo 27 (3) da Lei do Imposto de Renda Corporativo da China, 2007, em vigor em 01/01/2008).

²⁴⁵ O depósito de patentes é importante no sistema de pontuação usado para determinar se uma empresa se qualifica como de alta tecnologia, conforme diretrizes e Medidas administrativas do Ministério da Ciência e Tecnologia, com cadastramento a cada três anos.

na China²⁴⁶ (Mangelson et al., 2021; Yang, 2022), com vistas a alcançar as metas do governo central²⁴⁷.

No entanto, a preocupação (e pressões do exterior) com “comportamentos anormais” e pedidos de patentes atípicos, vem levando a emissão de orientações pelo CNIPA desde 2007, medidas disciplinares a partir de 2018, limitação dos incentivos dos governos locais apenas para patentes concedidas, limitadas a até 50% das taxas efetivamente pagas e suspensão do financiamento federal em 2021. Calvin e Leung (2020), mencionam a alta “taxa de descarte” das patentes de invenção chinesas - expiração antes de completada sua vigência legal – de 37%, frente a apenas 15% nos EUA, como indicador da baixa qualidade dessas patentes.

O 14º Plano Quinquenal de Desenvolvimento chinês (2021-25) estabeleceu a eliminação gradual dos subsídios dos governos locais ao patenteamento até 2025 e otimizar as políticas de incentivo para patentes de alto valor, priorizando qualidade, mais que quantidade, além de objetivos de curto e longo prazo²⁴⁸. O CNIPA vem atualizando os regulamentos contra pedidos de “patentes anormais” desde 2021, promulgando a esperada “Decisão sobre a Emenda dos Regulamentos de Implementação da Lei de Patentes da China em dezembro de 2023, com vigência a partir de janeiro de 2024 (Wang, 2024). A mudança de foco fica clara no Aviso 23/2021²⁴⁹, Plano Especial de Patentes para promover a inovação e o desenvolvimento de pequenas e médias empresas, por meio da transferência de tecnologia de empresas de menor

²⁴⁶ Patentes depositadas e concedidas fazem jus a vários outros benefícios, desde a promoção de produtos, solicitação de títulos acadêmicos pelos depositantes, candidatura de alunos do ensino fundamental e médio para níveis mais altos de escolaridade, programas de matrícula em universidades, dispensando exames de admissão, e até mesmo redução de penas judiciais (receber uma patente conta como um “serviço meritório importante”) (Yang, 2022).

²⁴⁷ O governo central da China empregou diversos incentivos para patenteamento de invenções por meio de políticas, diretrizes e mudanças legislativas. Metas estabelecidas desde o plano quinquenal de ciência e tecnologia do governo central, em 2006, são atualizadas nas sucessivas versões da Estratégia Nacional de Propriedade Intelectual (a partir de 2008, quando a PI foi alçada a prioridade nacional) (Aviso 18/2008 do Conselho de Estado) e da Estratégia Nacional de Desenvolvimento de Patentes do país (Lei, Sun e Wright, 2012), até 2020. A meta no 12º Plano Quinquenal de Desenvolvimento (2011-15) era 3,3 patentes de invenção para cada 10 mil habitantes, passando para 12 patentes/10 mil habitantes no 13º Plano (2016-20), cabendo registrar que a meta superada, com 15,8 patentes/10 mil habitantes no final de 2020. A política de inovação da China tem a peculiaridade de ser definida no plano federal, mas sua implementação se dá de forma descentralizada pelas diferentes províncias (Sun, Lei e Yin (2018), cujo maior ou menor êxito tem implicações econômicas e políticas junto ao governo central. Incentivos semelhantes existem em todas as províncias e cidades chinesas, levando ao crescente depósito de pedidos para obter os subsídios governamentais.

²⁴⁸ O Plano estabelece metas de curto prazo (até 2025) – valor adicionado das indústrias com utilização intensiva de patentes representar 13% do PIB; valor adicionado das indústrias de direitos de autor representar 7,5% do PIB; total anual de royalties da importação e da exportação de PI alcançar ¥ 350 bilhões (US\$ 54 bilhões); e meta de 12 patentes de invenção de alto valor para cada 10.000 habitantes – e de longo prazo (até 2035) – regime de PI da China situado entre os mais competitivos do mundo; a PI da China deve promover o desenvolvimento de inovação e empreendedorismo; o valor da PI deve ser enraizado na cultura nacional da China; e a PI da China sistema deve ser completo e permitir que o país participe de cooperação e governança global (Yang, 2022).

²⁴⁹ Disponível em: https://www.gov.cn/zhengce/zhengceku/2021-03/27/content_5596164.htm. Acesso em: 12 jul. 2024

porte e universidades, envolvendo empresas estatais, com transferência de recursos do governo central ¥ 100 milhões (US\$ 15,4 milhões) para cada província com melhores indicadores de desempenho na comercialização de PI.

5.3.1. O caso DABUS

Diversos ENPIs receberam pedidos de depósito de patentes em nome do sistema DABUS (*Device for the Autonomous Bootstrapping of Unified Sentience*), com as decisões disponibilizadas nos respectivos Websites. O caso é polêmico e pode se tornar paradigmático pelo pioneirismo em pedido de patente de invenção gerada autonomamente por IA.

O DABUS é um sistema conexcionista de IA desenvolvido por Stephen Thaler, da Universidade de Surrey/Inglaterra e fundador da Imagination Engines Inc, que o criador alega ser capaz de gerar autonomamente invenções. Não se pretende analisar os casos na íntegra, cujos pedidos foram submetidos em 17 escritórios e no âmbito do PCT, com muitos casos ainda pendentes de apelação (The Artificial Inventor Project)²⁵⁰ (Quadro 5.1). Pela legislação vigente, o envolvimento humano é um requisito em todo o mundo para a patenteabilidade de invenções geradas por IA, embora haja uma certa fragilidade nas decisões sobre o caso, conforme será visto nos casos do EPO, USPTO e CNIPA.

O caso reacendeu temas controvertidos da PI de invenções relacionadas a IA – a autonomia dos sistemas e/ou o reduzido envolvimento humano na geração das invenções –, contraposto unissonamente pelos escritórios nacionais com base no paradigma formal que estabelece que apenas humanos podem ser considerados inventores, mesmo quando isso não esteja explicitamente escrito nas leis de patentes, mas sim no arcabouço jurídico mais amplo, e sem qualquer avaliação sobre a efetiva autonomia do sistema.

Os pedidos de patente do DABUS (EP 18275163 e EP 18275174) foram negados pelo EPO, em janeiro de 2020. O depositante interpôs recurso, negada pela Câmara de Apelação Jurídica (Comunicado da Decisão, 2021), mantendo a decisão do EPO, pelo fato do inventor não ser uma pessoa física.

Os tratados internacionais preveem que o pedido de patente designe um inventor, mas deixam a interpretação do termo para as leis nacionais (EPC Implementing Regulations (Rule 19) e EPC (Artigo 81)). A EPC (Artigos 60, 62 e 81) menciona explicitamente o termo “inventor”, embora nem esta Convenção, nem seu Regulamento de Implementação, definam ou qualifiquem o termo. Segundo Stierle (2020:919), as Câmaras de Recurso estão tendo que lidar com a questão tardia da designação do inventor ter que ser uma “*natural person*”, que não

²⁵⁰ Disponível em: <https://artificialinventor.com/patent-applications/>. Acesso em: 12 jul. 2024

é um entendimento vinculante para o EPO e tribunais nacionais.

Quadro 5.1 - Situação dos pedidos de patentes em nome do Sistema DABUS em vários países

País	Patente/Pedido nº.	Situação	Justificação
África do Sul	ZA2021/03242	Aceito	
Reino Unido	GB1816909.4 e GB1818161.0	Negado e atualmente em apelação	A invenção só pode ser concedida a uma pessoa física, de acordo com as seções 7 e 13 da Lei de Patentes.
Europa	EP3564144	Negado e atualmente em apelação	O artigo 81 da CPE exige que um inventor seja uma pessoa com capacidade jurídica.
Europa	EP3563896	Negado e atualmente em apelação	O artigo 81 da CPE exige que um inventor seja uma pessoa com capacidade jurídica.
Alemanha	10 2019 128 120,2	Permitido devido à mudança do inventor para uma pessoa física.	O § 37(1) da Lei de Patentes exige que o inventor seja uma pessoa física.
Alemanha	10 2019 129 136,4	Negado	O § 37(1) da Lei de Patentes exige que o inventor seja uma pessoa física.
Coréia	KR 10-2020-7007394	Negado e atualmente em apelação	O artigo 33(1) estabelece que uma “Pessoa com Direito a uma Patente” é “[uma] pessoa que cria uma invenção ou um sucessor dela tem direito a uma patente ao abrigo desta Lei”.
Japão	JP 2020-543051	Negado e atualmente em apelação	Negado porque a Lei de Patentes não contempla pessoas não naturais como inventoras.
Nova Zelândia	776029	Negado e atualmente em apelação	A Lei de Patentes exige que o inventor seja uma pessoa física.
China	CN 2019800061580	Negado e atualmente em apelação	Os direitos de patente só podem ser concedidos a “sujeitos civis”, que são definidos como “indivíduos ou entidades que detêm direitos civis, cumprem obrigações civis e são afetados por resultados legais”.
Estados Unidos	16/524.350	Negado. Recursos esgotados	O inventor deve ser uma pessoa física.
Austrália	AU 2019363177	Negado. Recursos esgotados	O precedente exige que a invenção surja da mente de uma pessoa física.
Canadá	CA 3.137.161	Negado	As máquinas não têm direitos sob a lei canadense.
Arábia Saudita	521422019	Negado.	O inventor deve ser uma pessoa física.
Taiwan	TW108140133	Negado	As Diretrizes de Exame de Patentes estabelecem que o inventor deve ser humano.
Taiwan	TW108137438	Negado	As Diretrizes de Exame de Patentes estabelecem que o inventor deve ser humano.

Fonte: Cloud, Chankong e Hopfinger (2024)

Os pedidos de patente foram originalmente apresentados sem preencher o campo “inventor” nos formulários, apenas após exigência do EPO sendo incluído em nome do Sistema de IA conexionista DABUS, com o titular como seu empregador. A decisão alegou que nem requisitos substantivos de patenteabilidade, nem qualquer direito do público de saber quem é o verdadeiro inventor, sugerem que um sistema de IA pode ser um inventor sob o regime de patentes europeu.

A decisão se baseou em uma interpretação dos artigos da legislação que, no entanto, não mencionam explicitamente que o inventor deva ser uma pessoa física, nem proíbe ser um sistema de IA. Os termos da EPC não são explícitos quanto à exigência do inventor ser uma

pessoa física. O Artigo 90 (3) (5) da EPC prevê recusa de um pedido de patente quando não forem atendidas exigências do exame. O Artigo 81 da EPC, prevê que o pedido “deve designar o inventor”, cuja interpretação do termo cabe às leis nacionais. A Regra 19(1) do Regulamento de Implementação do EPC estabelece que a “designação [do inventor] deve indicar o nome de família, nomes próprios e endereço completo do inventor”, mas o EPO “não deve verificar a exatidão da designação do inventor” (Regra 19(2) de Implementação). O Artigo 62 da EPC estabelece que o requerente ou titular de uma patente europeia tenha o direito a ser mencionado como tal perante o EPO, enquanto o Artigo 60 dispõe que o direito à patente pertence ao inventor ou ao seu sucessor titular. Se o inventor for trabalhador assalariado, o direito à patente europeia será determinado de acordo com a lei do Estado em que o trabalhador exerce a sua atividade principal. Contudo, a EPC tem um papel limitado na determinação tanto da inventividade quanto da propriedade, abrindo espaço para futuras reinterpretações diferentes do caso em questão pela lei nacional e pelos tribunais nacionais.

Stierle (2020) observa que o requisito do inventor ser uma pessoa física parece um padrão aplicável internacionalmente, baseado em consultas do EPO a várias legislações nacionais dos estados membros e de relatório de uma mesa redonda de especialistas do IP5 sobre IA.

O autor faz dois questionamentos importantes sobre a decisão. A falta de capacidade da IA possuir direitos de inventor (tal como o direito de ser mencionada como tal), por não ter personalidade jurídica e pela diferença funcional entre nomes dados a coisas (sistemas de IA) e nomes dados a pessoas, razões alegadas como não atendendo aos Artigos 62 e 81 e a Regra 19(1) da EPC. O autor alerta que o EPO não tem competência para julgar capacidade jurídica, que cabe às legislações nacionais²⁵¹. O autor considera pouco convincente o argumento da distinção entre nome de coisas e pessoas, pois ambos têm a função de identificá-las e fazem parte da sua personalidade, permitindo-lhes, assim, exercer os seus direitos (protegidos pelas legislações nacionais).

Com a recusa em aceitar os pedidos de patentes, baseada exclusivamente em aspectos formais da interpretação da lei, o EPO sequer precisou examinar se o DABUS realmente inventou ou se um inventor humano usou o DABUS como ferramenta. Tal como na literatura, na prática a autonomia dos sistemas de IA parece ser hoje mais uma questão de “opinião” do que de uma efetiva avaliação, sem demandar que o depositante forneça provas da autoria da

²⁵¹ O autor cita dois casos, ambos de 2017, que iriam de encontro com a mencionada interpretação da lei. Na Estônia foi sugerido conceder personalidade jurídica a IA; e uma proposta de criação personalidade eletrônica para robôs autônomos foi aprovada pela CE ([P8_TA\(2017\)0051](#))

invenção, que, por seu turno, poderá vir a ser futuramente exigida pelos tribunais e riscos de direitos invalidados (Gervais, 2019:16-17). De fato, o caso do DABUS evidencia que a pergunta sobre a real autoria da invenção relacionada a IA não é feita, exceto se o pedido de patente nomear um sistema de IA como inventor. Quantas patentes de invenção relacionadas a IA podem ter sido concedidas de forma autônoma por sistemas de redes neurais artificiais, apenas por que o depositante não os indicou pelo risco de ter um pedido de patente invalidado?

O USPTO também negou o pedido de famílias de patentes²⁵² do DABUS nº 16/524.350 depositado em 2019, nos termos da Decisão sobre a petição (USPTO 2020)²⁵³. Ainda que não conste explicitamente da lei que o inventor deva ser uma pessoa física, esse entendimento foi baseado em artigos da lei que utilizam pronomes empregados para pessoas físicas (“*himself*” ou “*herself*”), da necessidade do inventor realizar uma declaração na submissão do pedido, que não poderia ser atendido por um sistema/máquina, na jurisprudência da decisão do caso *Univ of Utah v. Max Planck Gesellschaft zur Forderung der Wissenschaften*, onde o Tribunal de Apelações do Circuito Federal infere que as leis de patentes requerem que o inventor seja uma pessoa física, e da decisão *Burroughs Wellcome Co. v. Barrs Labs, Inc.*, onde é argumentado que para desempenhar a “um ato mental” inventores devem ser pessoas físicas, não firmas. Mas vale considerar o argumento do depositante na interposição do recurso: “*The referenced statutes are intended to compel an applicant to name a natural person even where the person does not meet the inventorship criteria*” (USPTO, 2020:6)²⁵⁴.

Pedidos de Patentes do DABUS também foram negados na China sob o mesmo argumento de um sistema de IA não poder ser listado como inventor (Wininger, 2022). Da mesma forma, foram recusados no *UK Intellectual Property Office* (UKIPO), no *German Patent and Trademark Office* (DPMA), no *Korean Intellectual Property Office* (KIPO). No *Australian Patent Office* a recusa do pedido foi revertida pelo tribunal, enquanto no *South Africa's Patent Office*, onde pedidos de patentes seguem o sistema de registro, não de exame, foram aceitos. Nos demais países, ainda aguardam decisão.

Invariavelmente, as decisões foram baseadas em aspectos formais e uma interpretação de muitos artigos das respectivas leis, sempre passíveis de revisão em tribunais, sem avaliar se

²⁵² Uma família de patentes é definida como um grupo de documentos de patentes, todos com o mesmo conjunto de prioridades (ou seja, grupo de documentos de patentes que protegem a mesma invenção).

²⁵³ Um outro caso interessante, cuja decisão é disponibilizada no Website do USPTO, é o Ex parte Hannun (ex parte Linden), 2018-003323 (2019), mas que não recebeu atenção da mídia. É um pedido de patente de tecnologia de reconhecimento de fala por aprendizado profundo, rejeitado no exame por ser considerado direcionado a uma fórmula matemática. A chinesa Baidu USA, LLC é listada entre as partes interessadas. Disponível em: [https://www.uspto.gov/sites/default/files/documents/Ex parte Hannun 2018-003323.pdf](https://www.uspto.gov/sites/default/files/documents/Ex%20parte%20Hannun%202018-003323.pdf). Acesso em: 10 mai. 2022

²⁵⁴ Alguns autores mencionam que o USPTO concedeu patentes a inventores não humanos, embora não explicitamente e não necessariamente com seu conhecimento (Gervais, 2019; Pearlman, 2018; Vertinsky, 2017).

sistemas de IA e máquinas podem inventar, se devem receber DPI e se podem qualificar como inventores. Em síntese, o conceito de inventividade e autonomia da IA não é sequer avaliado e o problema fica em suspenso. Stierle (2020) defende a necessidade de uma reflexão mais densa, citando Auditor do UKIPO que negou os pedidos de patente do DABUS, mas reconheceu a legitimidade do debate sobre a invenção da IA, em vez de ficarmos arbitrariamente presos à legislação atual. Stierle (2020:924) conclui que:

“the current approach of the European patent system might squeeze today’s innovation processes into the old-fashioned paradigms of yesterday’s patent system. Therefore, further studies and discussions on the capacity of AI to invent as well as on the pros and cons of AI inventorship are more than welcome”.

Sierle (2021) detalha possíveis implicações de uma eventual mudança no sistema, mesmo com riscos de emaranhados de patentes (*patent tickets*) e a necessidade de harmonização internacional.

5.4. O que muda com a IA generativa?

A fase dos sistemas de IA embarcados não suscitaram grandes dilemas em PI, uma vez incorporados a hardwares. Desafios efetivos vieram quando se desvincula dos equipamentos físicos, principalmente com os peculiares sistemas de redes neurais profundas que proliferaram no mercado na década de 2010 e, principalmente, com o surgimento daqueles baseados em modelos generativos. .

Com tal foco, muitos estudos teóricos e conceituais sobre DPI relacionados a IA e bancos de dados subjacentes foram desenvolvidos desde a década passada, recebendo atenção especial da Organização Mundial da Propriedade Intelectual (OMPI), com estudos, elaboração de documentos técnicos sobre os desafios ao sistema de PI (WIPO, 2020a), consulta pública (WIPO, 2020b) e dez seminários ou sessões anuais ou bianuais, entre 2019 e 2024, em fóruns da OMPI (*WIPO Conversation on Intellectual Property and Frontier Technologies*). O tema também mobilizou a agenda dos principais escritórios nacionais de PI a partir da virada da década, como o IP5²⁵⁵, fórum dos cinco maiores escritórios nacionais de propriedade intelectual – *United States Patent and Trademark Office (USPTO)*, *China National Intellectual Property Administration (CNIPA)*, *European Patent Office (EPO)*, *Japan Patent Office (JPO)* e *Korean Intellectual Property Office (KIPO)* –, bem como nos escritórios alemão, britânico, canadense, entre outros.

O debate, que parecia se tornar insípido, é retomado de forma mais aprofundada com o surgimento da IA generativa, ainda sem um posicionamento claro e recomendações conclusivas

²⁵⁵ Sobre o IP5 ver: <https://www.fiveipoffices.org/about>. Acesso em: 12 out. 2024

no sentido de uma orientação geral e harmonização internacional de direitos de PI sobre IA. Atualmente, as regras existentes apresentam diferenças entre os países, ainda que tais invenções transcendam fronteiras nacionais, tanto em termos da sua criação e captura de valor, quanto de impactos indubitavelmente globais dessas tecnologias cada vez mais poderosas. Boa parte da dos temas mais importantes está contida em documentos do Congresso americano, em propostas de revisão das diretrizes nacionais de exame de patentes, ainda não promulgadas, e nas discussões e documentos nos fóruns *Intellectual Property and Frontier Technologies*²⁵⁶ e *Standing Committee on the Law of Patents (SCP)*²⁵⁷.

O regime legal de PI estruturado para as tecnologias da IA preditiva é intrincado e híbrido, contemplando múltiplos ativos. Esses ativos protegem insumos (bases de dados) e produtos da IA (as invenções propriamente ditas, assistidas, implementadas ou geradas pelos sistemas de redes neurais profundas) (Drexel et al., 2020; Gervais, 2019; 2020; Schirru, 2020; Hartmann e Henkel, 2020), abarcando instrumentos do direito autoral (que protegem as bases de dados de treinamento dos modelos) e patentes (sobre as invenções da IA)²⁵⁸. O extenso regime legal abarca também direitos *sui generis* sobre investimentos para construção de bases de dados essenciais para as invenções da IA em algumas jurisdições (União Europeia, Singapura, Japão) – sem esquecer o uso generalizado de medidas de restrição técnica sobre bases de dados digitais.

O desenho industrial, por seu turno, se aplica a invenções de IA que envolvem também dispositivos físicos, como o Google Glass e possivelmente os óculos inteligentes da Meta (Tsang e Lee, 2022). O registro de marcas também é importante, com as marcas das cinco BigTechs americanas atualmente liderando o ranking das mais valiosas do mundo, encabeçadas pela Apple cuja marca alcançou mais de US\$ 1 trilhão em 2024 (Kantar, 2024)²⁵⁹, inclusive o uso frequente de marcas submarinas (Fink et al., 2022)²⁶⁰. Uma parte do conhecimento é

²⁵⁶ Disponível em: https://www.wipo.int/about-ip/en/frontier_technologies/. Acesso em: 14 out. 2024.

²⁵⁷ Disponível em: <https://www.wipo.int/policy/en/scp/>. Acesso em: 16 out. 2024.

²⁵⁸ Invenções geradas por IA não são passíveis de proteção por PI – da mesma forma que obras criadas por IA –, ao contrário daquelas inventadas por um ser humano. As invenções assistidas ou implementadas por IA, enquanto uma ferramenta, têm direito a patentes (da mesma forma, a criação humana é pré-requisito para proteção por direito autoral). Sobre esses conceitos, ver Drexel et al. (2020). O caso DABUS, já comentado, reacendeu temas controversos da PI de invenções relacionadas a IA – a autonomia dos sistemas e/ou o reduzido envolvimento humano na geração das invenções –, recusado unisonamente pelos escritórios nacionais de PI com base no paradigma formal que estabelece que apenas humanos podem ser considerados inventores, mesmo quando isso não esteja explicitamente escrito nas leis nacionais de patentes, apenas no arcabouço jurídico mais amplo, e sem qualquer avaliação sobre a efetiva autonomia do sistema.

²⁵⁹ Disponível em <https://www.kantar.com/inspiration/brands/revealed-the-worlds-most-valuable-brands-of-2024>. Acesso em: 21 jul. 2024

²⁶⁰ O registro de marcas divulga informações estratégicas de produtos futuros e a entrada em segmentos do mercado, o que significa um *trade-off* entre proteção e divulgação, contornado pelas gigantes da tecnologia – Apple, Amazon, Google, Cisco, Intel e IBM – por meio de marcas submarinas. Com isso, adiam o registro em

mantida em sigilo, por meio de *trade secrets*, que em algumas jurisdições possui legislação específica, como o *US Defend Trade secrets Act*²⁶¹ e a EU Directive 2016/943²⁶².

Por esse esquema, há vários direitos de PI, o que corrobora a hipótese da sua habitual complementaridade, e que convivem harmoniosamente com o regime *open-source*, que auxiliam na promoção do uso e difusão dos sistemas de IA e na captura de efeitos de rede, algo fundamental para a definição do *design* tecnológico dominante, sem a renúncia à sua propriedade, mas não sem um conjunto de incertezas.

5.4.1. Os novos desafios para invenções e criações da IA

Os ativos de PI, sobretudo patentes (e também) direito autoral, começaram a implicar em incerteza jurídica potencial, devido a aspectos precariamente abordados, recolocados em pauta com a chegada da IA generativa. A aplicação do arcabouço legal de PI às invenções peculiares da IA obrigou os escritórios nacionais de PI a tentar esclarecer os dilemas remanescentes por meio de simples orientações administrativas para exame dos pedidos de patentes de IA, sem qualquer mudança no arcabouço legal – com exceção da China, que alterou sua lei de patentes mencionando explicitamente os recursos algorítmicos. Isso está criando insegurança jurídica potencial, uma vez que atos administrativos não são impositivos aos tribunais e não garantem o *enforcement* de patentes da IA.

Muitos pedidos de patentes relacionados à IA preditiva e, principalmente, generativa, ainda não se converteram em patentes concedidas²⁶³, o que impede avaliar o grau de litígio na área. Contudo, Ferrandis e Lizarralde (2022) argumentam que as empresas podem preferir deixar de patentear por hesitar entrar em disputas, uma vez que invenções de IA como matéria patenteável ainda é incerta. Com a concessão de muitos pedidos de patentes, é grande o potencial de litígio (Chau et al., 2022; Handler, 2022), com risco de direitos invalidados e uma insegurança jurídica que enfraquece o regime de proteção legal e, certamente, a apropriabilidade no momento crucial de implementação da IA em setores *downstream*.

A “autonomia”, a “não explicabilidade” e a opacidade dos modelos, considerados

escritórios importantes, que dispõem de sistemas online “pesquisáveis”, antecipando eventual bloqueio de rivais por meio de um primeiro registro, inclusive por empresas de fachada, em lugar distinto de onde pretende usar as marcas, como países como Honduras, Jamaica e outros. Ainda que o princípio da territorialidade da PI impeça que um direito concedido em um país seja usado para litigar contra infração em outro país, pode ser usado para impedir a concessão de um direito em outro país por formar a “técnica anterior”.

²⁶¹ Public Law 114–153/2016. Disponível em: <https://www.congress.gov/114/statute/STATUTE-130/STATUTE-130-Pg376.pdf>. Acesso em: 21 jul.2024

²⁶² Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L0943&from=PT>. Acesso em: 21 jul.2024

²⁶³ Menos de 1% das 340.000 famílias de patentes relacionadas à IA enfrentaram litígios até a publicação de WIPO (2019).

caixas-pretas, cuja estrutura interna muda com o aprendizado/adaptação com os dados, são problemas que se aprofundam com o tipo de aprendizado empregado, a escala e a complexidade dos modelos generativos, reabrindo o debate sobre autoria da invenção, elegibilidade e requisitos legais para patenteamento, particularmente da suficiência descritiva e atividade inventiva. Esses dilemas, que já desafiavam o regime de PI – como abordado por uma parte da literatura e exemplificado pelo famoso caso DABUS – têm levado o Congresso americano a aprofundar o debate sobre as questões de PI das invenções geradas por IA.

O regime legal de patentes em IA nos EUA – o teste em duas etapas para elegibilidade que é empregado desde a orientação da Suprema Corte emitida para o caso *Alice Corp v. CLS Bank*, em 2014²⁶⁴ –, é considerado pelo Congresso confuso e sem consistência, o que exige esforços para sua aplicação pelos juízes dos tribunais. O teste tem efetivamente um alto grau de subjetividade, que aliado à complexidade dessas invenções estaria criando um ambiente de incerteza jurídica.

Por conseguinte, o Congresso vem tentando assumir diretamente o tema da elegibilidade para patentes das invenções da IA por meio de projetos de lei, apresentados em 2023, pelos senadores Thom Tillis (repblicano) e Chris Coons (democrata) – respectivamente, o *Patent Eligibility Restoration Act*²⁶⁵, que visa reduzir a subjetividade dos critérios de elegibilidade para patentes de IA, e o *Promoting and Respecting Economically Vital American Innovation Leadership Act*²⁶⁶, que objetiva acelerar a inovação nos EUA frente à rivalidade chinesa²⁶⁷.

Além disso, desde 2022 o Senado americano pressiona para um melhor esclarecimento dos DPIs relacionados a IA, enviando Carta²⁶⁸ aos escritórios de patentes (USPTO) e copyright (USCO) onde solicita uma comissão nacional para análise profunda dessa tecnologia e avaliação das questões de PI subjacentes. Aventou até mesmo a necessidade de alteração na legislação existente (por exemplo, a criação de um eventual direito *sui generis*) para incentivar invenções geradas por IA não elegíveis para patentes sob a lei atual – aparentemente algo similar à singularidade legal mencionada por Gervais e Nay (2023). Discussões a portas

²⁶⁴ A decisão da Suprema Corte sobre o caso *Alice Corp. v. CLS Bank International*, em 2014, foi incorporada às regras de exame de patentes de software e IA pelo USPTO no Manual of Patent Examining Procedure, como mencionado.

²⁶⁵ Disponível em <https://www.tillis.senate.gov/services/files/4B41CBF2-57AB-4E8E-9E93-7D714A7AAB40>. Acesso em: 17 jul.2023

²⁶⁶ Disponível em https://www.coons.senate.gov/imo/media/doc/prevail_act_bill_text.pdf. Acesso em: 17 jul.2023

²⁶⁷ Essas iniciativas tiveram início em 2019 com a divulgação do quadro sobre a reforma da elegibilidade de patentes (Coons, 2019) lançado pelos congressistas americanos e os projetos de lei agora reapresentados (Zingg, 2021).

²⁶⁸ Disponível em <https://www.copyright.gov/laws/hearings/Letter-to-USPTO-USCO-on-National-Commission-on-AI-1.pdf>. Acesso em: 29 nov.2022

fechadas no Senado foram feitas no “*AI Insight Forum*”, que é parte da iniciativa “*SAFE Innovation*”, realizando nove reuniões entre setembro e dezembro de 2023.

A Ordem Executiva 14.110/2023 acatou as recomendações do Congresso e solicitou a elaboração de orientações atualizadas sobre elegibilidade de patentes de IA generativa, reforçando a discussão do tema desde a criação da *AI/Emerging Technologies Partnership*²⁶⁹. Novas Orientações para exame de pedidos de patentes assistidas por IA (USPTO 89 FR 10043)²⁷⁰ (*Inventorship guidance*), divulgada em 13/02/2024, estabeleceram a necessidade de determinar se a contribuição humana para a inovação é significativa o suficiente para se qualificar para uma patente quando a IA também contribuiu, além dos requisitos formais de patenteamento. Outra nova orientação do USPTO (89 FR 58128)²⁷¹ sobre invenções assistidas por IA, divulgada em 17/07/2024 (*Subject Matter Eligibility*)²⁷², formaliza em ato administrativo o controvertido teste Alice (válido para invenções de software e de IA) e fornece instruções para examinadores e partes interessadas sobre como determinar se a contribuição humana para uma inovação é significativa o suficiente para se qualificar para uma patente quando a IA também contribuiu²⁷³. Por fim, novas orientações para registro de direitos autorais de obras contendo material “gerado” por IA (Copyright Office 37 CFR Part 202)²⁷⁴, divulgadas em 16/03/2023.

De fato, as peculiaridades da IA generativa e seu avanço em ritmo acelerado parecem exigir orientações ainda mais claras (e mais objetivas). Principalmente após episódios como a concessão de direito autoral, em setembro de 2022, à autora Kristina Kashtanova, criadora da história em quadrinhos “*Zarya of the Dawn*”²⁷⁵, cujas imagens foram integralmente geradas a partir de *prompts* pelo modelo de difusão Midjourney (Copple, 2023). A decisão foi revista em

²⁶⁹ Disponível em <https://www.uspto.gov/initiatives/artificial-intelligence/ai-and-emerging-technology-partnership-engagement-and-events>. Acesso em: 29 nov.2022

²⁷⁰ Disponível em <https://www.federalregister.gov/documents/2024/02/13/2024-02623/inventorship-guidance-for-ai-assisted-inventions>. Acesso em: 14 fev.2024

²⁷¹ Disponível em: <https://www.federalregister.gov/documents/2024/07/17/2024-15377/2024-guidance-update-on-patent-subject-matter-eligibility-including-on-artificial-intelligence>. Acesso em: 11 ago.2024

²⁷² Esta orientação era há tempos aguardada, uma vez que sob o atual paradigma legal *Alice / Mayo*, uma invenção é inelegível para proteção de patente se for direcionada a uma exceção judicial (ou seja, uma ideia abstrata, uma lei da natureza ou um fenômeno natural) e não tiver elementos adicionais que representem significativamente mais do que a exceção judicial

²⁷³ A natureza não vinculativa das orientações do USPTO sobre tribunais tem sido uma preocupação recorrente nas invenções que envolvam redes neurais artificiais e a nova orientação sobre elegibilidade para patentes foi recebida com apreensão por diversas organizações, como *American Intellectual Property Law Association (AIPLA)*, *Council for Innovation Promotion (C4IP)*, *Pharmaceutical Research and Manufacturers of America (PhRMA)*, *Intellectual Property Owners Association (IPO)*, *Association of University Technology Managers, Inc. (AUTM)*, que parecem preferir apoiar aprovação do *Patent Eligibility Restoration Act*. (McDermott e Brachmann, 2024)

²⁷⁴ Disponível em <https://www.govinfo.gov/content/pkg/FR-2023-03-16/pdf/2023-05321.pdf>. Acesso em: 30 mar. 2023

²⁷⁵ Disponível em: <https://www.kris.art/tutorials>. Acesso em: 28 out.2023

outubro, com o registro limitado apenas à história em si, fruto da criatividade e originalidade da autora, excluindo as imagens do Midjourney. O legado, entretanto, foram muitas dúvidas sobre a real capacidade do USCO de identificar conteúdo criado por IA nos pedidos de registro de direitos de autor, além de levantar dúvidas sobre quantas invenções “geradas” por IA também poderiam estar obtendo direitos proprietários por pessoas físicas que simplesmente fizeram uma demanda por *prompt* a um sistema de IA, ou, alternativamente, quantas invenções valiosas deixarão de ser patenteadas por uma contribuição humana considerada insuficiente.

A nova abordagem em gestação ressalta que não é claro se o novo conteúdo de texto, imagem, código etc., gerado por IA, pode ser protegida por DPI e quem seria seu detentor, lembrando que os casos em que não seja possível podem contar com instrumentos contratual que regem seu uso, como nos termos de uso do regime *open-source*. Nos casos de saídas de IA em obras passíveis de proteção por direito autoral, a nova orientação do USPTO exige uma contribuição humana, que não se restringe a um simples *prompt* do usuário, podendo não estabelecer direitos autorais uma vez que apenas influencia a saída. WIPO (2024b) lembra uma decisão do Tribunal da Internet de Pequim de reconhecer direitos autorais para um usuário sobre uma imagem gerada por IA meramente ajustando prompts e parâmetros para que a imagem refletisse sua escolha estética e julgamento, o que introduziria incerteza jurídica sobre criações da IA generativa. Países como Índia, Irlanda, Nova Zelândia, África do Sul e Reino Unido, também concederiam direitos autorais a obras geradas por computador, sem autor humano, e a Ucrânia teria introduzido direitos para “objetos não originais” gerados por programas de computador (WIPO, 2024c). Omite, contudo, que o USPTO também concedeu direitos autorais sobre a obra integral “*Zarya of the Dawn*”.

Em linhas gerais, considerando as questões mais relevantes para os propósitos desta tese, se relacionam às patentes, cujas futuras orientações pelos respectivos ENPIs, refletidas em documento recente da OMPI, podem criar incertezas ainda maiores (WIPO, 2024b).

A abordagem que parece ganhar simpatia pelo mundo é o detalhamento dos tipos de saídas da IA, reproduzindo, em termos gerais, a tipologia de invenções relacionadas a IA proposta por Drexler et al. (2020), ainda que acrescida de uma quarta categoria relacionada aos modelos e algoritmos de IA. Essas categorias em discussão, que consideram a inventividade, autoria e com implicações em termos de patenteamento, seriam: 1) invenções assistidas por IA, correspondendo àquelas em que a IA é usada como simples ferramenta no processo inventivo, como na seleção de substâncias candidatas para descoberta de medicamentos; 2) invenções baseadas em IA, em que sistemas de IA são parte do processo inventivo, como nos dispositivos de tradução que incorporam aprendizado profundo; 3) invenções geradas por IA, ou seja,

desenvolvidas de forma autônoma por IA, sem intervenção humana, como as supostas saídas do sistema DABUS; e 4) modelos ou algoritmos, em que a invenção é a própria tecnologia central de IA, como técnicas, modelos ou algoritmos de IA novos ou aprimorados (WIPO, 2024b).

O primeiro caso envolveria questões sobre quem seria o inventor (o provedor dos dados, ou o desenvolvedor do modelo ou seu proprietário), com implicações em termos do conceito de pessoa habilitada na arte (PHOSITA) e ampliação do nível da etapa inventiva.

O segundo, levaria a questões sobre requisitos de divulgação da invenção e dos dados, inclusive para que o PHOSITA possa reproduzir a invenção.

O terceiro caso envolveria o problema da contribuição humana para a invenção (ou, alternativamente, a concessão de DPI para um sistema de IA), com implicações sobre a própria racionalidade do sistema de patentes.

Por fim, a última categoria, potencialmente a mais polêmica, envolve o patenteamento dos próprios modelos de IA, vistos como invenções em si mesmos. Com base na discussão da seção 5.3, o modelo deixa de desempenhar o papel de insumo das invenções da IA, passando a ser um produto, uma invenção. Modelos e seu código, pela legislação vigente, pertencem ao campo do direito autoral em sua expressão, mas pela proposta em discussão passariam para o campo do direito patentário. Quanto as ideias por trás desses modelos, seus códigos e algoritmos, parecem indubitavelmente simples métodos matemáticos, conhecimento fundamental que, pelas leis e convenções de PI, seriam verdades científicas não passíveis de qualquer tipo de proteção intelectual. Por sim, levantam muitas perguntas sobre qual seria o efeito técnico de modelos ou algoritmos, um dos requisitos formais para patenteamento de invenções.

O restante da discussão parece retornar às primeiras reuniões da OMPI, em 2019 e 2020, com a pergunta central envolver alternativas entre patentear, manter em segredo comercial ou a necessidade de criação de um direito *sui generis* (WIPO, 2024b), sem sequer mencionar a tendência crescente dos modelos de IA generativa em regime *open-source*. Adicionalmente, a pergunta sobre o inventor e o grau de autonomia dos sistemas de IA permanecem sem resposta.

Ainda que com algumas pequenas variações entre os países, por operarem sob diferentes códigos legais, essas futuras orientações para exame de patentes dos vários ENPIs parecem convergir para esta abordagem sumarizada em WIPO (2024b), como sugerem documentos de

discussão do escritório coreano²⁷⁶, do escritório brasileiro²⁷⁷, no fórum dos BRICS (2022) e informações não oficiais sobre futuras diretrizes do escritório chinês. A exceção parece os Estados Unidos, onde as últimas orientações continuam centradas na necessidade de identificar a contribuição humana da invenção e corroborar o controvertido teste Alice.

No entanto, o diabo mora nos detalhes. O tratamento distinto das invenções da IA em diferentes países parece decorrente não de diferenças substantivas nas respectivas legislações, mas nas diretrizes de exame, meros atos administrativos emitidos por ENPIs, que não constituem regulamentação substantiva, sem força e efeito de lei.

5.4.2. O *open source* da IA como um quasi-DPI

O *open source* parece ser uma tendência que veio para ficar. Originalmente um ecossistema de inovação descentralizado, em que milhares de “devs”. contribuíam para projetos, facilitando a evolução tecnológica e a diversidade de soluções, inicialmente restrito às startups e à academia, passou a constituir um importante instrumento de competição estratégica das empresas líderes²⁷⁸, que abraçaram o *open source*, disponibilizando modelos para a comunidade de “devs.” nas bibliotecas de treinamento da IA preditiva e, posteriormente, para o público em geral com a IA generativa.

O tema do *open source*, contudo, é ainda relativamente obscuro para o público de fora das comunidades e se torna um entre os muitos temas controvertidos relacionados à PI, apesar de ser um instrumento cada vez mais complementar, não antagônico, aos direitos de propriedade.

O regime *open-source* não significa que os modelos estão em domínio público e sejam totalmente livres, nem necessariamente gratuitos, mas se relaciona à liberdade de uso. O que os diferencia do modelo proprietário é a disponibilização do código-fonte para o usuário, que desse modo pode reaproveitá-lo no desenvolvimento e distribuição de seus próprios sistemas e aplicativos. O programa executável, entretanto, por envolver tempo e recursos adicionais, é usualmente cobrado por meio de assinaturas pagas.

O *open source* é uma alternativa ao regime fechado por sigilo comercial, mas não

²⁷⁶ Documento “Patent protection for AI outputs at KIPO”. Nov 6, 2024. Disponível em: https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_conv_ge_2_24/wipo_ip_conv_ge_2_24_ss06.pdf. Acesso em: 20 nov. 2024

²⁷⁷ Ver Webinar: Construindo patentes relevantes, da Inova Unicamp, em 16 de outubro de 2024. Disponível em: <https://www.youtube.com/watch?v=uNb4vkGXLz0>. Acesso em: 18 out. 2024. Ver também BRICS (2022).

²⁷⁸ O open-source pareceu se tornar um paradigma da IA nos últimos anos, disseminado na comunidade de “devs.” de IA, com 65 mil projetos de código aberto no GitHub em 2023 (Corso, 2024), mas também uma tendência no campo empresarial, havendo inclusive um programa de subsídios da grande empresa de capital de risco, Andreessen Horowitz, para IA de código aberto. Disponível em: <https://a16z.com/supporting-the-open-source-ai-community/>. Acesso em: 28 out. 2024

prescinde de direitos autorais expressos nos seus “Termos de uso”, padronizados e não negociáveis, o que é conhecido como distribuição sem atrito. As ações de usar, reproduzir ou distribuir do modelo *open-source* são suficientes para indicar a aceitação dos termos da licença²⁷⁹. No entanto, muitos desses termos não descrevem claramente direitos e obrigações a serem seguidas.

Cada licença segue regras próprias, inseridas em duas grandes categorias: as permissivas, com poucas restrições ao uso do código-fonte, com usuários livres para criar trabalhos derivados, comercializá-los e impor proteções de PI; e as *copyleft*s²⁸⁰, com condições a serem seguidas, como a vedação de imposição de PI e obrigatoriedade de licenciamento do material derivado em condições similares.

A StabilityAI, por exemplo, que enfrenta ações de direito autoral sobre as bases de dados usadas, disponibilizou modelo e pesos do Stable Diffusion em código aberto, com licenciamento permissivo, concedendo todos os direitos sobre as imagens geradas aos usuários (desde que não sejam ilegais ou prejudiciais). Os modelos LLaMA (incluindo o novo 3.1), da Meta AI, por outro lado, são disponibilizados em *open source*, exceto para empresas com mais de 700 milhões/mês de usuários ativos, para as quais é requerida licença diferenciada, para uso do modelo em produtos comerciais.

Diferem das licenças típicas do software livre nos termos da *Open Source Initiative (OSI)*²⁸¹, pois não há um padrão único para as licenças de código aberto da IA, mas diferentes tipos de licenças, com condições distintas, inclusive relacionadas aos conjuntos de dados de treinamento. Muitas vezes, as licenças são incompatíveis entre si (Rajbahadur et al., 2022), que Bernier e Thorogood (2020) consideram que em algum momento exigirão padronização, por questões de interoperabilidade e uma vez que a combinação de licenças distintas pode criar ambiguidades jurídicas, custos e riscos potenciais.

A disponibilização do código-fonte permite aos usuários reaproveitá-lo no desenvolvimento e distribuição de novos sistemas, em condições similares ou não, e sem eliminar a possibilidade de cobranças por assinatura. Licenças de código aberto não excluem necessariamente PI por patentes, ainda que sejam um ponto de controvérsia. Algumas licenças

²⁷⁹ Ao analisar diversas licenças de código aberto, Ferrandis e Lizarralde (2022) constataram que algumas são mais explícitas do que outras sobre quais ações implicam na “aceitação”.

²⁸⁰ Há dois tipos básicos de *copyleft*: o forte, que exige a distribuição da obra licenciada ou da obra derivada nos mesmos termos; e o fraco, que permite a combinação do trabalho licenciado com trabalhos sob termos diferentes (Hall, 2017).

²⁸¹ A *Open Source Initiative (OSI)* definiu (por consenso) o significado de licenças do software com características específicas, tais como não discriminar pessoas ou grupos ou campos de atuação, nos termos da *Open Source Definition (OSD)*. Disponível em <https://opensource.org/definition-annotated/> e <https://opensource.com/law/13/1/which-open-source-software-license-should-i-use>. Acesso em: 10 jun. 2023

são omissas, enquanto outras abordam explicitamente o tema, contemplando concomitantemente uma concessão de patentes no texto da licença. Contudo, como alertado pelo Google²⁸², “a mera inclusão da palavra “patente”, ou não, em uma licença de código aberto, não é suficiente para identificar se há uma efetiva concessão de direitos de patente e, mesmo quando há, o escopo da concessão pode não estar bem definido”.

Entretanto, as licenças de código aberto são cada vez mais usadas como elementos competitivos estratégicos na busca pela construção de ecossistemas e definição do *design* dominante da IA. Como resumido por Shafter (2016) há quase uma década, a IA de código aberto auxilia as empresas a permanecer na vanguarda da tecnologia “...elas não estão dando as chaves do seu sucesso: elas estão abrindo caminho para seu próprio futuro”.

O código aberto visa maximizar a difusão da inovação por meio de licenças projetadas em torno do conceito de distribuição, em lugar da recompensa direta por meio do licenciamento de PI (monetização da inovação), com a atração (efeito de rede) como meio de competir no mercado. Desse modo, corresponde a um instrumento que facilita a difusão e adoção em massa dos modelos, além de reduzir custos de transação – uma vez que os termos de cada licença são os mesmos para diferentes usuários, dispensando negociação entre as partes -, ampliando a inovação e estendendo o ecossistema de “devs”. É um instrumento de atração, aprisionamento e controle com papel central no desenvolvimento e liderança do mercado.

A coexistência de estratégias de patenteamento e *open source* tem longa trajetória complementar em TICs, como no software, onde empresas com grandes portfólios de patentes compartilham o código fonte. No entanto, a literatura recente destacou o código aberto como elemento competitivo estratégico em contextos que dependem de efeitos de rede e onde vigoram estratégias empresariais que combinam DPIs, segredos comerciais e licenciamento de código aberto para promover a disseminação da tecnologia em ritmo acelerado (Ferrandis e Lizarralde, 2022; Calvin e Leung, 2020). Com efeito, Alphabet/Google, Microsoft, Amazon, Meta AI, IBM, Alibaba, Tencent e Baidu, acumulam patentes e simultaneamente compartilham pesquisas em publicações, nas principais conferências da área e em comunidades de “devs.” de código aberto, algumas vezes até licenciando patentes a título gratuito.

Ferrandis e Lizarralde (2022) destacam que as BigTechs abrem seus códigos como forma de alcançar benefícios em mercados adjacentes, nos quais mantêm o patenteamento em componentes essenciais da tecnologia, em diferentes pontos da cadeia de valor (licenciamento duplo, onde além da versão *open-source*, há a versão comercial que tem desempenho otimizado

²⁸² Ver Google Open Source. Disponível em: <https://google.github.io/opencasebook/patents/>. Acesso em: 10 jun. 2023.

a partir de recursos extras) - como o modelo *free/ freemium* antes mencionado – e/ou de mercados contíguos (como o caso do código “não tão aberto” do Android do Google).

Atualmente, o mais importante mercado adjacente são os serviços de computação em nuvem e fornecimento de data center providos pelas empresas líderes (Blind et al., 2021), que atendem uma demanda alta e com perspectivas de expansão, considerando o potencial de ampla aplicação da IA em outros setores, como qualquer tecnologia de propósito geral.

Ao contrário das patentes, as licenças *open-source* não excluem terceiros de “pegar carona” na inovação, mas não significam acesso incondicional, havendo exclusão dissuasiva (a perda do direito de uso para aqueles que não cumprirem os termos de uso dos modelos) ou passiva (desincentivo de uso de PI pelos usuários por meio de cláusulas de reciprocidade, não afirmação e retaliação). Adicionalmente, disponibilizar apenas o código dos complexos modelos da IA generativa, que dependem de muitos outros detalhes técnicos, não significa abrir completamente o caminho para a inovação por seguidores. Rikap (2023b) questiona a real capacidade dos modelos *open-source* competirem com os modelos proprietários, por serem menores (em número de parâmetros); não poderem ser usados diretamente, demandando desenvolvimento de aplicativos de interface com o usuário; e exigirem adaptação para atender necessidades de terceiros *downstream* (que demanda capacidades e talentos escassos).

Calvin e Leung (2020) enfatizam o caráter híbrido das estratégias empresariais predominantes na IA, conjugando DPIs, como grandes portfólios de patentes, segredos comerciais e simultaneamente o compartilhamento do conhecimento e lançamento de modelos em código aberto. Blind et al. (2021) argumentam que essas estratégias têm objetivos de curto prazo - ganhar força por meio de uma plataforma de IA aberta - e objetivos de longo prazo - padronizar e *commoditizar* a tecnologia, controlando suas camadas essenciais e seus mercados.

Nos últimos dois anos, entretanto, os inovadores pioneiros parecem ter reorientado suas estratégias, passando a omitir cada vez mais detalhes técnicos importantes dos modelos, que parecem mesmo transformar o *open-source* em *open for business* à medida que a escala dos modelos avança exponencialmente. Isso fica claro no mencionado vazamento de documento interno do Google no qual a empresa mostra preocupação com o *open-source* ter fugido do controle e os modelos proliferarem pelo mundo.

Do ponto de vista jurídico, o intenso debate atual envolvendo *open source* e abertura dos modelos de IA é, portanto, um tanto equivocado, misturando forma de licenciamento (que comporta licenças de uso com diferentes formatos) e disponibilização de informações técnicas. Por exemplo, os usuários podem usar e interagir com o aplicativo ChatGPT, mas não podem reprogramar ou usar o modelo GPT-3,5 para inicializar seus próprios modelos de linguagem

(Luitse e Denkena, 2021). A Meta mistura coisas diferentes, "código aberto" e "recursos disponíveis para alguns usuários sob algumas condições", segundo esclarece Maffulli (2023).

Calvin e Leung (2020) atribuem as estratégias corporativas híbridas a uma variedade de objetivos empresariais, desde patentear por pressões competitivas até abrir modelos de IA para construir reputação, atrair talentos e incentivar clientes a usar produtos pagos, usando acordos seletivos de licenciamento de código aberto, permitindo participar da comunidade enquanto mantêm a ameaça legal por patentes.

O patenteamento seguiria uma estratégia defensiva contra ações de terceiros (Simonite, 2018) ao invés da intenção de litigar ofensivamente por eventual infração, diante dos altos custos financeiros e reputacionais que envolvem (caso da disputa entre Apple e Google em smartphones, em 2011, cujos valores superaram o investimento em P&D dessas empresas) (Duhigg e Lohr, 2012). Este seria o caso de permitir que terceiros participem do “*Google Open Patent Pledge*”²⁸³, *pool* de patentes, com o compromisso de não se envolverem em litígios contra a empresa. A estratégia de patenteamento pode também visar proteger a parte do conhecimento que lhe cabe na inovação aberta dos ecossistemas da IA, bem como acordos de compartilhamento de patentes – como parece ser o caso do Google e Tencent visando facilitar sua entrada da primeira no mercado chinês (Reuters, 2018).

Patentear pode não ser interessante pelos altos custos com taxas anuais para manter ativo um extenso portfólio e do requisito de divulgação e publicidade das patentes, comprometendo estratégias de sigilo, além de colidir com a filosofia da comunidade de “devs.” de IA, contrária ao patenteamento, sobretudo de métodos teóricos amplos, vistos como verdades matemáticas. De fato, patentes sobre técnicas fundamentais de IA e amplas aplicações potenciais significa desencorajar a inovação, uma vez que privatiza elementos básicos e exclui terceiros que poderiam desenvolver outros modelos de IA (Zingg, 2021), como os modelos fundação.

O regime *open source* tem desvantagens, pelo compartilhamento de segredos competitivos, inclusive com empresas rivais, e renúncia de receitas com licenciamento, mas vantagens como explorar efeitos de rede, atrair talentos e usuários/clientes potenciais para versões aprimoradas pagas, além da consolidação de *design* tecnológico (Calvin e Leung, 2020).

Como o compartilhamento do conhecimento é hoje parcial no *open source* da IA, possibilitando executar, mas não reproduzir a maioria dos modelos de código aberto do zero, ganhou força um movimento para definir o tipo de licença padrão de código aberto para IA de modo encerrar a ambiguidade atual. A proposta lançada pela *Open Source Initiative* (OSI), em

²⁸³ Disponível em: <https://www.google.com/patents/opnpledge/>. Acesso em: 10 jun. 2023

28 de outubro de 2024, versão 1.0²⁸⁴, fornece um padrão de licenciamento do que pode ser considerado um modelo de redes neurais profundas de código aberto, fruto de parceria com a Mozilla.

Por tal proposta, para que os modelos possam ser efetivamente reproduzidos, usados, estudados e modificados, além de analisados e avaliados, deverão contemplar todos os elementos-chaves do sistema de IA, tais como a disponibilização do código fonte (lembrando que existe diferença entre o código que é escrito, o compilador usado e o que está na licença), mas também seus detalhes técnicos, incluindo o código fonte do sistema (documentação e o software específico do sistema que implementa a funcionalidade principal), a especificação de arquitetura, dos parâmetros e hiperparâmetros (arquivos binários para os pesos do modelo, valores de configuração necessários para treinar os parâmetros do modelo, como taxa de aprendizado, etc.), o código fonte do treinamento (parte do sistema de software que transforma hiperparâmetros), os *frameworks* de software (bibliotecas e estruturas nas quais o código-fonte é construído), o tipo e quantidade do hardware empregado e, como ponto mais polêmico em termos técnicos e legais²⁸⁵, o conjunto de dados (não apenas aqueles usados no treinamento, mas também testar e ajustar o modelo) (Rubinstein, 2024; Corso, 2024).

A adoção de licença com tal nível de detalhamento certamente poderá ter impactos irrefutáveis sobre a tendência à disseminação do open-source e sobre o regime legal de PI, subsidiando o exame de pedidos de patentes e as exigências de suficiência descritiva, inclusive para que PHOSITA possa reproduzir o modelo.

5.4.3. Direito autoral, *fair use* e gargalos de bases de dados

A importância de amplos conjuntos de dados para o desenvolvimento e desempenho dos modelos de redes neurais profundas foi amplamente tratada pela literatura, com a BigData analisada em estudos sobre a IA preditiva, em meio a temas como privacidade, proteção de dados, vieses algorítmicos e o poder de mercado que conferem às empresas líderes.

A importância do aprendizado/treinamento dos modelos com massivas bases de dados (pessoais e não-pessoais), o tornam o insumo chave para a identificação de padrões e capacidade de fazer classificações, recomendações, previsões ou criação de conteúdo (novos dados) a partir de inferências probabilísticas. Negar isso é desconhecer a essência dos sistemas de redes neurais profundas.

No entanto, os modelos da IA generativa mostraram uma nova realidade dos dados, que

²⁸⁴ Open Source AI Definition 1.0, disponível em: <https://opensource.org/ai>. Acesso em 28 out.2024

²⁸⁵ As questões mais controversas envolvem as diferentes licenças sobre dados e os problemas em termos de direitos autorais analisados na subseção seguinte.

passaram a contemplar o emprego de dados sintéticos, criados artificialmente pelos próprios modelos e compartilhando propriedades estatísticas dos dados humanos de origem (Hradec et al., 2022). Com isso, a inserção dos conjuntos de dados no ecossistema da IA deixa de estar restrita apenas ao papel de insumo (entrada) para o treinamento e ajuste dos modelos, tornando-se também um produto (saída), gerando economias de escala e escopo em todo o processo de criação de valor dos modelos (Tyagi, 2024) e tornando mais complexos os direitos de propriedade.

Primeiro, tornam cada vez mais improvável a proteção de dados enquanto insumo pelo direito autoral, conforme comentado na seção 5.2 deste capítulo. Segundo, porque criados a partir de dados originais de terceiros, podem violar direitos autorais (e, quando forem dados pessoais, ferir a legislação de proteção de dados, ainda que usualmente não revelem atributos pessoais). Terceiro, porque as saídas da IA são invenções geradas por IA e envolvem, por conseguinte, outros mecanismos de proteção, como patentes, inclusive pelas novas abordagens que poderão tratar os modelos em si como invenções.

A escala crescente da IA generativa demandou conjuntos de dados ainda mais extensos e de domínios variados para treinamento, acuracidade e desempenho dos modelos, habilitando maior conjunto de tarefas e capacidade de criar conteúdo e centrais para o progresso técnico da IA. A demanda por dados de treinamento, que emprega mineração de textos e dados, é atualmente atendida por uma combinação de dados gerados por humanos e gerados sinteticamente a partir de dados humanos originais, eventualmente protegidos por direitos autorais.

O acesso a amplos conjuntos de dados é uma das principais fontes de assimetria entre empresas nascentes e grandes empresas do setor e uma das principais vantagens competitivas daquelas que estruturaram ecossistemas de plataformas digitais, com acesso a fluxos contínuos de dados da ampla rede de usuários/clientes, anonimizados, que possibilitaram a estruturação de gigantescas bases proprietárias.

Essas bases, típica vantagem do pioneiro, permitiram avanços em novos modelos generativos por incumbentes, enquanto empresas nascentes que conseguiram se destacar com modelos notórios de IA enfrentaram essa barreira à entrada por meio do emprego de práticas de “raspagem” de dados de linguagem e de imagem de toda a Internet.

Muitos dos dados utilizados pelas *startups* são propriedade de terceiros e estão levando a incontáveis e variadas ações judiciais por violação de direitos autorais, como da renomada empresa que se especializou na estruturação de bases de dados, Getty Images, contra a StabilityAI, por uso indevido da sua base de imagens no treinamento do modelo Stable Difusion,

e contra a Midjourney, acusada de uso de cinco bilhões de imagens sem autorização (Oratz et al., 2023). Grandes gravadoras processam até mesmo startups pouco conhecidas, como a Suno AI, por usar gravações protegidas no treinamento de seus modelos (Carman e Metz, 2024), ou a Ghostwriter, criadora por meio de IA da canção “*Heart on My Sleeve*”, acusada de violação de direitos das músicas e vozes dos cantores Drake e The Weeknd (Shanfeld, 2023). Inúmeras ações judiciais de romancistas (Cho, 2023), roteiristas de cinema e televisão (Kilkenny e Cho, 2023) e editores como NY Times, acumulam-se sobre a OpenAI pelo uso de suas obras no treinamento do ChatGPT (Allyn, 2023)²⁸⁶.

Disputas alcançaram até modelos *open-source*, como o GitHub Copilot, com Microsoft (proprietária do GitHub) e OpenAI respondendo a ação civil pública por não citar autores de programas de código aberto usados no treinamento do modelo de geração de código de computação, com estudos comprovando que o Copilot gera resultados semelhantes aos dados de “entrada” de terceiros (Henderson et al., 2023), mas que terminou em acordo e arquivamento da ação pela importância do repositório no compartilhamento de modelos na comunidade de “devs” (The Economist, 2023f)²⁸⁷.

Duas visões antagônicas permeiam essas ações nos tribunais. Ainda que possam envolver direitos legítimos de criadores, o acesso a grandes conjuntos de dados é essencial para o treinamento dos modelos. Não se trata de ferramentas digitais de cópia e plágio como sugerem muitos críticos, que acusam os inovadores da IA de explorarem obras de terceiros, mas sistemas computacionais com benefícios potenciais para a economia e para a sociedade, onde os dados são insumo crítico. A cópia ou reprodução, se ocorrer, decorre de erros de anonimização e falhas técnicas, com o uso desses dados para reconhecimento de padrões limitado ao treinamento dos modelos, não à inferência que define suas saídas²⁸⁸.

Nos Estados Unidos, esses casos começam a ser avaliados sob a doutrina do *fair use* (Lemley e Casey, 2020), chamada *fair dealing* no Canadá, enquanto na União Europeia o debate é se o uso desses dados estaria inserido entre as exceções “*text and data mining*” do direito autoral (Geiger, Frosio e Bulayenko, 2018), cuja legislação está sob revisão pelas alterações introduzidas pelo *Digital Market Act* (Ricchiuti, 2023; Margoni e Kretschmer, 2022).

²⁸⁶ O ChatGPT também enfrenta investigação por violação da legislação de proteção de dados (GDPR, sigla de *General Data Protection Regulation*) na Itália, França, Irlanda, Polônia e Espanha, com uso proibido em alguns.

²⁸⁷ Dados pessoais obtidos da Internet também estão sujeitos às leis de privacidade de dados, que não serão aqui discutidas por fugirem do escopo da propriedade intelectual.

²⁸⁸ A avaliação é feita por meio de testes de violação que mostre similaridade, que se tornam difíceis no caso dos dados sintéticos que se distanciam dos dados de origem. Tyagi (2024) observa a importância do AI Act (artigo 53 (1) (d)) que forneceu uma salvaguarda ao estabelecer a obrigatoriedade dos provedores de modelos GPAI elaborarem e disponibilizarem publicamente resumo suficientemente detalhado sobre o conteúdo usado para o treinamento do modelo.

De acordo com a doutrina do *fair use*, o uso de obras protegidas por direito autoral é considerado justo quando: i) serve a um propósito social valioso; ii) o material de origem é transformado a partir do original; e iii) não afeta o mercado principal dos detentores de direitos autorais. Em outras palavras, a racionalidade do *fair use* (da mesma forma que as exceções do direito autoral) é o entendimento de que direitos proprietários não podem se sobrepor ao benefício maior para a sociedade (Lorenz, Perset e Berryhill, 2023).

Equiparar modelos generativos à reprodução não autorizada de obras de terceiros, como o famoso caso do Google Books, mesmo este agraciado pela doutrina do uso justo, demonstra desconhecimento do desenvolvimento e funcionamento dos modelos generativos. Considerar os *tokens* (palavras ou pedaços de palavras em modelos de linguagem ou *pixels* em modelos de imagem)²⁸⁹, convertidos em representações numéricas vetoriais e devidamente codificados, como uma reprodução do material de origem para saída dos modelos e parte da obra protegida confirma desinformação sobre as tecnologias de IA e do que constitui uma obra passível de proteção pelo direito autoral.

O que surpreende, entretanto, é que somente agora, com a IA generativa, surja o questionamento sobre o uso de dados no treinamento dos modelos, que remontam a IA preditiva. Há vinte e cinco anos, o primeiro pedido de patente de Sergey Brin, cofundador do Google, já mencionava a intenção de extração de informações semânticas de bases de dados da Web, com o título “*Extracting patterns and relationships from ‘scattered’ databases such as the World Wide Web*”²⁹⁰ (alterado para “*Extracting information from a database*” na patente concedida US6678681B1).

Provavelmente, a declaração imprudente da OpenAI, que afirmou ter usado dados de toda a Internet no treinamento do ChatGPT, instigou esse tipo de indagação, despertando interesses pecuniários diversos, por grandes produtores, editoras e empresas especializadas na construção de bancos de dados (como Getty Images, Scale AI, Contextual AI, Snowflake, Weaviate, PineCone, Neon), um mercado crescente e lucrativo (The Economist, 2023a).

O cenário é incerto e controvertido, inclusive ocarionando propostas de mudanças legislativas nacionais. Na França, a *Proposition de Loi* nº 1630/2023²⁹¹ altera o código de propriedade intelectual francês, transformado o direito autoral numa espécie de bens de clube

²⁸⁹ *Tokens* são partes de texto que são alimentadas em LLMs para processamento. Uma regra prática razoável é que um *token* corresponde a 0,8 palavras em inglês, em média. No caso de imagens e vídeo, supõe-se que uma imagem e um segundo de vídeo correspondam a cerca de 30 *tokens* de texto (Villalobos et al., 2024).

²⁹⁰ Disponível em: <https://www.seobythesea.com/Extracting-Patterns-and-Relations-from-Scattered-Databases-Such-as-the-World-Wide-Web.pdf>. Acesso em: 05 ago.2021

²⁹¹ Disponível em: https://www.assemblee-nationale.fr/dyn/16/textes/l16b1630_proposition-loi.pdf. Acesso em: 06 dez.2023

(Ostrom e Hess, 2007). Nas obras criadas por IA, sem intervenção humana direta, direitos de propriedade pertenceriam aos detentores de direitos sobre os dados empregados no treinamento dos modelos e a inserção de seus nomes como autores, com a gestão dos direitos exercida por sociedades de autores ou outras organizações de gestão coletiva. Nas obras geradas por IA a partir de obras cuja origem permanece incerta, a gestão dos direitos caberia ao órgão de gestão coletiva, a ser criado e custeado por meio de taxa vinculada, garantindo “remuneração justa e equitativa pela sua exploração”, com vistas a incentivar a inovação e promoção da diversidade artística, lembrando a proposta de Senftleben (2023). A Itália, no projeto de lei de regulação nacional da IA, também impõe emenda à lei de direitos autorais (Cooper e Somaini, 2024), enquanto no Japão prevalece o entendimento de que a lei em vigor, de 2018, permite que “devs.” de modelos usem obras protegidas, mesmo com fins comerciais (The Batch, 2023).

5.4.4. Impactos sobre o funcionamento do sistema de PI

A proliferação da IA generativa pelo mundo começa a levantar preocupações sobre possíveis impactos sobre o sistema de PI. Documento não oficial de um grupo do escritório australiano, disponibilizado em uma rede social, vislumbra cenários decorrentes da projeção de crescimento explosivo de pedidos de patentes (IPA Venture Team/IP Australia 2023).

De acordo com o documento, o sistema de patentes deverá enfrentar rupturas com a mudança de paradigma da IA generativa, que estaria transformando a “invenção com ferramentas” em uma “ferramenta de invenção”. O sistema de patentes, forjado para incentivar invenções humanas deverá sofrer pressões ao se deparar com conteúdo e invenções geradas por IA. Inventores humanos passariam a conviver com conteúdo inventivo criado por IA, com efeitos sobre a relevância dos atuais requisitos de patenteamento, em especial novidade e atividade inventiva, desafiando repensar o que é permitido proteger, os limites de inventividade, a duração e o escopo da proteção pelo sistema atual.

Uma potencial enxurrada de pedidos de patentes congestionaria o sistema, criando gargalos no exame de pedidos com detalhes técnicos complexos e extensos, listas crescentes de técnicas anteriores, inclusive com pedidos redigidos por ferramentas especializadas de IA generativa (como PatentPal, Legal Robot, AllPriorArt), acelerando a criação de conteúdo publicado que se tornaria parte do estado da arte. Isso poderia levar a tentativas de tirar proveito do sistema, com expansão de mecanismos de *patent ticket*²⁹², *patent trolls*²⁹³ ou *defensive publication* (ou *defensive disclosure*), com a publicação de informações para estabelecer o

²⁹² A respeito ver <https://www.upcounsel.com/patent-thicket-example>. Acesso em: 10 set. 2023.

²⁹³ Sobre o conceito: <https://www.britannica.com/topic/defensive-patenting>. Acesso em: 10 set. 2024

estado da técnica e o requisito de novidade. A consequência seria a inundação do sistema de patentes, criando um ambiente desafiador, com oposições recorrentes e litígios, ampliando a incerteza do sistema. O impacto sobre o estado da técnica anterior poderia reduzir o incentivo à divulgação de conhecimento através do sistema de patentes, que passariam a ser mantidas por *trade secrets*, com implicações sobre a disseminação do conhecimento para a sociedade.

Os autores sugerem a revisão dos procedimentos tradicionais de duas formas. Por meio de intervenção regulatória/legislativa, proibindo especificações longas, limitando o número de depósitos por determinado período e filtrando conteúdo gerado em massa por IA, além de um requisito de uso ou licenciamento obrigatório (similar às marcas). Outro mecanismo, preferencial pelo tempo exigido para mudanças formais, seria por intervenção tecnológica, bloqueando e inserindo marcas d'água digitais imperceptíveis em imagens, áudios, vídeos, etc., criados por IA.

Seria profícuo considerar alternativas como a proposta de Ramalho (2018) de criar um direito de PI separado, um direito conexo, ou *sui generis*, um “*disemination right*” para criações da IA e, também, para invenções da IA, com escopo e duração compatíveis com o ritmo acelerado dessa tecnologia, com requisitos de divulgação que melhorem a transparência dos produtos e abertura dos insumos da IA. Com isso, evitaria que inovadores humanos se vejam competindo com máquinas pela proteção da PI.

5.5. Considerações Finais

Neste capítulo foi analisado o regime jurídico de PI relacionado a IA de redes neurais profundas, a partir das regras de TRIPS, legislações nacionais/regionais e diretrizes de exame de patentes dos principais ENPIs (USPTO, EPO e CNIPA), além de alguma literatura especializada.

O regime legal de PI da IA é extenso e intrincado, sobrepondo e maximizando direitos que protegem bases de dados e invenções da IA, por patentes, direito autoral, marcas e outros instrumentos, frequentemente usados em estratégias defensivas, instrumento de barganha, proteção da parte do conhecimento na inovação aberta e valorização de ativos, mas que convive harmoniosamente com o código aberto, ainda que importantes detalhes técnicos dos modelos sejam mantidos em sigilo. Isso explica o patenteamento agressivo paralelamente à manutenção de esquemas de código aberto, como também apontado por autores como Calvin e Leung (2020) e Cappellen (2021). Contudo, ainda há muita incerteza sobre como a patenteabilidade da IA, o código aberto e o litígio por infração evoluirão no futuro.

No seu conjunto, parece claro que o complexo e híbrido regime de PI das invenções da

IA sobrepuja e amplia direitos, sem atender de forma adequada a todos os requisitos legais, em especial de patentes, tornadas potencialmente incertas, com probabilidade significativa de serem anuladas nos tribunais, fazendo com que as firmas passem a depender cada vez mais de outros instrumentos estratégicos.

Tal regime legal da IA herdou algumas das peculiaridades do software convencional, que acabou acomodadas por meio de um duplo mecanismo de proteção, que contempla em especial direitos autorais sobre *BigData* (e direitos *sui generis* em algumas jurisdições) e direitos de patentes sobre as chamadas invenções implementadas por computador, “produtos” dos sistemas de IA, desde que tenham claro efeito técnico. O exame dos pedidos de patentes, baseado em exemplos, caso a caso, sem uma regra geral clara, pode levar a tratamentos desiguais não por diferenças intrínsecas nas invenções, mas pela forma como os pedidos foram apresentados, pelos detalhes nas diretrizes de exame ou mesmo pela interpretação dos tribunais.

As modificações “não explicáveis” dos modelos da IA ao longo do processo de treinamento com amplos conjuntos de dados, que melhoram seu desempenho e precisão técnica, conferem algum grau de autonomia relativamente à programação original, inviabilizando um atendimento cartesiano aos requisitos de patenteamento.

As firmas têm recorrido a patentes de invenção que crescem exponencialmente desde a década passada, contemplando modelos de redes neurais profundas, sistemas biométricos de reconhecimento facial sob escrutínio pela regulação da IA e até conjuntos de dados de treinamento responsáveis pelo desempenho crescente dos modelos e fonte das invenções da IA.

Os amplos conjuntos de dados de treinamento dos sistemas de IA, que diferem radicalmente dos bancos de dados de software tradicional, cada vez mais compostos por gigantescos conjuntos de dados cuja análise transcende a capacidade humana, sendo organizados e analisados pelos próprios sistemas de IA e *BigData*, com isso tropeçando no requisito de originalidade do direito autoral, agraciados com direitos *sui generis* em algumas jurisdições, enquanto em outras patenteados como se fossem a invenção em si. Restrições técnicas digitais são amplamente utilizadas pelas firmas para limitar o acesso aos dados e constituem uma proteção eficaz contra o uso não autorizado e cópia por terceiros, maximizando a proteção.

Nas patentes, os requisitos legais de inventividade (e autoria) e suficiência descritiva são desafios no caso da IA. A descrição detalhada da invenção no ato do depósito (suficiência descritiva), que seria a contrapartida para a sociedade em troca do privilégio ao inventor, parece simplesmente impossível de ser atendida no atual estado da arte da IA, onde o funcionamento interno não é explicável, visão que converge com Ebrahim (2021). Esses

sistemas de IA baseados em modelos matemáticos/estatísticos e algoritmos, de natureza abstrata, se tornaram elegíveis para patentes se a invenção tiver um efeito técnico adicional. A explicabilidade e abertura da caixa preta da IA, atualmente tão caros para sociedade, não parecem relevantes no exame de pedidos de patentes de IA.

O requisito da novidade, aparentemente menos controvertido, é potencialmente problemático se sistemas de IA começarem a liderar o avanço do conhecimento e balizar o estado da arte que deve ser ultrapassado por inovações subsequentes, particularmente preocupante se sistemas de IA passarem a ser a fonte de invenções em outros setores da economia. Por outro lado, a necessidade de demonstrar aplicação prática e produção demonstrável, a partir dos usos setoriais citados pelo inventor na reivindicação acabam sendo hoje uma informação incompleta, pois pela natureza de tecnologia de propósito geral/habilitadora, nem todas as aplicações possíveis são conhecidas *ex-ante* e dependerão de extensa experimentação e desenvolvimento de tecnologias complementares. E por um prazo de 20 anos, para uma tecnologia que avança em ritmo tão veloz.

Os algoritmos são incluídos entre os assuntos não elegíveis para proteção de DPI, uma vez que são conceitos abstratos como teorias científicas ou métodos matemáticos. No caso de algoritmos dos sistemas de IA essa afirmação parece ainda mais verdadeira, com sistemas de IA de caráter abstrato. As ConvNets têm sido consideradas solução técnica elegível para patentes, mesmo que pareçam mais próximas do conhecimento científico puro, o que parece mais verdadeiro no caso das novas arquiteturas de redes neurais da IA generativa, criando conteúdo diverso e dados sintéticos de forma cada vez mais inescrutável. Até bases de dados de treinamento de modelos de aprendizado autossupervisionado e por reforço têm sido consideradas elegíveis para patentes. Segundo Hilty, Hoffmann e Scheuerer (2020:27) “... *most AI applications lack a theoretical justification for creating exclusive rights. If this fact is ignored, such legislation could ultimately lead to dysfunctional effects that have negative impacts on social welfare.*”

A análise do caso DABUS mostrou um dos pontos mais polêmicos da IA. Mesmo sem exigência explícita nas leis de patentes do inventor ser pessoa física, a recusa se apoia em outros *corpus* jurídicos e na abordagem teórica do *mainstream* de que invenções precisam de incentivos legais, sem qualquer avaliação de mérito sobre a real capacidade de sistemas de IA poderem criar autonomamente invenções. A autonomia dos sistemas de IA e a racionalidade econômica do sistema de patentes (que antes de estimular a inovação e remunerar o inventor, podem apenas proteger valiosos ativos empresariais estratégicos, usadas como moeda de barganha em trocas de tecnologia ou proteger uma parte diminuta do conhecimento exigido

para uma invenção) parecem definitivamente conflitantes, no presente e no futuro próximo. Decisões baseadas em interpretações subjetivas certamente abrem brechas para reinterpretações e contestações futuras dessas patentes pelos tribunais, que poderão até concluir que uma invenção é de domínio público se não houver um inventor humano.

Não obstante, a legislação de PI parece imune às peculiaridades e complexidades das invenções da IA, não produzindo mudanças, com exceção da inclusão de um artigo na Lei Chinesa de Patentes relacionado aos algoritmos. Resolução do Parlamento Europeu, sobre direitos de PI em IA ([P9_TA\(2020\)0277](#)) apenas mencionou que deveria recepcionar as mudanças na legislação antitruste e regulatória da economia digital, BigData e IA.

O tratamento de questões controvertias das invenções da IA vem sendo promovido por simples diretrizes de exame, meros atos administrativos que não constituem regulamentação substantiva, sem força e efeito de lei, mesmo que abordem situações que não se restringem a procedimentos formais, mas envolvem questões profundas da própria racionalidade do sistema de PI. Nesses casos, deveriam ser objeto de leis nacionais ou, preferencialmente, tratados e convenções internacionais, como o Tratado de Direito de Patentes (PLT), adotado em 2000 com vistas a harmonizar e agilizar procedimentos formais com relação a pedidos de patentes nacionais e regionais, algo urgente no caso das patentes de IA.

No entanto, as leis existem não como um fim em si mesmo, mas como meios para definir as "regras do jogo". Regras que deveriam ser mudadas quando o jogo é outro. O risco de reinterpretação de decisões pelos tribunais cria incertezas jurídicas e problemas de *enforcement*, enfraquecendo o extenso regime de PI da IA. Tais incertezas parecem reforçar o *open-source* em IA - com detalhes técnicos dos modelos mantidos em sigilo - e uma nova geração de plataformas digitais que fornece serviços na nuvem (IaaS, AIaaS, DaaS etc.), por meio de assinaturas regidas por instrumentos contratuais, de natureza mais fluida e não padronizada, com possíveis impactos sobre a proteção com base em PI. É necessário, contudo, análises mais aprofundadas sobre esses modelos e possíveis desdobramentos sobre DPI, relacionados às invenções e bancos de dados da IA, principalmente dados sintéticos, o que abre um campo profícuo para pesquisas futuras.

CAPÍTULO 6 – ALGUMAS EVIDÊNCIAS DA IA

6.1. Evolução dos modelos de IA e seus principais protagonistas

Acompanhar a evolução da inovação da IA a partir do conjunto das BigTechs americanas se tornou a prática usual. No entanto, há diferenças importantes entre esse grupo de empresas e entre elas e as empresas emergentes que se destacaram com a IA generativa. As BigTechs são tratadas como um bloco uníssono em virtude de empregarem plataformas digitais que empregam, sob a lógica do “capitalismo de plataforma” (Srnicek, 2017) ou “capitalismo de vigilância” (Zuboff, 2019). São empresas que apresentam, cada vez mais, estratégias diferenciadas, ocupam segmentos de mercados distintos e possuem dinâmicas particulares, fenômeno até agora pouco percebido pela literatura.

O capítulo apresenta amplas evidências da produção em IA e seus principais responsáveis, elaboradas a partir da ampla e atualizada base de dados aberta do instituto de pesquisa multidisciplinar Epoch²⁹⁴. A base permite identificar um número impressionante de mais de mil modelos de aprendizado de máquina desenvolvidos desde a década de 1950 até os dias atuais. Abarca desde modelos preditivos e de classificação até os LLMs, de vídeo, robótica, multimodais, entre outros surgidos mais recentemente. Construída a partir de extenso levantamento de dados atualizados sobre os modelos, discriminados por nome, versões, domínios, desenvolvedores, parcerias, referências bibliográficas e detalhes técnicos importantes, tais como número de parâmetros, quantidade de dados de treinamento e capacidade computacional empregada, entre outros, mesmo que não disponíveis para a integralidade dos modelos.

É formada com dados da base construída pela própria Epoch²⁹⁵, complementada pela base de publicações científicas da OpenAlex²⁹⁶ e das principais inovações algorítmicas²⁹⁷ do *survey* realizado por Zao et al. (2023), posteriormente atualizadas pela organização de pesquisa. As bases são analisadas conjuntamente com vistas a classificar os modelos conforme sua notabilidade, quais sejam, aqueles com maior representatividade em três métricas: alta contagem de citações, melhoria no “estado da arte” e importância histórica.

Esses dados amplos são tratados de forma detalhada e rigorosa, possibilitando uma visão praticamente completa sobre os modelos de IA baseados em aprendizado de máquina e sua evolução ao longo do tempo, originando estudos de alta qualidade e originalidade.

²⁹⁴ Disponível em: <https://epochai.org/>. Último acesso em: 14 mar.2024

²⁹⁵ Disponível em: <https://epochai.org/data>. Último acesso em: 14 mar.2024

²⁹⁶ Disponível em: <https://openalex.org/>. Último acesso em: 14 mar.2024

²⁹⁷ Disponível em: https://docs.google.com/spreadsheets/d/1L_j7OaX19HXWWIx_apKvWo2OteY1XOB7FamaLEd_p0s/edit#gid=1298304064. Último acesso em: 14 mar.2024

Consequentemente, possibilita uma visão minuciosa sobre os principais inovadores em IA, os modelos que se tornaram referências acadêmicas e representaram avanço no estado da técnica.

A análise dos dados mostra que a partir de meados da década passada a importância da academia no desenvolvimento dos modelos no mundo foi suplantada pela indústria, que se torna inequívoca na década atual com a IA generativa e rompendo as fronteiras das empresas gigantes do setor. Não obstante, à medida que os modelos aumentam em escala, em termos de diferentes métricas (número de parâmetros, quantidade dos dados de treinamento e capacidade computacional empregada), o protagonismo é indiscutivelmente das maiores empresas e umas poucas emergentes, como será visto. O somatório de modelos por desenvolvedor – isoladamente ou em parceria com outras organizações – permite identificar as empresas que se destacam e mostra a assimetria entre as *BigTechs* entre si e entre elas e as empresas emergentes.

Muitos modelos foram criados em parceria de empresas e instituições acadêmicas (ou das empresas entre si), embora a importância da pesquisa colaborativa divirja por empresa, principalmente em virtude da notabilidade do modelo. Para uma amostra de 896 modelos extraídos da base completa da Epoch, cujo conjunto de informações é mais completo, quase 40% foram desenvolvidos em parceria, constatado com base na afiliação dos autores dos artigos de divulgação dos modelos. Esta proporção varia, entretanto, para cada *BigTech*, com destaque para Microsoft, onde 52% dos modelos foram desenvolvidos em colaboração com outras organizações; 42% dos modelos da Meta AI e 31% do Google (incluindo sua subsidiária DeepMind). A pesquisa colaborativa é menos frequente entre as empresas emergentes - apenas 12% dos modelos da OpenAI; 11% da Stability AI e nenhum da Anthropic, InflectionAI, MistralAI e AI21 Labs, entre outras.

A importância da pesquisa colaborativa é menor quando considerados somente os modelos notáveis, predominantemente desenvolvidos por meio de pesquisa “isolada” das empresas, com menor incidência de autores com afiliações diversas.

As universidades chinesas são férteis na invenção de modelos de IA, em especial a Tsinghua University, com 28 modelos, muitos em parceria, com *BigTechs* chinesas (Alibaba e Huawei) e também americanas (Microsoft, Meta e Nvidia). A organização chinesa de pesquisa BAAI também se destaca, com a criação de 18 modelos. Google/DeepMind tem uma proporção relativamente menor de pesquisa colaborativa com instituições chinesas. Seus principais parceiros são a Carnegie Mellon University (CMU), o University College London (UCL) e Stanford University, sem esquecer das canadenses Université de Montréal e University of Toronto – que tem como acadêmico o ex-cientista chefe do Google Brain, Geoffrey Hinton.

Foi selecionado um subconjunto dos modelos lançados pelas empresas a partir de 2010,

subdividido em dois subperíodos com vistas a identificar as características dos modelos generativos e inovadores pioneiros: um primeiro subperíodo, de 2010 a 2016, majoritariamente composto por modelos da IA preditiva (embora incluam uns poucos modelos generativos, como baseados em arquitetura GAN, desenvolvida em 2014), e um segundo subperíodo, de 2017 até os dias atuais, que captura a onda de modelos da IA generativa. A escolha do início deste segundo subperíodo foi devida a ser o ano de publicação do artigo seminal dos Transformers, divisor de águas da IA generativa²⁹⁸.

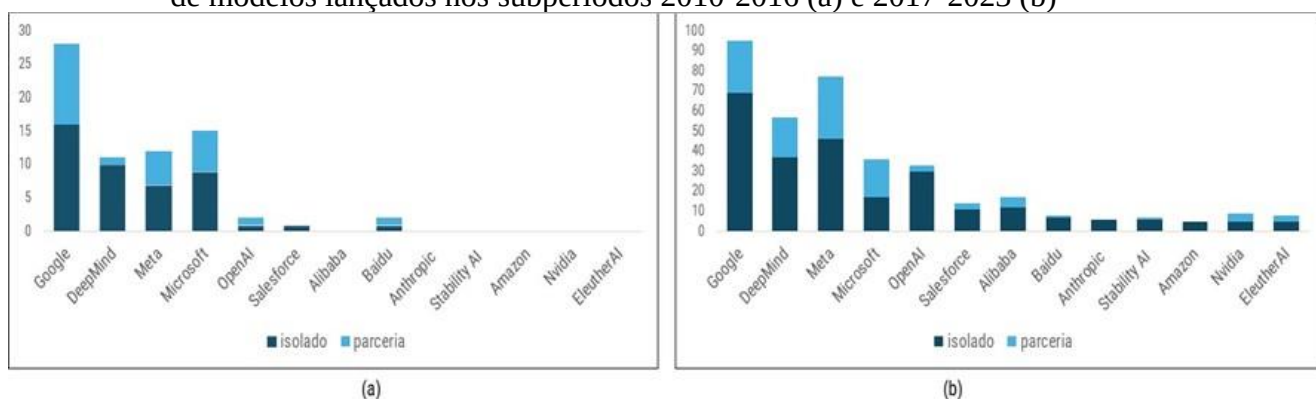
No nível da indústria, o Gráfico 6.1.1 (a) mostra o claro protagonismo do Google no primeiro subperíodo, seguido pela subsidiária DeepMind – aqui discriminada visando mostrar sua importância para o grupo Alphabet –, Microsoft, Meta e OpenAI (empresa criada em 2015, que já se destacava no primeiro subperíodo junto as *BigTechs* fundadas em 1998, 2010, 1975 e 2004, respectivamente). A liderança do Google se mantém no segundo subperíodo, mesmo com a expansão notável da Meta, que supera Microsoft, e da OpenAI. Muitos modelos foram desenvolvidos em parceria, principalmente nos casos da Meta e Microsoft, como pode ser observado no Gráfico 6.1.1 (b), com menor frequência no caso do Google, DeepMind e OpenAI. A menor participação de modelos em parceria pela DeepMind no primeiro subperíodo é comparável a OpenAI no segundo subperíodo, reforçando a hipótese da maior relevância das parcerias à medida que a empresa amadurece. O gráfico incorpora novos protagonistas na corrida da inovação da IA generativa no segundo subperíodo, como Salesforce, as chinesas Alibaba e Baidu, Anthropic, Stability AI, Nvidia e a organização de pesquisa EleutherAI²⁹⁹. Muitas empresas emergentes estão desenvolvendo modelos de IA generativa, mas apenas um ou, no máximo, dois modelos cada uma.

Cottier, Besiroglu e Owen (2023) identificaram como líderes da IA, o Google, a OpenAI e a Meta, que desenvolveram fração expressiva das invenções algorítmicas que sustentam os LLMs, além de responsáveis pelas maiores execuções de treinamento e liderando métricas de impacto de citação. Ainda que as instituições acadêmicas chinesas tenham produzido o maior número de artigos de IA na última década (Maslej et al., 2023), os autores concluem que a liderança científica chinesa não se refletiu em performance empresarial, com superioridade das gigantes americanas comparativamente às rivais chinesas Tencent e Baidu.

²⁹⁸ A data do lançamento de um modelo toma como *proxy* a data de publicação do artigo mais antigo (incluindo pré-impressão) de introdução da respectiva inovação algorítmica, disponibilizada na base de dados da Epoch (na maioria são artigos publicados na arxiv.org).

²⁹⁹ A HuggingFace seria um outro caso a destacar, mas não foi incluída no gráfico porque seus modelos são lançados predominantemente em parcerias (amplas) e não isoladamente – métrica usada para a seleção dos principais *players*.

Gráfico 6.1.1– Principais empresas desenvolvedoras de modelos de IA, em termos do número de modelos lançados nos subperíodos 2010-2016 (a) e 2017-2023 (b)



Fonte: Elaborado com base nos dados da [Epoch Database](#)

Nota: As 13 empresas discriminadas nos gráficos foram aquelas que lançaram o maior número de modelos de forma isolada no período completo, sem incluir os dados de 2024 por não representar um ano fechado

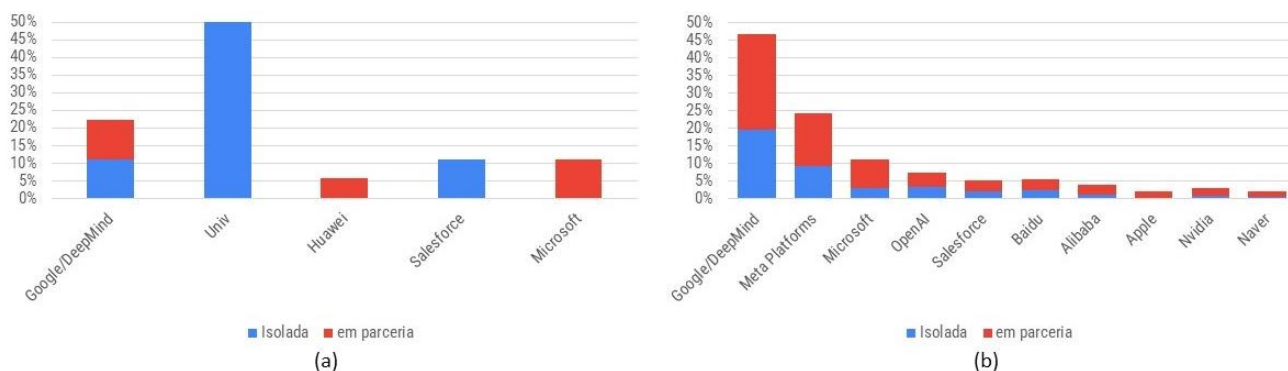
A Microsoft, apesar de liderar a produção científica corporativa (com maior número de publicações) e até pouco tempo também as patentes de IA, dava indícios de defasar no avanço do estado da arte em IA (SOTA, sigla de *state of the art*) em termos de inovações algorítmicas e lançamento de modelos, o que parece explicar seu recente recurso a parcerias. De fato, seus LLMs lançados de forma isolada, em 2023, Phi-1,5 e Phi-2 (Gunasekar et al., 2023), e a nova versão de 2024 (Phi-3 14B) (Abdin et al., 2024) não representaram avanço no estado da arte, segundo a classificação da Epoch, sugerindo que os modelos com notabilidade SOTA da Microsoft ficaram no passado, exceção para o modelo multimodal MetaLM (2022). Modelos recentes em parceria evidenciam uma reorientação estratégica, quando parece virar o jogo. Primeiro, por meio da aquisição do GitHub, principal repositório *open-source*, tentativas de aquisição do Discord, aquisição da Semantic Machines e, finalmente, a parceria com a OpenAI, iniciada em 2018. A voracidade da Microsoft não para por aí, incluindo recentemente Mistral AI, Inflection AI, além da internacionalização de infraestrutura na nuvem.

O Gráfico 6.1.2 (a) e (b) evidencia as empresas que desenvolveram o maior número de modelos SOTA, de acordo com a Epoch, nos dois subperíodos, medidos pela representatividade desses modelos em relação ao total de modelos lançados. No primeiro subperíodo, os modelos SOTA eram desenvolvidos predominantemente pela academia, com destaque para poucas empresas (a), enquanto no segundo subperíodo as empresas predominam, com claro destaque para Google/DeepMind, seguido pela Meta.

No Gráfico 6.1.3 (a) e (b) é mostrado o percentual dos modelos mais citados em relação ao total, por organização, nos dois subperíodos, evidenciando uma perda de importância relativa das universidades a partir de 2017. A comparação entre os dois gráficos mostra que apesar das universidades deixarem de responder pelo avanço no estado-da-arte da IA (Gráfico 6.1.2), que passou para as empresas, elas ainda continuam responsáveis pelos modelos mais citados

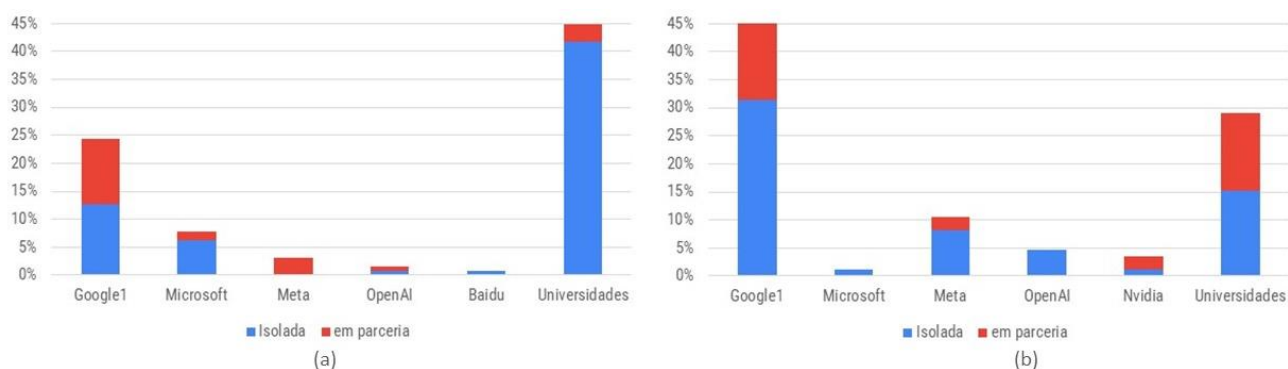
(Gráfico 6.1.3) depois do Google/DeepMind, empresa cuja missão parece se confundir com a da academia³⁰⁰.

Gráfico 6.1.2 – Principais empresas desenvolvedoras de modelos de IA que representaram melhoria no estado da arte (SOTA improvement), em termos de percentual no total de modelos lançados nos subperíodos (a) 2010-2016 e (b) 2017-2023



Fonte: Elaborado com base nos dados da Epoch Database

Gráfico 6.1.3 – Principais empresas desenvolvedoras de modelos de IA altamente citados, em termos de percentual no total de modelos lançados nos subperíodos (a) 2010-2016 e (b) 2017-2023



Fonte: Elaborado com base nos dados da Epoch Database

Nota: ¹ Inclui DeepMind

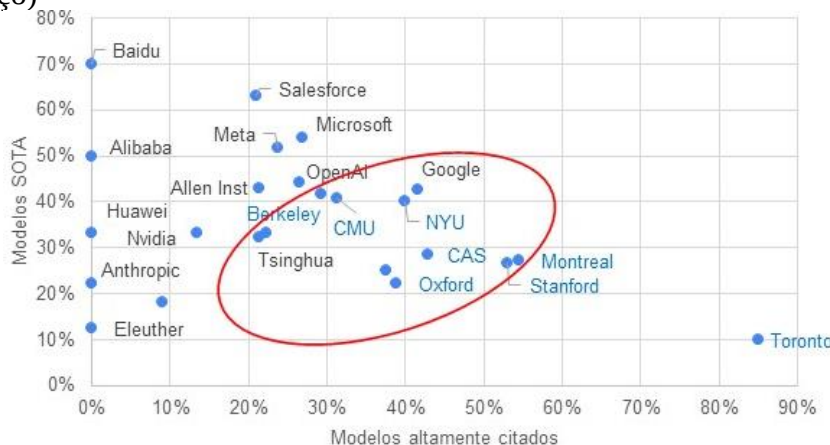
O Gráfico 6.1.4 mostra as organizações que contemplam os melhores modelos em melhoria no estado da arte e maiores citações, uma medida de qualidade dos modelos. É interessante notar que o Google definitivamente desempenha papel típico das universidades, caminho que parece começar a ser seguido pela OpenAI³⁰¹, considerando a posição dessas duas empresas no gráfico similar às universidades.

³⁰⁰ Embora o maior número de publicações seja da Microsoft, os artigos mais citados são do Google, Meta e DeepMind (Cottier, Besiroglu e Owen, 2023). Em 2022, a DeepMind foi responsável por 12% dos artigos de pesquisa de IA mais citados, só superada por Meta e Google em impacto de pesquisa, de acordo com a startup de pesquisa de IA Zeta Alpha. Disponível em: <https://sifted.eu/articles/deepmind-talent>. Acesso em 03 jan.2024

³⁰¹ Dados da WIPO (2024a) mostram que Alphabet/Google é a única empresa no ranking das 20 principais organizações em publicações científicas de IA generativa entre 2014 e 2023, atrás apenas da Academia Chinesa de Ciências e das universidades Tsinghua e Stanford. É líder no ranking de citações de publicações, com quase o dobro da segunda colocada (Berkeley). Por este indicador outras empresas – Meta, DeepMind, Nvidia, OpenAI, Microsoft e Twitter – também entram no ranking das 20 organizações líderes.

Surpreende a ausência da Apple entre as líderes no lançamento de modelos, o que poderia sugerir seu atraso na corrida da IA generativa frente às demais *BigTechs*. De fato, uma estratégia refratária à divulgação e abertura dos modelos sempre caracterizou a Apple (Stone e Vance, 2009), que parece só mudar recentemente como forma de atrair e reter talentos. A empresa anunciou em 2023 estar em fase de testes com seu *chatbot*, além de desenvolver no mesmo ano o Ferret-13B multimodal, em parceria com a Columbia University, em código aberto no GitHub (You et al., 2023), bem como do MM1 (McKinzie et al., 2024).

Gráfico 6.1.4 – Medida de qualidade dos modelos desenvolvidos pelas principais organizações (isoladamente e em parcerias) pela percentagem de modelos que representaram melhoria no estado da arte (SOTA) e de modelos mais citados, em termos de percentagem do total, em 2010-2024 (até março)



Fonte: Elaborado com base nos dados da Epoch Database

Embora pareça ter demorado a avançar na IA generativa, a Apple poderá surpreender ao incorporar modelos de código aberto para rodar em seus dispositivos e na atualização do *chatbot* Siri. A linha de pesquisa da empresa parece sugerir uma abordagem mais pragmática e comercial, pela busca de arquiteturas de modelos compatíveis com telefonia móvel, principal carro-chefe da empresa. Uma dessas novas técnicas usa memória flash para armazenar dados do modelo, o que permitirá o avanço na execução de LLMs nos iPhones, que hoje esbarram nas limitações dos dispositivos móveis por exigirem robusto poder computacional e de memória. A técnica permite que modelos de IA funcionem até o dobro do tamanho da memória disponível do iPhone, representando aumento de 4 a 5 vezes na velocidade nos processadores padrão. Isso possibilitará executar LLMs avançados em qualquer de seus dispositivos - iPhone, iPad, MacBook ou mesmo, no futuro, no smartwatch Apple Watch (Alizadeh et al., 2024), viabilizando um modelo de negócios que alcance usuários finais.

A discreta estratégia da Apple parece focar na busca de soluções técnicas para alcançar um modelo de negócios viável no contexto do seu ecossistema, por meio de parcerias visando incorporar aos seus aplicativos modelos generativos protagonistas, como as muitas negociações

parecem indicar - com Meta AI, Google, OpenAI, Anthropic, Perplexity (empresa fundada por ex-pesquisadores da OpenAI, Meta e Quora), que integrarão no iPhone e Macbooks seu novo produto *Apple Intelligence*, anunciado em junho de 2024. É parte das negociações a participação, junto com a Microsoft, Nvidia e Thrive Capital, na nova rodada de financiamento da OpenAI, que elevará o valor de mercado da OpenAI a US\$ 100 bilhões (Peters, 2024).

Outra ausência é da Amazon, que apesar de desenvolver produtos (como o Rufus), parece enveredar por uma estratégia distinta do Google, OpenAI e Meta AI na corrida pelo lançamento de modelos. Mais orientada para desenvolvimento de hardware, suporte a organizações *open-source* (como HuggingFace) e parcerias com as emergentes Anthropic, StabilityAI e AI21 Labs, para disponibilização desses modelos como serviço na nuvem no *Amazon Bedrock*. Tal estratégia pragmática parece igualmente focada no desenvolvimento de modelo de negócios (e condições de apropriabilidade da inovação), no qual modelos da IA serão simples *commodities*.

Algumas empresas emergentes não são meras coadjuvantes, dividindo o protagonismo no lançamento de modelos disruptivos da IA generativa com Google, DeepMind e Meta. Não somente OpenAI, mas também Anthropic, Mistral AI, AI21 Labs e outras poucas participantes ativas da corrida. Embora essas (poucas) emergentes tenham avançado em modelos que se destacam na mídia, a maioria não representou melhoria no estado da arte e nem são referências em citações para modelos subsequentes. Muitas emergentes desenvolvem modelos derivados das líderes e muitas *startups* criam meros aplicativos, o que não é o cerne do avanço tecnológico no campo da IA. É importante realçar este ponto, pois demandam recursos financeiros e capacidades muito diferentes e que distinguem a inovação da sua difusão, como mostrado na pilha de tecnologia (Figura 4.6, do Capítulo 4). Coube às *startups* explorar as possibilidades mais amplas da IA generativa e dos *chatbots* (O'Reilly, 2024), mas mesmo que a história mostre casos de inovadores incrementais que conseguiram suplantar inovadores pioneiros, é significativo o *gap* dos simples desenvolvedores de aplicativos em capacidade infraestrutural, recursos financeiros e técnicos.

Embora o lançamento acelerado de modelos – na maioria LLMs - parecesse romper as fronteiras das gigantes do setor, os últimos dois anos indicaram menor protagonismo de empresas emergentes à medida em que cresce a escala dos modelos, como no caso dos multimodais. Esses modelos maiores mostram protagonismo efetivamente assimétrico das *BigTechs*, com predominância incontestável do Google (incluindo a subsidiária DeepMind), seguida pela Meta, Amazon e as *BigTechs* chinesas, em especial Alibaba e Baidu, além da

OpenAI (graças ao suporte da Microsoft) e umas poucas emergentes (que também mantêm laços com as *BigTechs* desconhecidos do público em geral).

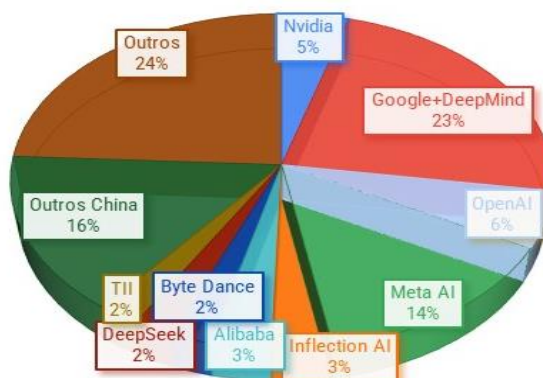
Poderosos modelos multimodais são recentes, o primeiro identificado em 2020, o Perceiver IO, desenvolvido pela DeepMind. Desde então avançam rapidamente, somando 53 modelos até 4 de março de 2024 (Epoch), passando a representar um quarto dos novos modelos desse ano, com a maioria (dois terços) ainda LLMs. Três quartos dos modelos multimodais estão concentrados em poucas organizações - onze modelos do Google/DeepMind (quatro deles versões da família Gemini), apenas um em parceria com a academia; cinco da OpenAI (versões do GPT-4 e do CLIP), quatro da Microsoft (três em colaboração com universidades, inclusive Peking University) e vinte liderados por organizações chinesas (com destaque para Alibaba, desenvolvedora ou co-desenvolvedora de nove modelos; BAAI, com seis modelos e quatro da Tsinghua), três da Anthropic, dois da Salesforce, dois da Baidu, um da Meta e as demais empresas com um modelo cada.

As empresas que dominam os modelos de fronteira da IA parecem ser aquelas que apostam nos modelos gigantes, computacionalmente intensivos. Foram identificados 94 desses modelos (até março de 2024) na base de dados da Epoch, desenvolvidos por poucas empresas, destacando-se Google/DeepMind, Meta AI, OpenAI, Nvidia³⁰², Inflection AI e Alibaba (Gráfico 6.1.5). Um percentual significativo de modelos está incluído na categoria “Outros” (24%), composto por diversas organizações pulverizadas que desenvolveram, cada uma, apenas um ou dois modelos, mas que evidenciam um interesse generalizado nos computacionalmente intensivos, mesmo por empresas de menor porte. Destaca-se também “Outros China”, por organizações chinesas, categoria que inclui muitas empresas emergentes. Em média, os modelos da IA de grande escala usam cerca de 100 vezes mais computação do que outros modelos contemporâneos de IA (OECD, 2023).

O Gráfico 6.1.6. discrimina 20 modelos mais intensivos em computação de treinamento, estimado pela Epoch, onde se destacam o Gemini Ultra, seguido pelo GPT-4 (treinado com 42% da computação utilizada para o modelo do Google), modelos chineses, o emiradense Falcon 180B e o LLama2 70B (Meta), além de modelos de empresas emergentes, unicórnios como Anthropic e Inflection AI, empresas avaliadas em US\$ 16 bilhões e 4 bilhões, respectivamente, e a própria OpenAI, avaliada em US\$ 80 bilhões, cuja capacidade financeira para arcar com pesados custos de computação pareceria improvável.

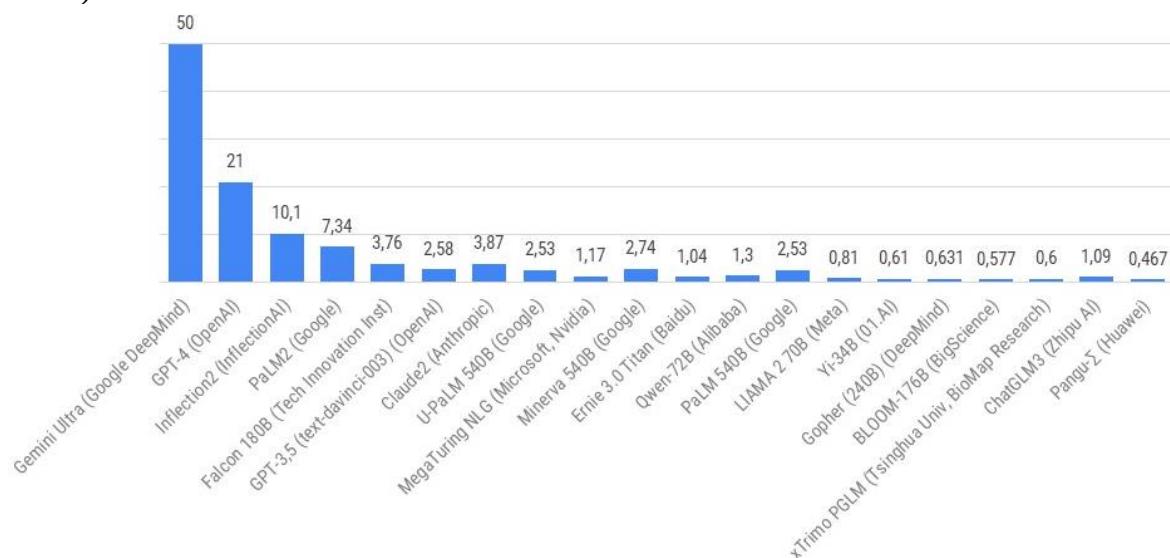
³⁰² A Nvidia desenvolveu o modelo Nemotron e também modelos em parceria com HuggingFace, Microsoft, etc.

Gráfico 6.1.5 - Principais desenvolvedores dos modelos de IA intensivos em computação (percentual do número total desses modelos)



Fonte: Elaborado a partir da base de dados Epoch e dados rastreados por Rahman, Owen e You (2024)

Gráfico 6.1.6 - Principais modelos de IA generativa em termos de computação de treinamento (estimada), em bilhões de petaFLOP (10^{15} operações de ponto flutuante estimadas na literatura de IA)



Fonte: Elaborado a partir da base de dados Epoch

A improbabilidade no mundo dos negócios muitas vezes é contornada por investimentos de terceiros, como capitalistas de risco e, cada vez mais, vínculos na casa dos bilhões de dólares dessas empresas emergentes com as gigantes do setor, por meio de alianças que se sobrepõem umas às outras, conforme exemplos mostrados na Figura 6.1.1.

Figura 6.1.1 - Alianças sobrepostas das BigTechs com empresas emergentes da IA generativa.

	Microsoft	Amazon	Google	Meta	Nvidia	Apple
OpenAI	✓	✓ ¹				✓ ³
Adept	✓				✓	
AI2I Labs		✓ ¹	✓		✓	
Anthropic		✓ ^{1,2}	✓			✓ ³
Inflection	✓				✓	
InWorld	✓			✓		
Runway			✓		✓	
Synthesia			✓		✓	
Typeface	✓		✓			
Mistral AI	✓					
HuggingFace		✓	✓		✓	

Fonte: Adaptado de CBInsight. Disponível em: <https://www.cbinsights.com/research/report/big-tech-future-charts/>. Acesso em 15 dez.2023

¹ AWS – também StabilityAI

² Também investimentos de US\$ 4 bilhões, com compromisso de uso de chips proprietários da Amazon

³ Apple Intelligence (iPhone, Macbooks etc.) – também Perplexity, LLaMA Meta, Google Gemini...

As empresas menores que se destacam em modelos generativos dependem da capacidade do núcleo infraestrutural das *BigTechs* – as americanas Amazon, Microsoft e Google e as chinesas Alibaba e Tencent – e até de incumbentes de alta tecnologia, como IBM, Oracle e Salesforce, que ofertam, de forma padronizada e automatizada, na forma de serviços, recursos de armazenamento e de rede a partir de recursos de computação de propriedade do provedor.

Isso reflete e aprofunda a clara tendência no setor, em que as grandes empresas de tecnologia com expressivos recursos de computação em nuvem alavancam ativos para alcançar uma maior posição em IA. A Microsoft parece o exemplo mais evidente, pelos vínculos com OpenAI e agora Inflection e Mistral AI, mas é quase uma regra no setor.

Muitas startups avançaram até agora com recursos de investidores de risco e das próprias *BigTechs*, em créditos de computação. Muitas tentaram avançar na construção de capacidade computacional própria – Inflection, com 22 mil H100 GPUs; Cohere, com mais de 16 mil H100 por cluster de GPUs; e Anthropic, com clusters com dezenas de milhares de GPUs (The State of AI Report Compute Index, 2024). Contudo, parecem não suportar esse desproporcional peso financeiro, como a Inflection e a Character AI, cujos “cacos” estão sendo recolhidos por Microsoft e Google, além do futuro imprevisível da StabilityAI, podendo indicar que a previsível consolidação das *startups* já pode ter começado.

As *BigTechs*, enquanto *Cloud Service Providers (CSP)*, fornecem recursos computacionais para novos *players*, além de controlar repositórios como o GitHub que suportam “devs.” ao redor do mundo. Isso evidencia a importância dos seus investimentos

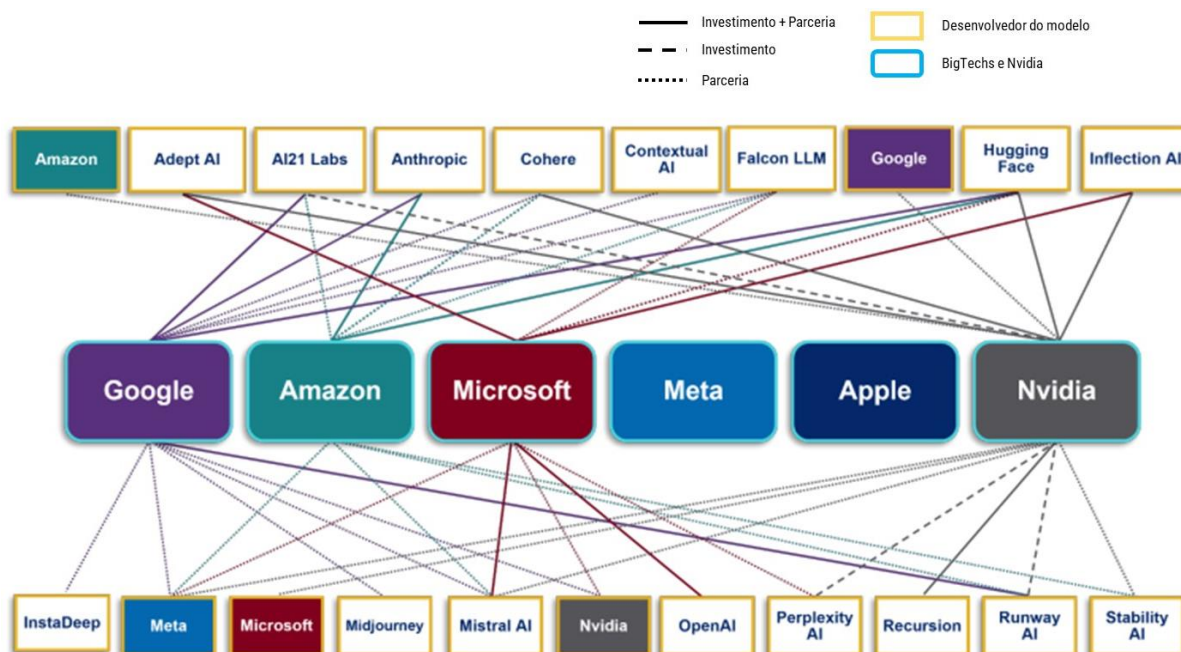
expressivos e crescentes em infraestrutura de hardware na fase da IA em que modelos são cada vez mais dependentes de rápidos avanços em hardware de computação poderoso e especializado (Khan e Mann, 2020).

Praticamente todas as *startups* e mesmo alguns laboratórios de pesquisa de IA em código aberto, operam por meio de créditos de computação, compartilhando receita e mantendo acordos contratuais com gigantes da tecnologia (McLaughlin e Gardizy, 2023). O suporte financeiro e infraestrutural das *BigTechs* às empresas emergentes inovadoras da IA generativa está criando vínculos complexos, mesmo sem participação acionária ou parcerias formais.

O assimétrico poder computacional do núcleo estrutural das *BigTechs* não significa que sejam totalmente autossuficientes. Elas mesmas são dependentes de um ecossistema global de contratantes que abarca projetistas de chips (Nvidia, AMD, etc.), empresas de semicondutores (TSMC, Qualcomm), montadoras de servidores (Supermicro, Inventec) e data centers (por exemplo, Equinix). Em especial a Nvidia, que controla a maior parte do mercado global de processadores gráficos usados em aprendizado de máquina (The Economist, 2023d) e tem como maiores clientes Amazon, Google e Microsoft. Isso reforça a noção de ecossistemas cooperativos, com interrelações e interdependências cada vez mais amplas e variadas.

A Figura 6.1.2, elaborada pela CMA, agência antitruste britânica, mapeia de forma não exaustiva as relações entre as *BigTechs* americanas e a Nvidia com empresas emergentes inovadoras de modelos de fundação. As relações envolvem parcerias e/ou investimentos, identificados em fontes públicas, dessas empresas entre si e com empresas menores, seja pelo fornecimento da computação para desenvolvimento dos modelos, seja pelo fornecimento do modelo (para distribuição no mercado pelas BigTechs/Nvidia), não incluindo subsidiárias integrais. Correspondem a teias de alianças em um sub-ecossistema emergente. Mais uma vez o caso da Apple chama atenção, não sendo encontrados documentos que mostrem sua atuação em modelos de fundação e relações com outras empresas.

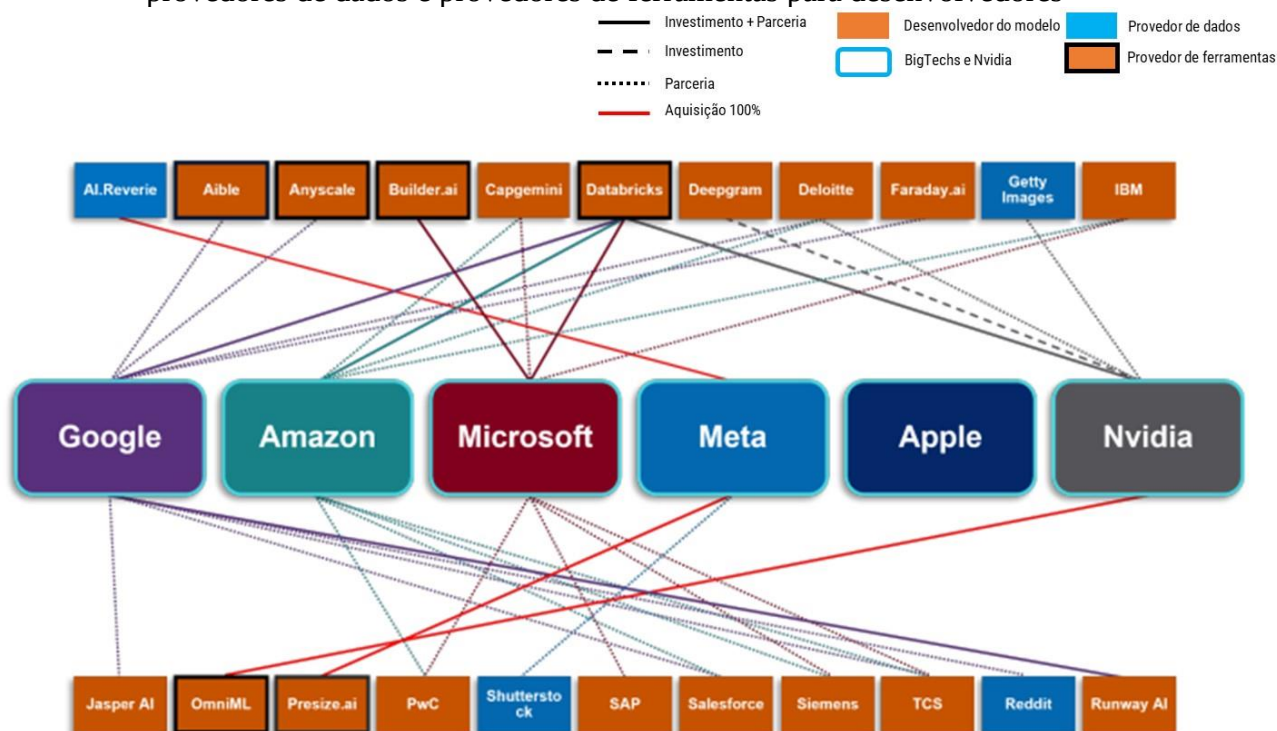
Figura 6.1.2 - Relação entre BigTechs e Nvidia e desenvolvedores de modelos de fundação



Fonte: CMA (2024:31)

A Figura 6.1.3, por seu turno, mapeia as relações do ecossistema mais amplo das *BigTechs*/Nvidia com implementadores dos modelos, provedores de serviços voltados para o cliente alimentados por modelos de fundação, fornecedores de dados ou ferramentas para desenvolvimento/implementação dos modelos.

Figura 6.1.3 - Relação entre BigTechs/Nvidia e desenvolvedores de modelos de fundação, provedores de dados e provedores de ferramentas para desenvolvedores



Fonte: CMA (2024:32)

A infraestrutura na nuvem e *gateways* fornecidos pelas líderes gigantes do setor permite que empresas menores desenvolvam modelos e produtos de IA, mas ao custo de seu aprisionamento às funcionalidades e *design* do provedor, característica típica do modelo de negócios de plataformas digitais, com as empresas líderes competindo para ser a infraestrutura fundamental para IA (Rahman e Thelen, 2019).

O código aberto foi capaz de avançar na IA generativa porque as *startups* puderam obter esse ativo complementar essencial (infraestrutura computacional) controlado pelas *BigTechs* (Widder, West, Whittaker, 2023), aprofundando a dependência estrutural típica do modelo de plataformas digitais.

A Tabela 6.1.1 sintetiza os detalhes técnicos dos modelos mais intensivos em computação, compreendendo computação de treinamento, parâmetros, conjunto de dados de treinamento, tipos e quantidade de hardware utilizado no treinamento, com vistas a ilustrar a indiscutível caminhada rumo a escalas gigantescas, não passíveis de serem acompanhadas pela maioria das empresas emergentes.

Tabela 6.1.1 - Detalhes técnicos de modelos intensivos em computação, por desenvolvedor e ano de lançamento

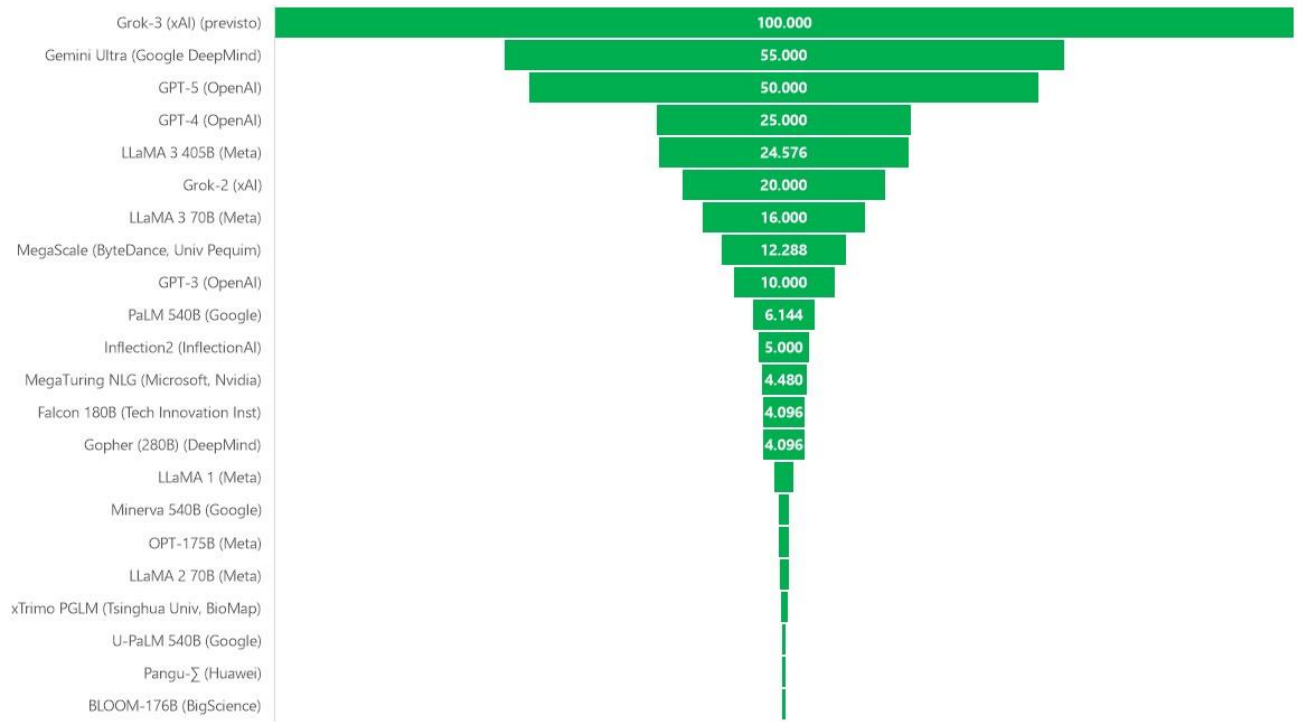
Modelo (desenvolvedor)	Ano	Capacidade de computação de treinamento	Parâmetros	Tamanho do conjunto de dados de treinamento	Hardware de treinamento	Quantidade de hardware
Gemini Ultra (Google DeepMind)	2024	50 bilhões PetaFLOP	n.d.	n.d.	Google TPU v4	55000
GPT-4 (OpenAI)	2023	21 bilhões PetaFLOP	n.d.	4.900.000.000.000	NVIDIA A100 SXM4 40 GB	25000
MegaScale (ByteDance, Univ Pequim)	2024	12 bilhões PetaFLOP	530 bilhões	n.d.	NVIDIA A100	12288
Mistral Large		20 bilhões PetaFLOP	n.d.	n.d.	NVIDIA H100 PCIe	n.d.
Inflection2 (InflectionAI)	2023	10,1 bilhões PetaFLOP	n.d.	n.d.	NVIDIA H100 SXM5	5000
PaLM2 (Google)	2023	7,34 bilhões PetaFLOP	340 bilhões	2.700.000.000.000	Google TPU v4	n.d.
LIAMA 3 70B (Meta)	2024	6,3 bilhões PetaFLOP	70 bilhões	15.000.000.000.000	NVIDIA H100 SXM5	16000
Claude2 (Anthropic)	2023	3,87 bilhões PetaFLOP	n.d.	n.d.	n.d.	n.d.
Falcon 180B (Tech Innovation Inst)	2023	3,76 bilhões PetaFLOP	180 bilhões	2.626.000.000.000	NVIDIA A100 SXM4 40 GB	4096
Grok-1	2023	2,9 bilhões PetaFLOP	314.000.000.000	n.d.	n.d.	n.d.
Minerva 540B (Google)	2022	2,74 bilhões PetaFLOP	540 bilhões	613.875.000.000	Google TPU v4	1024
GPT-3,5 (text-davinci-003) (OpenAI)	2022	2,58 bilhões PetaFLOP	175.000.000.000	n.d.	NVIDIA A100 SXM4 40 GB	n.d.
U-PaLM 540B (Google)	2022	2,53 bilhões PetaFLOP	540.000.000.000	n.d.	Google TPU v4	512
PaLM 540B (Google)	2022	2,53 bilhões PetaFLOP	540.000.000.000	585.000.000.000	Google TPU v4	6144
Qwen-72B (Alibaba)	2023	1,3 billion PetaFLOP	72.000.000.000	n.d.	n.d.	n.d.
MegaTuring NLG (Microsoft, Nvidia)	2021	1,17 bilhões PetaFLOP	530.000.000.000	202.500.000.000	NVIDIA A100 SXM4 80 GB	4480
ChatGLM3 (Zhipu AI)	2023	1,09 bilhões PetaFLOP	130 bilhões	1.050.000.000.000	n.d.	n.d.
Ernie 3.0 Titan (Baidu)	2021	1,04 bilhões PetaFLOP	260 bilhões	668.000.000.000	Huawei Ascend 910	n.d.
LIAMA 2 70B (Meta)	2023	810 milhões PetaFLOP	70 bilhões	1.500.000.000.000	NVIDIA Tesla V100 DGXS	n.d.
Gopher (280B) (DeepMind)	2021	631 milhões PetaFLOP	280 bilhões	225.000.000.000	NVIDIA A100 SXM4 80 GB	1000
Yi-34B (01.AI)	2023	610 milhões PetaFLOP	34 bilhões	n.d.	Google TPU v3	4096
xTrimo PGLM (Tsinghua Univ, BioMap Research)	2023	600,01 milhões PetaFLOP	100 bilhões	n.d.	n.d.	n.d.
BLOOM-176B (BigScience)	2022	577 milhões PetaFLOP	176 bilhões	n.d.	NVIDIA A100 SXM4 40 GB	768
Pangu-Σ (Huawei)	2022	577 milhões PetaFLOP	176 bilhões	262.500.000.000	NVIDIA A100 SXM4 80 GB	384
Pangu-Σ (Huawei)	2023	467 bilhões PetaFLOP	1,085 trilhão	246.750.000.000	NVIDIA A100 SXM4 80 GB	512
					Huawei Ascend 910	512

Fonte: Elaborado a partir da Base de Dados da Epoch

O Gráfico 6.1.7 complementa a tabela anterior ao apresentar a evolução em número de chips utilizados no treinamento dos modelos de IA, complementando as informações da Base de Dados Epoch com dados levantados pela LifeArchitecture.ai, que serve como *proxy* da

corrida da IA que parece se transformar em corrida por chips e por modelos gigantes. De fato, parece haver uma correlação entre as empresas mais inovadoras em IA e as que mais se destacam em modelos computacionalmente intensivos, que poderão alcançar cem mil GPUs até o final de 2024, com o Grok-3, da xAI, empresa do Elon Musk.

Gráfico 6.1.7 – Número de chips empregados no treinamento dos modelos de IA mais intensivos em computação



Fonte: Elaborado a partir de Epoch Database e LifeArchitecture.ai³⁰³

A consultoria SemiAnalysis analisa minuciosamente a capacidade computacional da IA, com informações surpreendentes, não usualmente divulgadas pelas empresas. Descreve o Google como a empresa mais rica em computação do mundo, com uma infraestrutura estimada vinte vezes mais poderosa que a OpenAI até o final de 2024 (Patel e Nishball, 2023). Após as aquisições programadas para 2024, a Meta deverá ocupar a segunda posição no ranking de maior número de GPUs H100 do mundo.

6.2. Evolução das patentes de IA e seus principais protagonistas

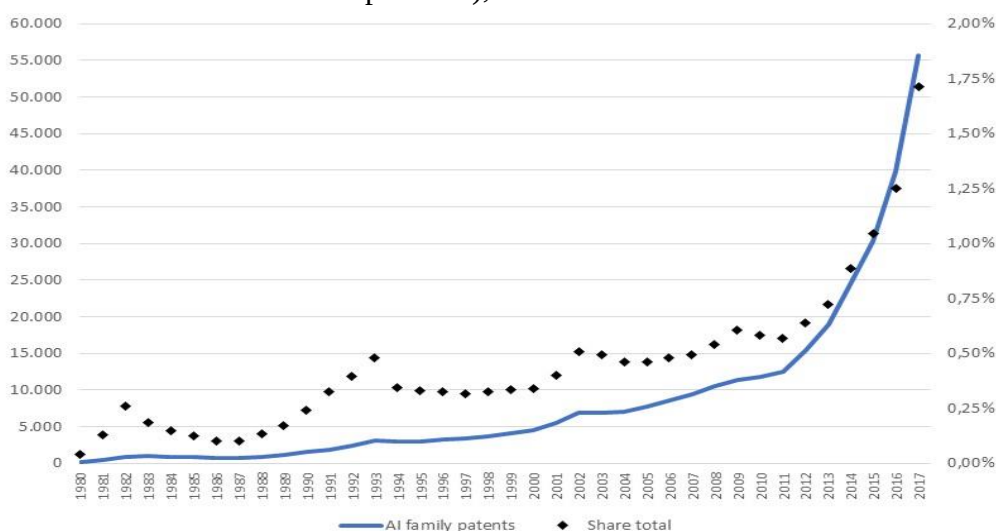
Muito antes da expansão dos sistemas de redes neurais profundas na década passada, as tecnologias de IA recebiam proteção por meio de patentes de invenção de IA “embarcadas” em conjuntos maiores compostos por hardware, software e os próprios sistemas de IA, quando a indústria ainda não havia passado para uma arquitetura modular e os principais *players* eram empresas verticalmente integradas.

³⁰³ <https://docs.google.com/spreadsheets/d/1O5KVQW1Hx5ZAkcg8AIRjbQLQzx2wVaLl0SqUu-ir9Fs/edit#gid=1381013921>. Acesso em: 16 jun.2024

As mudanças na arquitetura da indústria de TICs levaram à especialização das empresas em módulos particulares, fragmentou as tecnologias e desmembrou a proteção da invenção em patentes separadas de hardware, de software e de sistemas de IA.

A consolidação dos modelos de redes neurais profundas levou ao crescimento expressivo das patentes de invenção de IA a partir de 2012, ampliando sua participação no total de pedidos de patentes no mundo. O Gráfico 6.2.1 mostra dados de pedidos de famílias de patentes de IA desde 1980 e estima sua participação no total dos pedidos de patentes, mostrando não somente o aumento significativo dos pedidos, que se torna exponencial a partir de 2012, como sua participação no total de pedidos de patentes passa de 1,2% em 2012 para 2,8% em 2017. Se considerados apenas os dados do USPTO, a participação dos pedidos de patentes de IA no total passou de 9% em 2002 para 16% em 2018 (ou 42% se consideradas todas as patentes onde IA aparece em todas as subclasses de tecnologia usadas pelo USPTO) (Toole et al. 2020:2).

Gráfico 6.2.1. - Número de pedidos de patentes de IA e proporção no total (famílias de patentes), 1980-2017



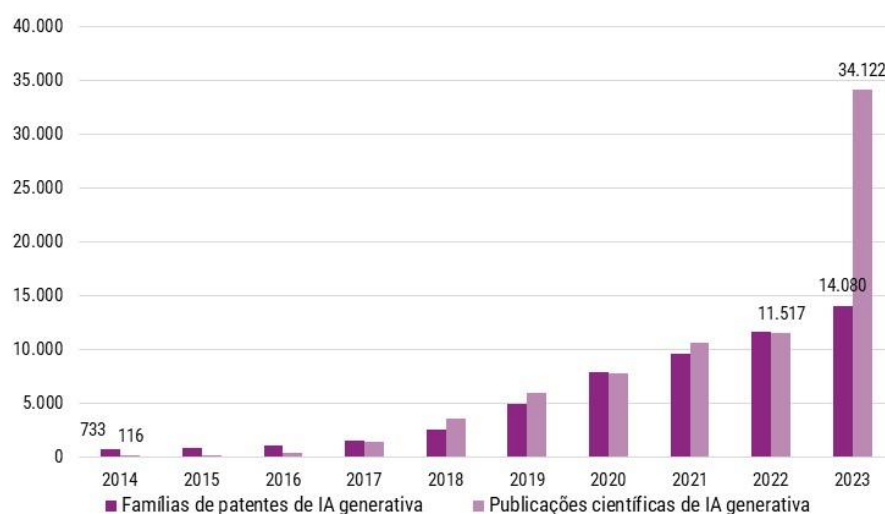
Fonte: Elaborado a partir de WIPO (2019) e <https://www3.wipo.int/ipstats/>. Acesso em 02 jan.2023

O número de pedidos de patentes relacionadas a IA avança desde a década passada (WIPO, 2019), com a IA preditiva e se torna exponencial com a IA generativa e a emergência de *breakthroughs* das novas arquiteturas de redes neurais, conforme retratado em sucessivas versões do *AI Index Report do Human Centered Artificial Intelligence* da Universidade de Stanford, a partir de dados do Center for Security and Emerging Technology. Em 2022, mostrou um aumento de 63% das patentes concedidas frente ao ano anterior e 31 vezes o número de patentes concedidas em 2010 (Maslej et al., 2024). A China assumiu a liderança, com o maior número de famílias de patentes, representando 70% do total mundial e com crescimento acelerado (50% ao ano), ultrapassando os EUA (com 11% do total), Coreia (8%), Japão (6%), Índia (3%), Reino Unido e Alemanha. O *gap* entre China e Estados Unidos é menor no caso

das publicações científicas, que respondem por 32% e 31% do total global, respectivamente, seguidos pelo Reino Unido (8%), Alemanha (5%), Canadá, Coreia, Japão, Índia, Austrália e França.

Relatório da OMPI, divulgado em julho de 2024, mostra informações específicas das patentes da IA generativa, que somaram 54.358 famílias de patentes publicadas (89% delas ativas no final de 2023), das quais 14.080 somente em 2023 (mais de quarto do acumulado no período, representando 6,14% do total de 229.228 patentes do conjunto da IA no ano, frente a 4,2% em 2017). O Gráfico 6.2.2 mostra a evolução das patentes e publicações científicas da IA generativa entre 2014 e 2023. Um crescimento ainda mais expressivo é esperado para os próximos dois anos, considerando o prazo de 18 meses entre a data do depósito e a publicação da patente. O aumento das publicações científicas relacionadas a IA generativa foi ainda mais significativo, com o número acumulado de mais de 75 mil publicações entre 2014-23, mais de 34 mil no ano de 2023.

Gráfico 6.2.2 - Famílias de patentes e publicações acadêmicas de IA generativa, por ano, 2014–2023

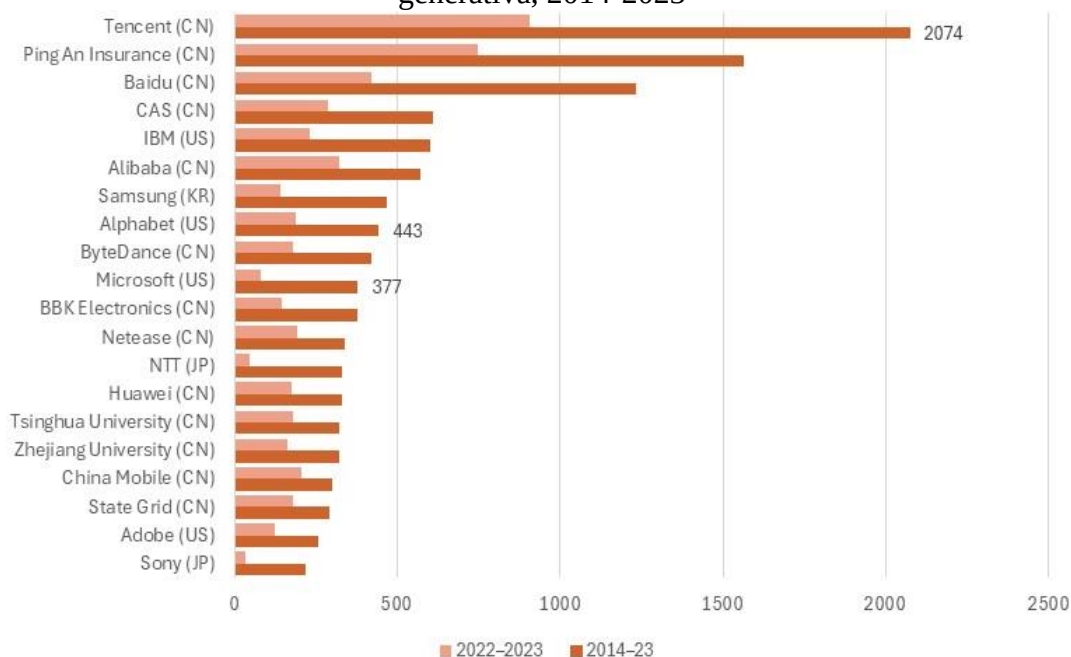


Fonte: WIPO (2024a)

Os principais detentores das patentes da IA generativa são empresas chinesas, com Tencent, Ping An Insurance e Baidu encabeçando a lista das vinte organizações com os maiores portfólios, além da Alibaba, ByteDance, Huawei e universidades chinesas. Apenas quatro empresas americanas, nomeadamente IBM, Alphabet/Google, Microsoft e Adobe, constam da lista, além de duas japonesas e uma coreana (Gráfico 6.2.3).

Segundo dados da Statista (2024), é possível perceber que a predominância de organizações chinesas no conjunto das patentes de IA e aprendizado de máquina começou em 2020, com a disputa pela liderança entre Baidu e Tencent.

Gráfico 6.2.3 – Vinte principais organizações detentoras de famílias de patentes de IA generativa, 2014-2023



Fonte: WIPO (2024a)

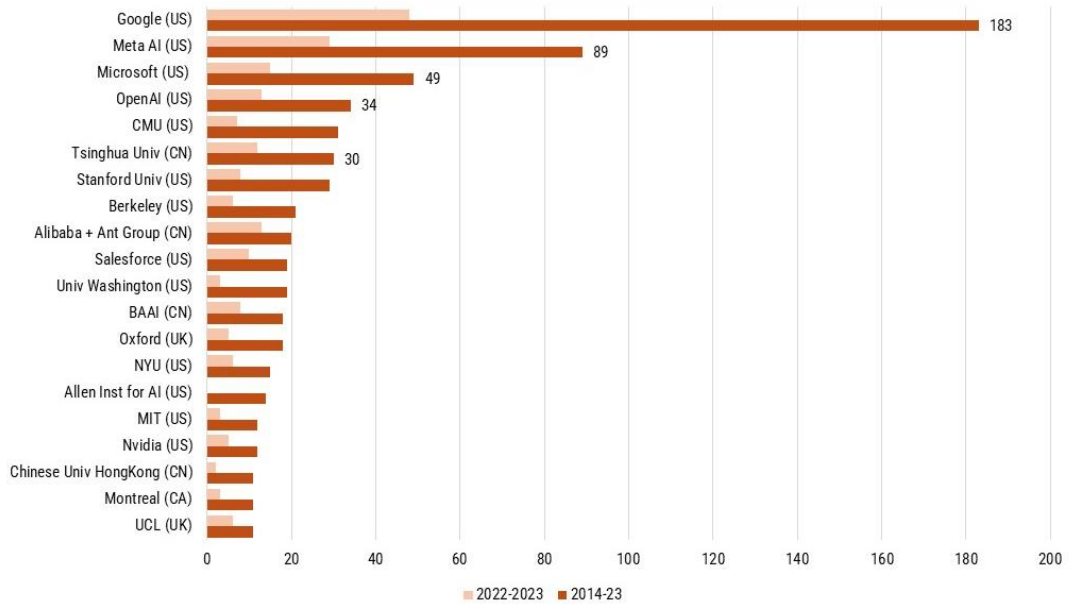
É impressionante o contraste entre o *ranking* das instituições líderes em patentes e daquelas que lideram o lançamento de modelos de IA generativa comentados na seção anterior. As vinte principais organizações que lançaram modelos são discriminadas no Gráfico 6.2.4 no mesmo período 2014-2023. A liderança irrefutável é das empresas americanas, nomeadamente Google (incluindo DeepMind), Meta AI, Microsoft e OpenAI. Entre as empresas chinesas o destaque é a Alibaba (incluindo a subsidiária financeira Ant Group), mas atrás de universidades americanas (CMU, Stanford, Berkeley) e da Tsinghua University.

O Gráfico 6.2.5 plota os dados dos dois indicadores com o objetivo de fornecer uma melhor visualização da discrepância entre as empresas líderes em patentes e no lançamento de modelos, com apenas Alphabet/Google, Microsoft, Tsinghua University e, em menor escala, Alibaba, se destacando em ambos os indicadores.

Esses resultados divergentes decorrem da distinção óbvia entre os dois tipos de indicadores, bem como ao fenômeno do *open-source* em IA na economia americana. Primeiro, o uso de patentes publicadas como indicador de desempenho tecnológico é usual, apesar de sujeito a questionamentos. O uso de patentes concedidas, se comparado a pedidos de patentes, pode evidentemente carregar vieses decorrentes do desempenho de cada ENPI e da prioridade da PI como política pública – como os inúmeros incentivos ao patenteamento pelos governos

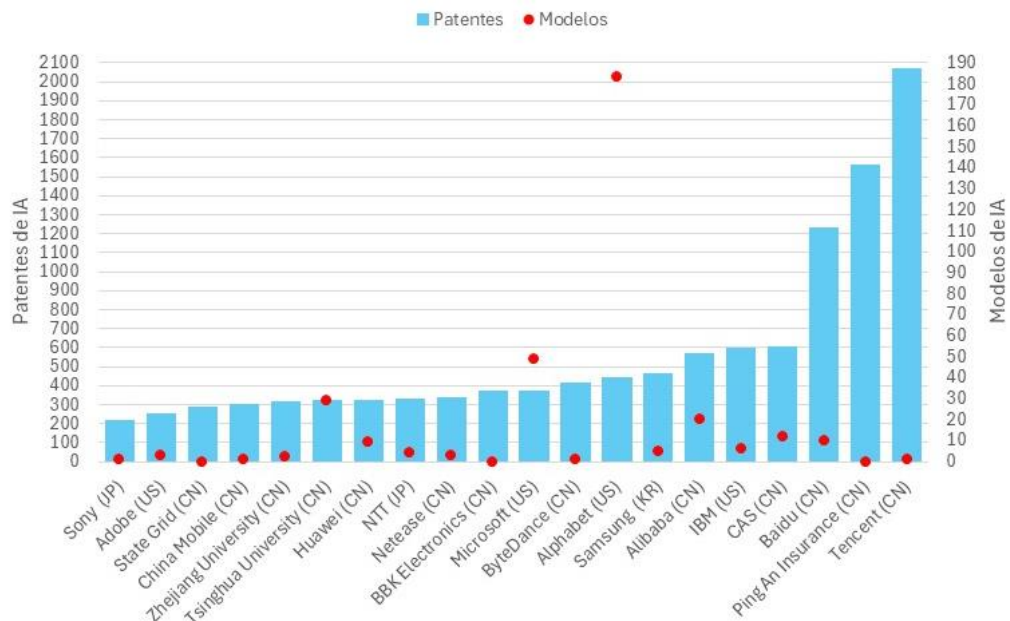
da China, comentados no Capítulo anterior - ao invés de refletir apenas a performance em desenvolvimento tecnológico.

Gráfico 6.2.4 – Principais desenvolvedores de modelos de IA generativa lançados em 2014-2023



Fonte: Elaborado a partir de Epoch Database.

Gráfico 6.2.5 – Número de patentes de IA generativa e número de modelos lançados pelos vinte principais detentores de patentes, 2014-2023



Fonte: Elaborado a partir de WIPO (2024a) e Epoch Database

Em segundo lugar, patentear ou não uma invenção depende de cada estratégia empresarial, não sendo, necessariamente, um requisito para a colocação de um produto/serviço novo no mercado. O caso da OpenAI, uma das empresas com desempenho notável no lançamento de modelos paradigmáticos da IA generativa é ilustrativo de uma estratégia que optou por não privilegiar o patenteamento de invenções até 2023. O relatório WIPO (2024a)

atribuiu essa estratégia à natureza sem fins lucrativos da empresa e sua disposição em fornecer partes significativas de sua tecnologia em código aberto – ainda que, como visto, a OpenAI possua uma subsidiária com fins lucrativos e o *open source* tenha se consagrado como estratégia também de empresas que objetivam o lucro - ou, alternativamente, protegendo partes de suas invenções com segredos comerciais que, de outro modo, seriam divulgados com as patentes. Keseris e Kovarik (2023), entretanto, verificaram a publicação de seis patentes americanas da OpenAI no primeiro trimestre de 2024 (três concedidas e três pendentes), o que indicaria a mudança para uma estratégia defensiva de patentes. Ao contrário, acredita-se que tal reorientação estratégica mais provavelmente está vinculada ao lançamento recente de versões aprimoradas dos modelos e à tentativa de monetização na busca de desenvolver um modelo de negócios.

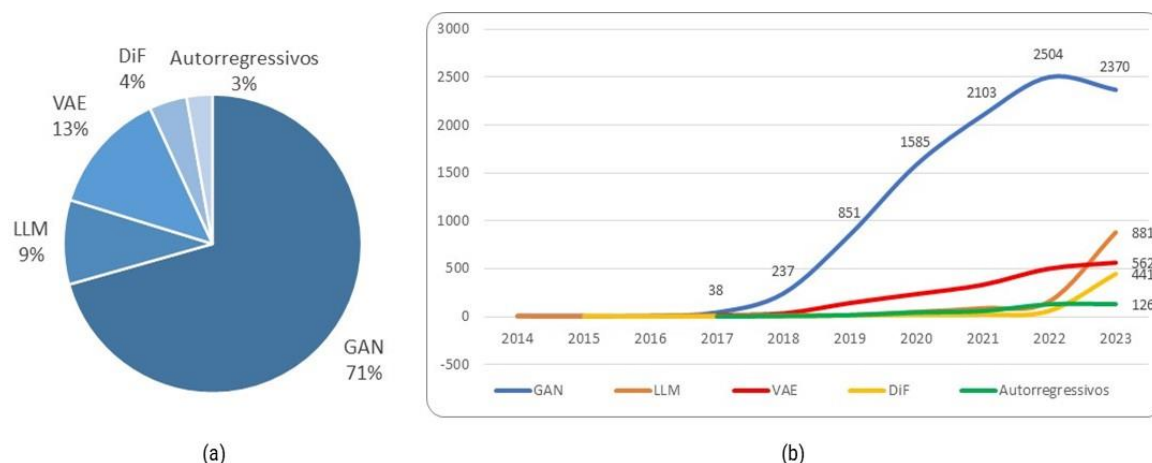
Terceiro, há uma especificidade da IA em relação a outros setores econômicos que deve ser levada em consideração, possivelmente atrelada ao regime *open-source*. O lançamento de novos modelos no mercado resulta de inovações algorítmicas, arquiteturas, treinamento e outros detalhes técnicos que caracterizam atividades mais próximas do avanço do conhecimento (e se refletem na expansão desenfreada de artigos científicos), ainda com dificuldades para identificar aplicações reais e usos específicos, dificultando o patenteamento (que exige a indicação de aplicação industrial), e condizente com o fato da maioria dos desenvolvedores da IA ainda não ter conseguido viabilizar a monetização das invenções e uma aplicação “matadora”.

A classificação da maioria das aplicações das patentes de invenção da IA generativa no setor de software e outros aplicativos (Apêndice A8 de WIPO, 2024:87-88) pode estar ligada às próprias incertezas quanto ao uso em setores *downstream*, como é comum em tecnologias de propósito geral, levando as empresas a indicarem aplicações industriais no seu setor de origem, que é o setor de software. Da mesma forma, o requisito novidade em relação ao estado da arte não seria, a princípio, passível de ser atendido por muitos modelos derivados (desenvolvidos a partir de modelos pré-treinados de terceiros lançados em *open source*), que não representaram avanço no estado da arte (melhoramento SOTA). Adicionalmente, a predominância de patentes do setor de software sugere que devem estar voltadas para proteção de invenções algorítmicas e modelos de IA que poderiam ser consideradas um conhecimento fundamental.

Quarto, existem peculiaridades inerentes ao sistema de patentes, como o prazo de 18 meses entre o depósito e a publicação da patente, que parecem inconciliáveis com uma tecnologia que evolui em ritmo tão rápido como a IA. Isso, ademais, compromete o uso do indicador de patentes para fins de prospecção e identificação de tendências que, em poucos anos

ou mesmo meses, podem ser completamente ultrapassadas por novas tecnologias e modelos totalmente distintos. Por exemplo, os dados de patentes de WIPO (2024a) mostram predomínio de patentes relacionadas às redes neurais GAN (71% do total) (Gráfico 6.2.6 (a)), cujo principal uso é imagem (visão computacional), que apresentam o maior crescimento até 2022 (Gráfico 6.2.6 (b)), ao passo que nos dados da Epoch (Gráfico 6.2.7 (a)) mostram a predominância dos LLMs (linguagem) (58% dos modelos), que apresentam crescimento mais acelerado no período (Gráfico 6.2.7 (b)). Com efeito, os LLMs, baseados em redes Transformers são a principal tendência desde a revolução provocada com o ChatGPT, corroborando que o lançamento de modelos é um melhor indicador de prospecção tecnológica. Além disso, as GAN foram suplantadas pelos modelos de difusão, o que torna a vantagem em patentes um indicador pouco promissor de seu uso em inovações concretas potenciais.

Gráfico 6.2.6 - Famílias de patentes globais em cinco principais tipos de modelos de IA generativa, acumulado 2014–2023 (a) e evolução 2014-2023 (b)



Fonte: Elaborado a partir de WIPO (2024)

Por fim, deve ser considerado o problema da falta de definições consolidadas no campo da IA, que pode se refletir nos resultados encontrados com dados de patentes dependendo da estratégia de busca escolhida pelo pesquisador. Por exemplo, foi realizada uma busca na base Patentscope da OMPI, dos depósitos de pedidos empregando os códigos IPC (*International Patent Classification*) identificados como mais frequentes de aprendizado profundo em três níveis³⁰⁴, combinados com a palavra-chave “*Transformers*” no título e no resumo dos pedidos, entre 2016-2022, gerando resultados bastante distintos em termos dos principais titulares, com

³⁰⁴ O Código IPC foi considerado em 3 níveis de hierarquia (seção, classe e subclasse). Por exemplo G06N, um dos códigos comuns na classificação das patentes de *deep neural networks*, compreende a seção (1º nível: G), a classe (2º nível: 06) e a subclasse (3º nível: N). Ou seja: G = física; 06: *Computing; calculating or counting*; e N = arranjos computacionais baseados em modelos computacionais específicos. Xin et al. (2021), por exemplo, identificaram as 20 principais IPC (*International Patent Classification*) de patentes médicas relacionadas à IA, cujas IPCs mais frequentes mudaram entre 1972-2010 e 2011-2017, que podem refletir desde mudanças nos critérios de classificação, até mudanças de paradigma da IA. Ver Anexo 1.

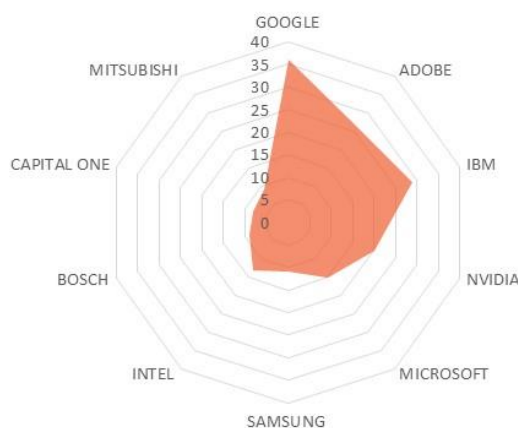
predominância clara das empresas americanas (Gráfico 6.2.8).

Gráfico 6.2.7 – Lançamento de modelos de IA por domínio, acumulado 2014-2023 (a) e evolução 2014-2023 (b)



Fonte: Elaborado a partir de Epoch

Gráfico 6.2.8 - Dez principais depositantes de pedidos de patentes relacionadas a redes Transformers, USPTO, 2016-2022



Fonte: Elaboração própria a partir de dados do Patentscope/WIPO

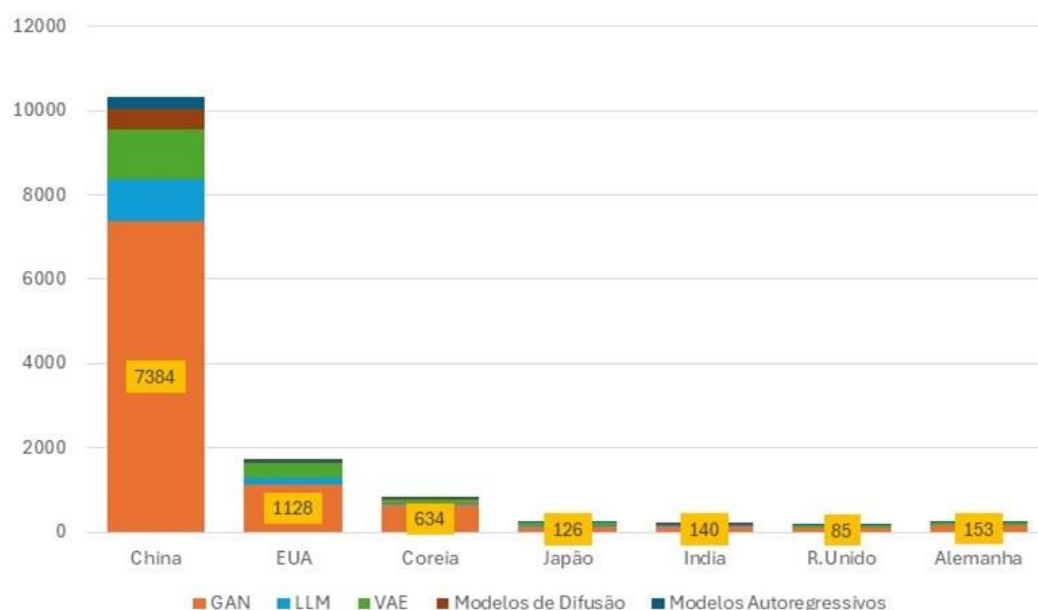
A escolha de efetuar a busca, em inglês, usando o tipo de rede neural, combinada com “*deep learning*”, pareceu mais recomendável para uma tecnologia ainda não madura, onde não há termos padronizados, ao invés de usar expressões genéricas como “*artificial intelligence*”, ou mesmo “*large language models*”, entre outras, que ganharam notoriedade na mídia, mas não são termos científicos reconhecidos e não parecem a nomenclatura empregada pelas empresas e inventores quando objetivam obter um direito legal³⁰⁵. Com base nessa metodologia de busca, os dez principais depositantes de patentes de arquiteturas *Transformers* no USPTO evidenciam mais uma vez o predomínio do Google, seguido por empresas que historicamente se destacam no campo da IA, como IBM e Microsoft, e novos grandes *players* da IA

³⁰⁵ A classificação de patentes ainda é uma tarefa realizada por especialistas e examinadores na maioria dos principais nacionais, mas já há sistemas de classificação a partir de processamento de texto usando aprendizado profundo em redes neurais. A identificação das IPCs empregada nesta tese foi feita por meio de pesquisa na base de dados da Derwent/Clarivate, que dispõe de informações do título e abstract dos pedidos de patentes de “*deep learning*”, por mineração de textos e dados, empregando o software gratuito [Iramuteq](https://www.aramuteq.com/), identificando então as classes mais frequentes.

generativa, como Nvidia, Intel, Adobe, etc.

Por outro lado, os resultados de WIPO (2024a) que mostram a preponderância de patentes relacionadas a arquitetura GAN, possivelmente refletem um viés da China pelo expressivo número de patentes das organizações do país, expressivamente maior do que outro país (incluindo Estados Unidos) que talvez valorizem esse tipo de redes neurais aplicadas a dados de imagem (Gráfico 6.2.9).

Gráfico 6.2.9 – Patentes de IA generativa por tipo de arquitetura e pelos países líderes, 2014-2022



Fonte: WIPO (2024a)

Em outras palavras, o uso de dados de patentes não seria a melhor escolha para a prospecção tecnológica, considerando o crescente emprego de mecanismos alternativos para apropriabilidade da inovação, além de não conseguir capturar os verdadeiros *breakthroughs* da IA generativa no mercado que são indiscutivelmente os LLMs, modelos de difusão, multimodais e de fundação. Por um lado, os prazos de sigilo das patentes são incompatíveis com uma tecnologia que evolui de forma tão acelerada, tornando dados de patentes permanentemente defasados; por outro, as peculiaridades do setor em termos de *open source* e as indefinições das aplicações da IA, lançam questionamentos adicionais sobre as patentes tanto como estratégia de apropriabilidade quanto indicador de tendências na área.

Quinto, parece prematuro afirmar que a China está definindo o ritmo do progresso técnico da IA como sugerem os dados de patentes, mesmo reconhecendo o avanço significativo das organizações chinesas em IA generativa, que tanto preocupa o governo americano a ponto de tentar fechar as torneiras do hardware por meio de restrições à exportação. No entanto, é preciso admitir que o número de patentes pode não ser o melhor indicador para medir a

liderança em IA, além de ofuscar a importância do *open source* para a disputa corporativa pela liderança e progresso técnico da área. As empresas americanas lideram o lançamento de modelos de IA, concomitantemente à publicação de artigos que os detalham, ainda que restringindo aspectos técnicos essenciais.

CAPÍTULO 7 – DISCUSSÃO SOBRE O REGIME DE APROPRIABILIDADE DA IA

A abordagem do PFI, desenvolvida para o contexto da economia industrial, parece capaz de fornecer um arcabouço para a análise do regime de apropriabilidade da inovação da IA. Tal estrutura foi revisitada em diferentes momentos, incorporando novas variáveis da economia digital, em particular, a importância dos dados e a complementaridade nos ecossistemas de plataformas digitais. Não obstante, foi tratada a apropriabilidade para o conjunto das tecnologias digitais emergentes, com ênfase em *smartphones* e tecnologias de interoperabilidade (*wireless*) (Teece, 2018a; Teece, 2016; Teece, 2012), sem descer aos detalhes que distinguem a IA de qualquer outra tecnologia conhecida e que demanda um olhar mais cuidadoso sobre este caso específico, particularmente diante dos novos desafios dos modelos generativos e que deverão se ampliar quando da implementação dessas tecnologias em setores econômicos a jusante.

A abordagem original mostrou a importância das estratégias das firmas para definir o *design* tecnológico dominante e a gestão de ativos/tecnologias complementares exigidos para a apropriabilidade da inovação, variáveis mais relevantes do que o sistema de PI e a natureza da tecnologia para o sucesso de uma inovação, particularmente em regimes fracos de patentes, em que há probabilidade significativa de direitos anulados ou contornados, tornando a proteção legal quase impossível. A captura de valor da inovação, após o padrão tecnológico estabelecido, passa a depender fortemente de um conjunto de ativos (e capacidades) complementares. De fato, pouca atenção foi dedicada a explicitar o regime jurídico de PI, a partir da suposição de que para uma ampla gama de tecnologias e indústrias, o regime de proteção por patentes é fraco (Teece, 1986:290). No entanto, as implicações econômicas são muito distintas em regimes de apropriabilidade fortes ou fracas. Nesse sentido, é importante dedicar alguma atenção ao tema.

No caso da IA, a principal característica do regime de apropriabilidade depende da competição pela captura de efeitos de rede, acumulação de grandes conjuntos de dados e capacidade computacional – esta última variável recebendo pouca atenção da literatura do PFI –, por meio da estruturação de ecossistemas de inovação e do modelo de negócios das plataformas digitais. Enquanto na economia industrial o processo de criação e captura de valor de inovações discretas ocorria sequencialmente (ou seja, a empresa devia primeiro *criar* valor antes de poder *capturá-lo*), nas inovações da economia digital, como a IA, a criação e captura de valor parecem ocorrer simultaneamente. Adicionalmente, a definição do *design* tecnológico dominante parece ocorrer paralelamente a decisões, essencialmente dinâmicas, relacionadas a ativos/tecnologias complementares e de estruturação de modelos de negócios. Esse regime parece, contudo, passar por mudanças importantes com o avanço recente da IA generativa.

No auge dos modelos preditivos, os sistemas de redes neurais profundas se tornaram o *design* tecnológico dominante, em um ambiente “coopetitivo” no qual as empresas competiam para consolidar posições em segmentos estanques do mercado, enquanto cooperavam no desenvolvimento de modelos e no avanço do conhecimento, por meio da coautoria de publicações científicas sobre as principais invenções algorítmicas e tipos de redes neurais artificiais, mesmo que com estratégias próprias quanto aos ativos (e tecnologias) complementares.

Tal regime de apropriabilidade inclui amplos direitos de propriedade intelectual, liderança no *design* tecnológico e estratégias quanto a ativos/tecnologias complementares, como os massivos e variados conjuntos de dados em bases proprietárias e, com o avanço dos modelos generativos, a internalização de infraestrutura computacional e hardware especializado por um conjunto reduzido de empresas que passaram a constituir o núcleo infraestrutural das *BigTechs*, americanas e chinesas, também orquestradoras de plataformas digitais. Esse subconjunto de empresas internalizou desde cedo os ativos complementares, enquanto outras optaram pela contratação de terceiros de partes da pilha de computação, como chips especializados, data centers, cabos ou satélites de transmissão de dados, entre outros.

Enquanto o núcleo infraestrutural integralizou esses ativos por meio de típica integração vertical a montante, criando poderosas estruturas de data centers, redes de transferência e interconexão de dados, supercomputadores e chips especializados, a maioria dos *players*, principalmente empresas de menor porte, os contratam das empresas líderes, de incumbentes de alta tecnologia, de fabricantes especializados, de titulares de tecnologias de interoperabilidade etc. e outros que compõem o ecossistema de inovação da IA, caracterizado por fortes assimetrias. O código aberto, por outro lado, passou a desempenhar papel central no caso de uma tecnologia que depende da difusão por meio dos efeitos de rede, instrumento fundamental para a (re) definição do (novo) *design* tecnológico dominante.

A nova fase dos modelos generativos delineou mudanças importantes no regime de apropriabilidade. O subconjunto infraestrutural das *BigTechs*, ao integralizar os ativos (tangíveis) complementares, em poderosas infraestruturas de computação na nuvem, progressivamente o veículo principal do avanço da indústria, composto por uma miríade de insumos e produtos conectados como “XaaS”, com potencial inédito de criação e captura de valor das suas próprias inovações e de terceiros, muitas vezes rivais, de menor ou maior porte empresarial. Não é por outro motivo que todas empresas importantes de TICs, da mesma forma que provedores de bancos de dados e chips, beneficiários indiretos da expansão desenfreada da

IA generativa, somam-se ao subconjunto das *BigTechs* para criar suas próprias infraestruturas de computação na nuvem.

Na IA preditiva a corrida tecnológica se traduziu em uma corrida por patentes, mas a chegada da IA generativa provocou mudanças profundas com o avanço sem precedentes do *open source* e o patenteamento começando a ser relegado a segundo plano (exceto na China), privilegiando mecanismos alternativos de PI (como *trade secrets*) e a apropriabilidade pelo controle dos ativos tangíveis da IA.

Com vistas a melhor explorar essas ideias, a seção 7.1 procura examinar o regime de apropriabilidade da IA preditiva, no contexto dos ecossistemas de plataformas digitais para, na seção seguinte, discutir as mudanças no padrão de inovação e competição que estão sendo imprimidos pelos modelos gigantes da IA generativa, relacionados aos massivos e cada vez mais variados dados de treinamento e, em particular, ativos tangíveis infraestruturais. A construção da principal vantagem competitiva da IA generativa, que avança espetacularmente nos últimos anos com a escalabilidade dos modelos, possibilitou às provedoras da gigantesca capacidade computacional requerida pela corrida tecnológica de modelos próprios, de subsidiárias integrais ou minoritárias e mesmo parceiros sem vínculos acionários, ainda que todos enredados em aprisionamento técnico, financeiro e computacional. Rivals de destaque, como a Meta AI, apostam em uma estratégia aparentemente diferenciada, mas que, na realidade, reproduz a experiência do núcleo infraestrutural, em especial Alphabet e Microsoft, nos vínculos com o regime *open-source* desde meados da década passada, sem prescindir da construção de uma infraestrutura de classe mundial com um dos maiores *clusters* de computação do mundo, como anunciado na divulgação dos resultados do ano passado no relatório³⁰⁶. Com isso, parece confirmar a hipótese de que a corrida por modelos e pelo *design* dominante da IA se converteu em uma corrida desenfreada por capacidade computacional.

Uma estrutura ampla e complexa de proteção da PI para invenções que empregam modelos de IA, conjugada ao controle sobre conjuntos de dados e capacidade computacional extrema asseguram um poder descomunal para as empresas líderes. É o controle sobre esses recursos que permite a abertura de seu código fonte, mantendo em sigilo parte do conhecimento incorporado em detalhes técnicos dos modelos, em um regime híbrido de apropriabilidade em que o regime de PI convive de forma harmoniosa com o *open-source* com vistas à difusão e consolidação dos sistemas de IA.

³⁰⁶ Disponível em: https://s21.q4cdn.com/399680738/files/doc_financials/2023/q4/META-Q4-2023-Earnings-Call-Transcript.pdf. Acesso em: 02 fev.2024

Este capítulo analisa a evolução do regime de apropriabilidade e as principais mudanças em curso entre os modelos preditivos classificadores e modelos generativos, conforme sumarizado no Quadro 7.1, com implicações sobre o padrão de inovação, competição e modelo de negócios.

Quadro 7.1 – Quadro comparativo da apropriabilidade da inovação da IA preditiva e IA generativa

IA Preditiva	IA Generativa
Sistemas exitosos de IA em plataformas digitais (serviços, varejo e setores intensivos em dados)	Indústria criativa e potencial de aplicação em amplos setores econômicos <i>downstream</i> e na sociedade
Previsão, classificação e recomendação a partir de reconhecimento de padrões nos dados	Geração de conteúdo de texto, imagem, áudio, vídeo, código de programação
Modelos menores	Modelos enormes (parâmetros)
Unimodalidade	Multimodalidade
Unitarefa	Multitarefas (grande êxito em NLP)
Redes neurais artificiais: ConvNets, RNNs	Redes neurais artificiais: Transformers, GANs, VAEs
Intensidade computacional: elevada	Intensidade computacional: gigantesca
Ativos complementares: amplos conjuntos de dados	Ativos complementares: amplos conjuntos de dados e, principalmente, infraestrutura computacional
Padrão de inovação: Ecossistemas corporativos de inovação (complementaridade no contexto multi-invenção de plataformas digitais)	Padrão de inovação: Ecossistemas de inovação estendidos (complementaridade com hardware)
<i>Design</i> tecnológico dominante consolidado: redes neurais profundas com diversos usos	<i>Design</i> tecnológico dominante em formação : disputa pelo <i>design</i> se mistura com disputa pelos efeitos de rede
Padrão de apropriabilidade híbrido: DPIs amplos e complexos + <i>trade secrets</i> + Bibliotecas <i>open-source</i> + ativos complementares	Padrão de apropriabilidade híbrido: DPIs + <i>trade secrets</i> + modelos <i>open-source</i> no mercado + assinaturas pagas → contratos + ativos complementares
Modelo de negócios: Plataformas digitais	Modelo de negócios: Plataformas na nuvem (XaaS)
Padrão de competição: liderança BigTechs (“coopetição”)	Padrão de competição: BigTechs, incumbentes de alta tecnologia, startups emergentes (maior competição)

Fonte: Elaboração própria

7.1. Apropriabilidade da IA preditiva

O regime legal de PI da IA preditiva é extenso, intrincado e híbrido, estabelecido desde que os sistemas de redes neurais profundas se transformaram no *design* tecnológico dominante e ativos (tecnologias) complementares, como extensas bases de dados, que são cruciais não somente para a captura de valor, mas para o próprio funcionamento dessa tecnologia, juntamente com a capacidade computacional. Dessa forma, sobrepõe e maximiza direitos, protegendo insumos (bases de dados) e invenções da IA, mas que convive simultaneamente com o código aberto, o que parece representar uma estratégia híbrida onde firmas gerenciam DPIs (patentes, direito autoral e, na Europa, também direitos *sui generis*, além de marcas submarinas), em geral em estratégias defensivas e não como incentivo à inovação. Isso explica o patenteamento agressivo que vigorou até recentemente, paralelamente à manutenção de esquemas de código aberto, como também apontado por poucos autores. Contudo, ainda há muita incerteza sobre como a patenteabilidade da IA, o código aberto e o litígio evoluirão no futuro.

No estágio de desenvolvimento das inovações da IA preditiva, desde o surgimento dos jogos complexos até as inovações das plataformas digitais, a IA pareceu migrar do estágio pré-paradigmático para paradigmático, em que as técnicas de aprendizado profundo em redes neurais artificiais, com crescente acuracidade de resultados, pareciam representar o *design* dominante da IA, inclusive começando a definir padrões relevantes.

Alguns autores (Tomic, 2018, Ferràs-Hernandéz, Nylung e Brem, 2023), entretanto, consideram que o *design* dominante da IA ainda não está bem definido, com tecnologias ainda em desenvolvimento, com padrões disputados pelas firmas concorrentes, incluindo os modelos *open-source* como um método de consolidar o *design* tecnológico. Aqui, ao contrário, considera-se que a disseminação de sistemas de redes neurais profundas, em bibliotecas de treinamento, contribuiu para sua transformação no *design* dominante, suplantando técnicas como a IA simbólica, com avanço de patentes da IA preditiva e amplo emprego de inovações exitosas largamente disseminadas nas plataformas digitais. Esses indicadores parecem evidenciar que o *design* já foi definido. Se havia alguma dúvida, contudo, sua integração nas plataformas de *Cloud computing* pareceu a pá de cal definitiva.

Acredita-se que, diferentemente da era industrial, o *design* tecnológico dominante da IA preditiva parece ter sido construído não por uma firma isolada, que alcançaria vantagens competitivas, mas de forma colaborativa pelas *BigTechs*, além de inegáveis contribuições de devs., a partir de interações no âmbito acadêmico, nas conferências e nas competições típicas do campo, sem esquecer das inovações de complementadores atrelados ao ambiente das plataformas digitais. Nesse estágio, o *design* tecnológico parecia caminhar no sentido de se converter em um padrão técnico. Segundo o gerente de padrões corporativos da Microsoft, Jason Matusow³⁰⁷ “*We cannot ignore that we have to apply global standards to get the maximum benefit based on responsible use of AI technology,*” de forma a desbloquear todo o potencial da IA.

Foi visto que padrões técnicos são relevantes em muitas indústrias e cruciais em indústrias com arquiteturas modularizáveis, onde partes do sistema são desenvolvidas separadamente por empresas especializadas e depois montadas, mas exigindo interfaces e padrões bem definidos e PI codificada com direitos bem definidos (Pisano, 2006b). Nas indústrias de TICs a definição de *designs* dominantes mostrou-se tema fundamental, com padrões para a interoperabilidade das muitas tecnologias envolvidas, além de influenciar as condições de apropriabilidade.

³⁰⁷ Disponível em: <https://www.iso.org/news/ref2763.html>. Acesso em 09 jan. 2022

Padrões e as patentes essenciais (SEPs) constituem um capítulo à parte na questão da apropriabilidade das inovações da IA. Por um lado, o controle do viés dos dados e da opacidade dos sistemas de aprendizado e algoritmos evolutivos certamente constituem um desafio para a definição de padrões técnicos. Por outro, essas tecnologias exigem compatibilidade e interoperabilidade, com conexão nunca vista entre celulares, computadores, plataformas e usuários. Nas futuras casas e fábricas inteligentes, nos veículos autônomos, na robótica e muitas outras aplicações da IA, serão efetivamente indispensáveis.

No entanto, se o arranjo legal relacionado a padrões, patentes essenciais e condições FRAND já é hermético, sua imposição a tecnologias ainda em desenvolvimento, como a IA, torna-se efetivamente desafiador e pede uma reflexão sobre implicações potenciais para a captura de valor das inovações.

Iniciativas para converter o *design* dominante de redes neurais profundas em padrões técnicos começaram a surgir, voltadas para responder às questões da baixa transparência dos sistemas, com riscos e danos potenciais, ao viés nos dados e *BigData*, e à interoperabilidade, por meio de tecnologias de telecomunicação sem fio.

Há um número não desprezível de propostas para estabelecimento de padrões para a IA por diversas organizações (SSOs), visando a confiabilidade da IA (*Trustworthy AI*, do ITU)³⁰⁸, a IA responsável (*Responsible AI standardization*, da ISO)³⁰⁹ e principalmente a explicabilidade da IA (XAI). Simultaneamente, despontam iniciativas de padrões para arquitetura de referência de *BigData*, com destaque para padrões avançados da ISO (série ISO/IEC-20547)³¹⁰.

As próprias BigTechs têm privilegiado princípios éticos da IA e possuem iniciativas próprias para interpretabilidade, explicabilidade e confiabilidade das técnicas de IA. Todas têm programas específicos com essa finalidade, como o People+AI Research (PAIR)³¹¹ do Google, o Responsible and Trusted AI³¹² da Microsoft, o SageMaker Clarify³¹³ da Amazon, entre outros. Essas firmas participam de diversas SSOs ao redor do mundo, de modo que essas iniciativas internas acabarão influenciando no estabelecimento de padrões.

³⁰⁸ Disponível em: <https://www.iso.org/news/ref2530.html>. Acesso em 09 ago. 2022

³⁰⁹ Disponível em: <https://www.iso.org/news/ref2763.html>. Acesso em 10 ago. 2022

³¹⁰ Disponível em: <https://www.iso.org/advanced-search/x/title/status/P,U/docNumber/20547/docPartNo/docType/0/langCode/ics/currentStage/true/searchAbstract/true/stage/stageDateStart/stageDateEnd/committee/sdg>. Acesso em 10 ago. 2022

³¹¹ Disponível em: <https://pair.withgoogle.com/>. Acesso em 10 ago. 2022

³¹² Disponível em: <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar6>. Acesso em 10 ago. 2022

³¹³ Disponível em: <https://www.amazon.science/blog/improving-explainable-ais-explanations>. Acesso em 10 ago. 2022

Iniciativas de padronização relacionadas a explicabilidade dos modelos da IA surgiram formalmente em 2017, por meio de estudos da Comissão Europeia visando o lançamento de projeto piloto relacionado a patentes essenciais em tecnologias selecionadas, dentre elas a IA. Empregam uma abordagem holística, com maior equilíbrio entre necessidades de titulares e implementadores de SEP, sobretudo firmas de pequeno porte. Foi visto que os principais problemas das SEPs são a definição da essencialidade (autodeclarada, com grande assimetria e riscos para empresas de menor porte), definições claras dos termos FRAND (valor dos *royalties*) e práticas anticompetitivas que causam frequentes litígios (Kostadinova, 2020).

A Comissão Europeia vem discutindo propostas de maior transparência das SEPs, incentivando *pools* de patentes e outras plataformas de licenciamento, maior colaboração entre SSOs e ENPIs e até criação de repositório de declarações de essencialidade feitas para SSOs. Sobre as condições de licenciamento FRAND, os debates contemplam princípios gerais para taxa agregada de *royalties* que seriam razoáveis, não discriminação entre implementadores, e licenciamento global, com base na criação de grupos de especialistas em licenciamento/avaliação de SEP visando troca de experiências e conhecimento econômico, jurídico e técnico para assessorar a Comissão no monitoramento dos mercados de licenciamento SEP.

Fischer (2020) também argumenta que o interesse em submeter tecnologias de IA a normas e padrões estava sendo motivado pelas questões de segurança, confiabilidade e preocupações éticas. Entre as iniciativas que menciona de padrões de “IA Explicativa” estariam a minuta elaborada pelo NIST (*US National Institute of Standards and Technology*)³¹⁴, “Quatro Princípios da Inteligência Artificial Explicativa”, submetida a consulta de *stakeholders* em 2020, com vistas a construir um entendimento global. Outra iniciativa é do *German Institute for Standardization* (DIN), elaborando um Roteiro de Padronização da Inteligência Artificial (“*IA Made in Germany*”)³¹⁵, submetido em 2019 à discussão de especialistas e partes interessadas visando identificar áreas da IA onde a padronização é mais urgente nos temas rastreabilidade, explicabilidade e confiabilidade.

Desde 2016, o IEEE³¹⁶ têm iniciativas de padrões em IA em usos como sistemas autônomos, ciências da vida e segurança cibernética, com foco na ética, transparência, privacidade de dados, explicabilidade dos sistemas e da *BigData*. Em 2020, iniciou o projeto

³¹⁴ Disponível em: <https://www.nist.gov/artificial-intelligence/ai-standards>. Acesso em 12 ago. 2022

³¹⁵ Disponível em: <https://www.din.de/en/din-and-our-partners/press/press-releases/artificial-intelligence-requires-standards-and-specifications-688888>. Acesso em 12 ago. 2022

³¹⁶ Disponível em: <https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html>. Acesso em 12 ago. 2022

de guia para arquitetura da IA explicável (XAI)³¹⁷, principalmente com usos em finanças, educação, direito, saúde, transporte e telecomunicações. Outro projeto é o Padrão para licenciamento de IA responsável³¹⁸, visando definir norma que descreva especificações de licenças de IA Responsável, definição padronizada de componentes e elementos de software da IA (código fonte e serviços) e identificação de violações e penalidades. O comitê de trabalho "ISO/IEC JTC 1/SC42"³¹⁹ foi a primeira iniciativa de comitê internacional de padrões da IA com objetivo de propor a padronização da IA; adaptar estratégia de patentes e orientar subcomitês de aplicações da IA.

Os padrões da IA, entretanto, não poderão ficar restritos apenas à explicabilidade. Padrões serão essenciais para a disseminação das tecnologias da IA, dependem de tecnologias *wireless* e seu uso em setores a jusante criará necessidades ainda maiores. Usos sofisticados como cirurgias remotas, avanços em telemedicina, comunicação massiva de máquina para máquina, realidade virtual, robôs colaborativos, veículos autônomos, fábricas, casas e cidades inteligentes, irão requerer tecnologias e padrões de interoperabilidade. A capacidade de conectar todos os dispositivos entre si, que depende da expansão de contínua interconectividade, cria valor além da inovação da IA (Ghafele, 2021). A implantação em curso do novo paradigma em conectividade móvel habilitado por tecnologias 5G e 6G será fundamental para a IA com impactos sobre a captura de valor das inovações. Tal como nas gerações anteriores, o *cluster* de tecnologias de interoperabilidade 5G, protegidas por SEPs assegurarão recursos revolucionários em termos de rapidez, latência e fluxo de dados, tornando possíveis aplicações inéditas da IA³²⁰.

Esses padrões e o sistema de patentes essenciais nas tecnologias de telecomunicações tiveram papel central em algumas indústrias, com investimentos *downstream* passando a ser feitos dentro do padrão³²¹. Um grupo seleto de firmas gigantes desse tipo de tecnologia (que hoje inclui Huawei e ZTE), construíram capacidades e portfólios SEPs, com potencial para influenciar a definição de novos padrões globais.

³¹⁷ Disponível em: <https://development.standards.ieee.org/myproject-web/public/view.html#pardetail/8406>. Acesso em 12 ago. 2022

³¹⁸ Disponível em: <http://licenses.ai/>. Acesso em 12 ago. 2022

³¹⁹ Disponível em: <https://www.iso.org/committee/6794475.html/>. Acesso em 12 ago. 2022

³²⁰ Teece (2021:28) esclarece que enquanto as quatro gerações anteriores de redes móveis usavam torres de celular com elevada extensão vertical e alto consumo energético para transmissão a longas distâncias, a geração 5G usa grande número de microcélulas menores, junto com torres de células, para criar uma cobertura de rede de altíssima velocidade (exigida, por exemplo, pela realidade virtual), atrasos mínimos (exigido em usos como cirurgia remota e outros de monitoramento remoto que demandam alta capacidade e atraso mínimo).

³²¹ Esse escopo é muito mais amplo do que se supõe, pois para alcançar benefícios dessas tecnologias, as empresas de infraestrutura precisam projetar seus equipamentos com base nelas, ser incorporadas em novos modems por fabricantes de chips e dispositivos, na atualização de provedores de serviços de rede.

Ainda que as capacidades tecnológicas chinesas sejam insistentemente questionadas³²² e a interferência governamental seja incomum no sistema de definição de padrões – desde a divulgação do *Made in China 2025*³²³ visa reorientar o crescimento chinês para bens e serviços de alta tecnologia, alcançando liderança nas novas tecnologias digitais e na definição de padrões globais conforme o *China Standards 2035*³²⁴ –, o caráter disruptivo e estratégico da tecnologia 5G leva a supor que a interferência e política governamental são hoje mais amplas no mundo, considerando a importância desses padrões/SEPs na disputa pela hegemonia global. O Relatório NSCAI (2021) não deixa dúvidas sobre as preocupações com o risco da liderança chinesa em IA e outras tecnologias emergentes. Relatório recente do governo alemão apresenta indicadores de que a liderança já é real em praticamente todos os *clusters* relevantes (EFI Report, 2022). Teece (2021) admite uma dimensão geopolítica³²⁵ associada à tecnologia 5G, pela sua importância econômica e segurança nacional, mas parece ignorar a importância dessa dimensão para outras tecnologias, como a IA.

O conjunto de tecnologias 5G autodeclaradas essenciais é gigantesco, contemplando as várias dimensões de dispositivos e infraestrutura de rede, inclusive porque “carregam” tecnologias de gerações anteriores. Em 2020, totalizavam 94 mil concessões e solicitações ativas (correspondendo a 25.700 famílias) (das quais, 64% são pedidos de patentes ainda não concedidas) de 5G SD-SEPs (sigla da expressão em inglês *self-declared SEPs*). Número que é estimado alcançar mais de 150 mil patentes até 2030.

As principais firmas detentoras de portfólios de famílias de patentes 5G SD-SEPs (pedidos e patentes concedidas) são mostradas no Gráfico 7.1. Metade dessas patentes declaradas essenciais em 5G são detidas por apenas quatro firmas: Huawei (17%), Qualcomm (12%), Samsung (11%) e ZTE (11%). No seu conjunto, há uma clara concentração de patentes autodeclaradas essenciais 5G na China (46% do total), seguida por EUA (20%), Coreia do Sul (20%) e Japão (10%). No entanto, a ponderação dessas patentes 5G pela taxa de essencialidade

³²² As desconfianças sobre as capacidades tecnológicas e patentes chinesas, em especial em IA, parecem ignorar a importância de dados massivos para a IA decorrente do gigantesco sistema de coleta e armazenamento de dados por meio do sistema chinês de *social credit score* (na verdade, um modelo de governança governamental baseada em dados, contendo dados de empresas, indivíduos, e outras entidades) e das plataformas digitais chinesas (que inclusive criaram seus próprios sistemas de classificação de confiança), que talvez constituam o maior BigData do mundo.

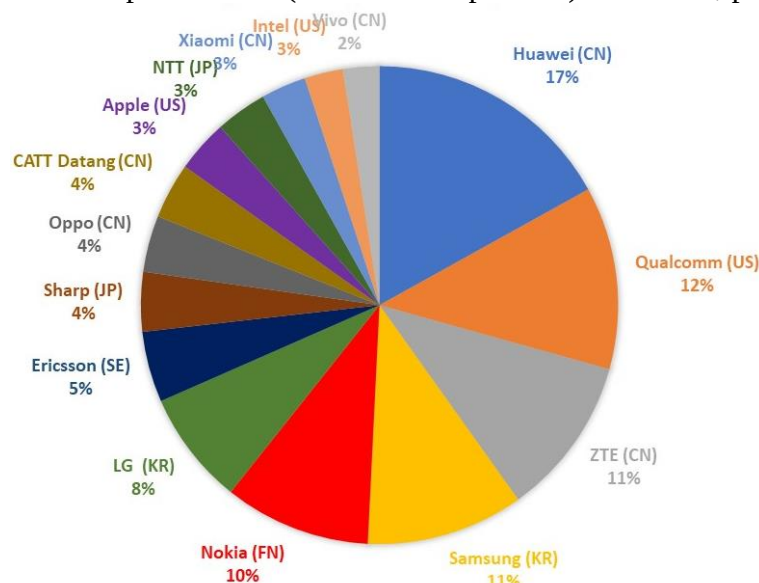
³²³ Disponível em: <https://www.china-briefing.com/news/made-in-china-2025-explained/>. Acesso em 20 ago. 2022

³²⁴ Disponível em: <https://www.china-briefing.com/news/what-is-china-standards-2035-plan-how-will-it-affect-emerging-technologies-what-is-link-made-in-china-2025-goals/>. Acesso em 20 ago. 2022

³²⁵ Essa questão geopolítica e a ascensão da China parecem estar merecendo grande atenção de David Teece, tendo sido abordada em trabalhos recentes (Teece 2022a; 2022b), inclusive em coautoria sobre o problema dos semicondutores (Guile et al., 2022).

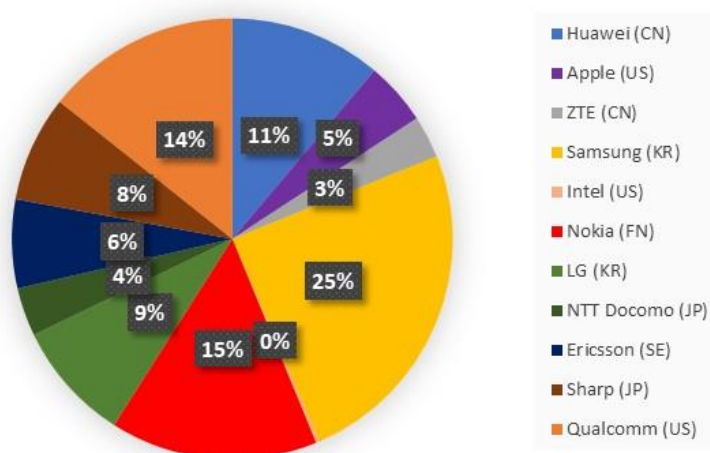
histórica – patentes com efetiva chance de se tornar o padrão - modifica a distribuição geográfica, conforme o Gráfico 7.2, que mostra a liderança da Samsung (25% do total), Nokia (15%), Qualcomm (14%), Huawei (11%), LG (9%), Sharp (8%), Ericsson (6%) e ZTE (3% do total). Esses dados também mostram uma maior concentração do portfólio – as três principais empresas com 54% do total das SEPs – e a maior importância geográfica da Coreia do Sul (43%), seguida por EUA (25%), China (18%) e Japão (14%).

Gráfico 7.1 – Famílias de patentes 5G (concedidas e pedidos) no mundo, principais empresas



Fonte: Elaborado a partir de IAM-Media/Iplytics Platform³²⁶.

Gráfico 7.2 – Famílias de Patentes 5G ponderadas pela taxa de essencialidade, principais empresas



Fonte: Elaborado a partir de IAM-Media/Iplytics Platform (Acesso em: 22 ago.2022)

³²⁶ <https://www.iam-media.com/article/who-leading-the-5g-patent-race-patent-landscape-analysis-declared-seps-and-standards-contributions>. Acesso em: 22 ago.2022

Teece (2021) argumenta que a comparação de portfólios de patentes de 5G só é possível se a qualidade das patentes for equivalente e se as jurisdições de licenciamento forem semelhantes, lembrando que estatísticas de patentes podem ser distorcidas pela construção de grandes portfólios de patentes apenas por motivação estratégica (semicondutores na década de 1990), por subsídios de governos no sentido de promover patentes de invenção e “invenção de patentes” (de baixa qualidade), como supõe ser o caso das patentes chinesas. Até mesmo citações de patentes, que indicam invenções prévias nas quais se baseiam, podem refletir conquistas tecnológicas, não valor comercial, também decorrentes de estratégias empresariais (citações estratégicas).

A partir da definição do *design* dominante, o êxito da inovação em IA passaram a depender de ativos complementares de firmas orquestradoras de plataformas digitais³²⁷, em particular conjuntos de dados integrados à tecnologia focal do provedor da plataforma, serviços de computação na nuvem, sistemas operacionais (como iOS da Apple e o Android da Alphabet), serviços como aplicativos de provedores de conteúdo das plataformas digitais (em geral, firmas *startups*), e também tecnologias de interoperabilidade (por exemplo, tecnologias 5G, protegidas por SEPs e licenciadas por empresas especializadas Samsung, Nokia, Qualcomm, Huawei, LG, Sharp, Ericsson e ZTE) e provedores de *chips* (adquiridos de terceiros mas, de modo crescente, desenvolvidos internamente pelas *BigTechs*, inclusive em parcerias como OpenAI/Microsoft ou firmas independentes, como Nvidia). O licenciamento das tecnologias de interoperabilidade protegidas por patentes essenciais de padrões, necessárias para a conexão das tecnologias digitais, absorvem uma parcela importante dos lucros com a inovação, considerando que mesmo com taxas de licenciamento baixas, o volume de SEPs envolvidas significa usualmente somas milionárias. Aliás, pela lógica do modelo PFI, os maiores riscos de gargalos/ competidores potenciais estariam nos segmentos de *chips* e computação na nuvem, onde firmas incumbentes de alta tecnologia começam a investir.

A Figura 2.5 do ecossistema de inovação da IA apresentada no Capítulo 2 identificou, em parte, os ativos complementares e a forma como são acessados (na parte inferior de cada retângulo, que mostra seus atuais provedores). Alguns deles, como os amplos conjuntos de dados (BigData), são ativos integrados às atividades das *BigTechs* e poderiam ser considerados parte da inovação focal. Ativos (na verdade, inovações) complementares podem ser integrados

³²⁷ Há muitas definições de plataformas digitais. Por conveniência metodológica usamos a mesma de Teece (2018a), como sendo qualquer combinação de hardware e software que estabelece padrões, interfaces e regras que permitem a empresas adicionar valor e interagir entre si e/ou com usuários (incluindo implementadores *downstream*), numa relação de “coopetição” (cooperação no curto prazo e competição no longo prazo) (Petit e Teece, 2020).

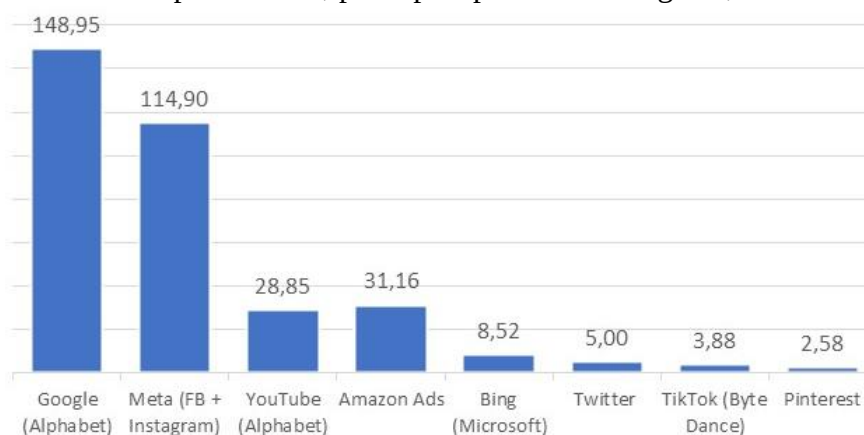
pela firma inovadora (como no caso do sistema operacional Android, que foi adquirido pelo Google possibilitando competir com a Apple na venda de aplicativos) ou adquiridos de terceiros por via contratual, como conteúdos, jogos, de vendas online da Apple Store e Google Play, além de capacidade computacional, adquiridos de terceiros durante a fase da IA preditiva. Também é possível que as bibliotecas de treinamento de modelos de aprendizado de máquina *free/freemium*, permitam a identificação de novos ativos, como novos dados acessados, melhorias nos modelos de IA e novos serviços potenciais, que possam vir a constituir novos ativos complementares, agregadores de valor às plataformas. Dessa forma, ainda que as plataformas digitais sejam contextos de captura de valor de múltiplas-inovações, ecossistemas mais amplos respondem pela criação de valor da inovação.

Paredes-Frigolett e Pyka (2022) consideram que as plataformas digitais mudaram a maneira como as firmas abordam o problema da criação e captura de valor, provocando mudanças no modelo padrão de como lucram com a inovação. As iniciativas de diversificação das orquestradoras de plataformas são estratégias de extensão de seus ecossistemas de inovação. No entanto, pelo menos na fase do desenvolvimento da IA, o modelo PFI é ainda capaz de explicar como as plataformas lucram com a inovação (e não uma relíquia do passado como supõem os autores), apenas exigindo uma visão mais flexível sobre a noção de ativo complementar e de como são acessados na economia digital. A dicotomia estanque “desenvolver *versus* adquirir”, típica da economia industrial, é flexibilizada por meio do debate antigo entre controle e propriedade e totalmente desmontada desde as GVCs.

As transações que envolvem ativos complementares nas plataformas digitais são baseadas em relações assimétricas, mas argumenta-se que não mais assimétricas do que muitas transações na economia industrial. Assemelham-se às parcerias de firmas farmacêuticas incumbentes e startups de biotecnologia que revigoraram a inovação em medicamentos a partir da mudança de paradigma, que prescindiram temporariamente tanto do controle quanto da propriedade, apenas passando a envolver participações acionárias (inicialmente minoritárias) nos estágios mais avançados (e promissores) da inovação, após a fase do “vale da morte” e de demonstração quando passariam a atrair concorrentes diretos ou do *venture capital*, até finalmente sua aquisição integral quando comercialmente viáveis e estratégicas. Essa foi a trajetória das pioneiras em biotecnologia farmacêutica, Genentech e Genzyme - até serem adquiridas pelas grandes farmacêuticas (a primeira pela Roche em 2009, enquanto a segunda pela Sanofi em 2011), restando apenas a Amgen autônoma entre as pioneiras – padrão que se consolidou com o aumento dos custos da inovação em medicamentos.

O conceito de apropriabilidade é distinto, mas se reflete na apropriação efetiva de valor³²⁸, razão para apresentar alguns fatos estilizados sobre a efetiva apropriação pelas BigTechs líderes da IA generativa. O modelo de receita pela monetização de dados pela publicidade online, ao contrário do senso comum, não predomina no setor. Entre as *BigTechs*, apenas Meta e Alphabet/Google dependem profundamente da receita de publicidade online, que representou 98% e 81,3%, respectivamente, da receita dessas firmas em 2021 (Gráfico 7.3). Apesar da importância das receitas da AppStore, o modelo de negócios da Apple ainda depende principalmente da venda de dispositivos de hardware; a Amazon depende das receitas de vendas *online* - e cada vez mais da computação na nuvem -; enquanto a Microsoft conseguiu alcançar a liderança em receitas de serviços de computação na nuvem, que suplantaram até mesmo o Windows³²⁹.

Gráfico 7.3 - Receita publicidade, principais plataformas digitais, US\$ bilhões 2021



Fonte: <https://fourweekmba.com/google-revenue-breakdown/>

Os anúncios online são a fonte quase exclusiva de receita da Meta, ao passo que a Alphabet tem buscado reduzir a dependência, possivelmente diante do escrutínio social e regulatório sobre o mecanismo que viabiliza serviços gratuitos aos usuários de plataformas em

³²⁸ Hurmelinna-Laukkanen e Yang (2022) propuseram o enquadramento conceitual que distingue apropriabilidade e apropriação. A apropriabilidade seria o potencial de se beneficiar de uma inovação, que se acumula por meio de instrumentos de apropriabilidade: isolando mecanismos de apropriabilidade e ativos complementares, enquanto a apropriação é a realização desse potencial, que se manifesta em benefícios privados e sociais quando os instrumentos são empregados em processos de exclusão, alavancagem ou divulgação.

³²⁹ As receitas da Amazon são provenientes de vendas *online* próprias (cerca de 50% da receita total em 2021) e de terceiros (22%), além da AWS (que engloba receitas da computação na nuvem) (13,2%) e da receita de serviços de assinatura prime – embora queda recorrente na receita de venda em lojas físicas. Na Microsoft, produtos de servidor e serviços em nuvem já representam a principal fonte de receita (34% do total, em 2022), suplantando a receita de outros produtos e serviços (23%) (em 2019, ambas representavam cerca de um quarto da receita) e do windows (12% em 2022, frente aos 18% em 2019), seguidos por receita de jogos, de assinaturas premium na plataforma de rede social (LinkedIn) voltada para relações profissionais (talvez não um terreno fértil para publicidade online), publicidade no Bing (6ª no ranking das receitas), venda de dispositivos, outros serviços e outros. Na Apple, mais de três quartos das receitas são provenientes da venda de dispositivos – onde o carro chefe é o iPhone (praticamente metade da receita) seguido pelo ApplePay (23,6%), sistema de pagamento por aproximação e carteira digital que permite aos usuários pagamentos usando dispositivos da Apple, além do iMac (8,9%) e Apple Watch (9,4% e crescente).

troca dos seus dados. Os anúncios *online* (do Search Engine, Youtube e de membros das plataformas de publicidade - AdMob, AdSense, Ad Manager - do Google) são a principal fonte de receita do grupo, embora a contribuição das receitas não-publicidade tenha aumentado de 3,2% para 18,7%, entre 2009 e 2021³³⁰. Contribuíram para isso o Google Play – viabilizado pelo Android, sistema operacional em código aberto desenvolvido para dispositivos móveis pela *startup* Android Inc. (para competir com o Symbian e o Windows mobile) adquirida pelo Google, se tonando um competidor a altura do iOS da Apple -, registrado no demonstrativo de resultados na rubrica “Google Others” (cuja participação total é de 11% na receita total) e o rápido aumento com o Google Cloud³³¹ (7,5% da receita). As principais apostas do Google/Alphabet ainda não têm papel destacado na geração de receita, registradas de forma agregada na rubrica “Other Bets” (apenas 0,8% da receita total em 2021)³³².

As plataformas digitais permitem que outros *stakeholders* nelas disponibilizem serviços *online* pagando taxas (em geral elevadas, estimadas em 30%, mas que pela pressão dos reguladores antitruste vem sendo reduzidas para empresas de menor porte), como ocorre com os jogos e apps disponíveis no Google Play e na Apple Store, gerando elevadas margens de lucros. A margem operacional estimada das lojas *online* seria de 78% no caso da Apple e 62% do Google (frente a margem de todas as operações, de 35% e 31%, respectivamente) em virtude de seus baixos custos de manutenção (The Economist, 2022), respondendo por parcela expressiva e crescente das receitas das plataformas.

São inovações que contribuem para a maior criação e captura de valor da inovação focal, sem que o titular da plataforma precise desenvolvê-los diretamente. Nos contratos entre o dono da plataforma, como Apple, e empresas participantes, constam cláusulas onde estas últimas concordam em, a qualquer momento, o titular da plataforma possa optar por oferecer o mesmo serviço diretamente, como visto. Isso subverte a hipótese implícita de Paredes-Frigolett e Pyka (2022) da importância das capacidades dinâmicas das partes interessadas que participam, uma vez que esse tipo de cláusula sugere que a Apple já possui tal capacidade (e pode decidir

³³⁰ Disponível em: <https://www.cnbctv18.com/information-technology/google-is-gradually-reducing-its-reliance-on-advertising-1063271.htm>. Acesso em: 12 mar. 2023

³³¹ A pioneira entre as plataformas de computação na nuvem foi a Amazon, lançada em 2006, seguida pelo Google (2008), Alibaba Cloud (2009) e Microsoft, que introduziu o Azure em 2010.

³³² *Other Bets* são investimentos estratégicos e ambiciosos de longo prazo da holding Alphabet. Destacam-se a X Developments LLC (antes Google X), unidade de P&D “*moonshot*”, criada em 2010 e responsável por projetos secretos de P&D, incluindo empresas nascentes adquiridas, que esteve à frente do Google Glass, Loom, Wing, projetos de energia (Dandelion, Malta, Makani), além de Waymo (veículos autônomos) e Intrinsic, muitos depois convertidos em empresas independentes, em que se destacam a Verily Life Sciences, DeepMind (hoje incorporada ao Google), Calico, Access, Nest Labs (automação residencial) e até os mencionados fundos próprios de venture capital, importante mecanismo para identificação de negócios e tecnologias promissores.

implementá-la a qualquer momento).

Também fornecedores terceirizados que pagam taxas pela venda online no Marketplace da Amazon, simultaneamente competindo com produtos próprios do negócio de varejo da Amazon – modelo de negócios sob escrutínio antitruste, onde há até propostas de divisão dos negócios de varejista e provedor de infraestrutura (em especial, computação em nuvem) da empresa.

Estes seriam casos em que há simultaneamente integração/contratação de ativos complementares, exigindo um olhar mais flexível do PFI; flexibilidade esta que exige uma visão dinâmica, uma vez que antigos devs. (complementadores) da plataforma, são muitas vezes por ela adquiridos (“cooperar para depois absorver” nos termos de Paredes-Frigolett e Pyka, 2022, ou pela competição entre plataformas, como parece ter sido o caso da aquisição pelo Google do YouTube, que antes usava a plataforma do MySpace), mas que não parece comprometer a abordagem do PFI, apenas exigindo maior dinamismo do modelo. No entanto, Paredes-Frigolett e Pyka (2022) têm razão ao afirmar que estratégias de integração podem sinalizar negativamente os demais *stakeholders* e prejudicar a governança do modelo de negócios, levando a perda de confiança e saída de provedores – o que afetaria uma das principais vantagens competitivas da plataforma que é o efeito de rede – embora os autores pareçam superestimar a capacidade das empresas menores “andarem pelas próprias pernas”, independente do ambiente de plataforma.

As cada vez mais frequentes disputas judiciais são apontadas como evidência da ineficácia do regime de apropriabilidade. No entanto, muitos litígios relacionam-se a práticas anticompetitivas como PlayStore (ou AppStore) contra devs/provedores de aplicativos, em particular jogos, como o caso paradigmático Google vs. Epic Games (aplicativo de jogos Fortnite), Activision Blizzard (videogames), Riot Games (produtora do *League of Legends*) e outros provedores de aplicativos. De fato, as disputas têm envolvido cifras milionárias em longas disputas, muitas vezes para evitar que fornecedores de ativos complementares se tornem competidores diretos e ameaça real para a sustentabilidade do modelo de negócios (por exemplo, com a intenção da gigante dos videogames Activision lançar sua própria loja de aplicativos – mas que acabou adquirida pela Microsoft). Em virtude dos *network effects* típicos das plataformas, a saída de provedores importantes de ativos complementares pode resultar em perdas significativas de usuários da plataforma. Em grande parte, o tema parece ganhar maior dimensão em função do ambiente restritivo (regulatório e antitruste) que vem se formando no mundo contra as plataformas digitais.

Mesmo assim, as recentes disputas com 24 grandes devs. (Infomoney, 2022) podem ser casos isolados de provedores que, de fato, ganharam relevância no mercado, mas que poderiam ser vistos como insignificantes no universo de centenas de milhares de provedores de ativos complementares das plataformas?³³³ Disputas efetivamente importantes parecem ser casos como o da Apple, que recentemente implantou o IDFA (sigla de *Identifier for Advertisers*) no iOS, sistema operacional da Apple (ou o AAID – *Android Advertising ID*, do Google) para permitir que usuários rejeitem a coleta de seus dados por aplicativos, o que tem significado um bloqueio claro ao Facebook, evidenciando a competição entre plataformas, bem como litígios relacionados às tecnologias de interoperabilidade e patentes essenciais de padrões.

De fato, o orquestrador da plataforma não precisará, em termos estratégicos, integrar ativos complementares enquanto tiver um fluxo contínuo de inovações sem precisar investir diretamente no seu desenvolvimento. O poder de mercado parece, assim, se expressar sobre o modelo de plataforma em seu todo, não sobre produtos/serviços específicos, com exceção daqueles efetivamente estratégicos – o mecanismo de busca, o sistema Android, entre outros, recorrendo mais uma vez ao exemplo do Google – e a *Alphabet* concentrando investimentos em P&D nas suas apostas (*Bets*) efetivamente disruptivas.

A *BigData* é efetivamente um exemplo paradigmático de ativo complementar da IA, que possibilita vislumbrar muitas oportunidades novas na economia digital e a exploração simultânea de várias trajetórias de inovação e diversificação das firmas, mantido em sigilo. O foco nos dados parecia estar presente na época da criação do Google, como o mencionado título do primeiro pedido de patente, deixando clara sua natureza de empresa *data-driven*. Um conjunto de dados mantido cativo, não disponibilizado nem nas bibliotecas gratuitas, nem para parceiros nas plataformas, em volume, variedade e qualidade.

A estratégia relacionada aos ativos complementares é especialmente importante na fase paradigmática, que no caso da IA acredita-se ter como *design* dominante os modelos de *deep neural networks*, mas que não parecem replicáveis/imitáveis em função do controle das inovações focais baseados em amplos conjuntos de dados de treinamento dos modelos organizados pela firma inovadora (que podem, portanto, ser vistos como parte da inovação da IA). A partir daí, é mais fácil replicar a tecnologia, considerando a hipótese do PFI de regimes de PI usualmente fracos, com o sucesso da inovação passando a depender mais das condições

³³³ É expressivo o número de “devs.” atrelados a essas plataformas, estimados em 800.000 a 900.000 devs que oferecem 2,5 a 3 milhões de aplicativos no Google Play e 500.000 a 600.000 devs provendo 1,8 milhões de aplicativos na AppStore, em 2020 (The Economist 2022a). As relações intrincadas não param por aí. O uso do mecanismo de busca do Google, opção padrão na maioria dos dispositivos Apple, envolve pagamentos pelo Google entre US\$ 8 e US\$ 12 bilhões ao ano (2 a 3% da receita total da Apple).

de acesso aos ativos complementares especializados e co-especializados. Investir nesses ativos envolve irreversibilidades e não podem ser facilmente acessados por contrato pelos altos riscos para o investidor. Nos regimes de apropriabilidade forte, o retorno da inovação é possível mesmo prescindindo de deter diretamente ativos complementares, com a estratégia de contratar, ao invés de integrar, se mostrando adequada uma vez que esses ativos estão disponíveis de forma competitiva. Em casos de regimes fracos, os inovadores são pressionados a definir estratégias de internalização dos ativos, pois o sucesso da inovação passa a depender do seu controle.

A *BigData* serve para treinar modelos de IA, mas não prescinde dos modelos *per se*, que precisam ser desenvolvidos. Quanto custaria para uma Alphabet/Google desenvolver tudo sozinha... e para quê, se pode lucrar com a inovação de terceiros? Para alguns, isso significa que o Google compartilha uma parte de seus lucros potenciais com inovações levadas ao mercado; para outros, é se apropriar de parte dos lucros de terceiros por meio das taxas cobradas pelo uso da plataforma, além da captura de dados e melhorias nos modelos. Nesse sentido, mais importante do que um possível monopólio de mercado, é seu monopólio sobre o conhecimento, recorrentemente reafirmado por Cecília Rikap em seus inúmeros trabalhos, mas que aqui se considera prescindir de instrumentos formais de PI.

As escolhas estratégicas mais complexas se relacionam aos ativos complementares cruciais para o modelo de negócios e para a inovação focal, como a computação na nuvem. As pioneiras Amazon e Alphabet/Google, logo seguidas pela Microsoft, sugerem algum grau de concorrência entre as plataformas, e não apenas cooperação, principalmente no caso de ativos de natureza infraestrutural. As plataformas na nuvem são mecanismos de prestação de serviços (SaaS, IaaS, DaaS, etc.) e uma fonte crescente de receita. Foi crucial para a reorientação estratégia da Microsoft rumo a inovação da IA generativa, depois de um relativo atraso frente às demais *BigTechs* pela posição confortável no mercado pelo sucesso do Windows.

Diferentemente das plataformas de transação, a computação na nuvem parece um ativo fundamental para acesso a dados (sobretudo empresariais) e que terá papel crucial na fase da implementação da IA em setores *downstream*. Não obstante, firmas incumbentes do setor de tecnologia também começam a investir em *cloud computing*, desde IBM, HP, Qualcomm, Oracle, entre outras, além das gigantes BigTechs Chinesas Alibaba e Tencent.

Pisano (2006a) lembrava que a digitalização, ao alterar o lócus central da inovação para componentes digitais (*chips*), afastou os lucros da inovação dos inovadores de produtos finais em direção aos fornecedores (inovadores) de *chips* especializados, capturando parcelas crescentes do retorno da inovação. Nas TICs, o *design* do chip e o software incorporado teriam

se tornado o principal ativo co-especializado. No caso da IA, o ativo complementar crucial volta a ser o hardware. Assim, daqui para frente, mais do que em qualquer outro momento da história, haverá uma diferença entre produzir “*potato chips*” e “*computer chips*” (Dosi et al., 2021:3). Além da insuficiência global da oferta e da concentração do mercado da IA há uma dimensão geopolítica relacionada a essa concentração. O *market share* de uma única empresa, a taiwanesa TSMC, é de 60%, seguida pela coreana Samsung, com 13%, ainda que contrabalançado pela holandesa ASML que detém 90% do mercado de sistemas litográficos exigidos para criar os circuitos dos *chips* de computador. São esses dilemas que motivam as recentes políticas industriais dos Estados Unidos e União Europeia e das discussões conjuntas com a IA no fórum bilateral US-EU TTC.

Além disso, existem problemas de ineficiência para chips para soluções baseadas em IA, que estariam estimulando esforços de inovação de produtores no sentido de projetar uma nova arquitetura de processador para as necessidades de IA. Prytkovae e Vannuccini (2021) vislumbram dois cenários possíveis: o surgimento de um novo projeto dominante na forma de um “chip plataforma” composto por núcleos heterogêneos; ou a fragmentação do mercado da indústria de semicondutores em sub-mercados com *chips* dedicados. Foi visto que firmas de IA já começaram a integração/diversificação com vistas ao desenvolvimento de *chips* e parcerias como Meta/Nvidia, Microsoft/OpenAI, e do consórcio Mobileye, Intel, Qualcomm e Samsung para *chips* em ConvNets. Da mesma forma, a China vem enfrentando o desafio de desenvolvimento de sua indústria de *chips* para IA em meio a restrições de acesso às tecnologias dos EUA (Ernst, 2020).

A competição típica do estágio paradigmático parece evidente, como do Google que com a aquisição do Android conseguiu organizar sua *Play Store*, competindo e hoje superando em vendas e número de devs. da *App Store*. De fato, as plataformas parecem mesmo coexistir em um ambiente de permanente coopetição, caminhando juntas na evolução do conhecimento (participando de redes de pesquisa, publicando em co-autoria, criando bibliotecas gratuitas de treinamento, entre outras), mas competindo eventual ou potencialmente em segmentos específicos, considerando a noção de mercado relevante global como é a atuação das plataformas. Casos notórios como o chinês TikTok (da ByteDance), que em curto espaço de tempo superou o Instagram/Meta, alcançando 1 bilhão de usuários, além de concorrentes diretos em vídeos curtos do Snapchat, do Kwai (já entre os apps mais baixados no Google Play e App Store) e do Zynn, que pertence à *startup* chinesa Kuaishou e tem a Tencent entre seus investidores.

O surgimento dos novos modelos de IA generativa por diferentes provedores pareceu reabrir, entretanto, a competição pré-paradigmática. Exatos dois meses após o lançamento do ChatGPT3³³⁴ pela OpenAI/Microsoft, foi seguida pelo ERNIE da chinesa Baidu, pelo BARD do Google e logo depois pelo Tongyi Qianwen da Alibaba, entre outros. E não parou por aí. A competição é obstinada pela demonstração do potencial dos modelos por trás desses *chatbots* da IA generativa, por incumbentes detentores e empresas nascentes.

Pisano (2017), oferece novos *insights* interessantes, para o regime de apropriabilidade. Foi visto que apesar da conexão entre a natureza da tecnologia e a arquitetura da indústria, esta pode ser modificada a partir de estratégias empresariais ou em decorrência de fatores institucionais ³³⁵, como a modularização da indústria de TICs. De fato, indústrias modularizáveis (equipamentos eletrônicos, software, automobilística, aeronáutica) possibilitam uma divisão de trabalho entre organizações especializadas em partes do sistema, desenvolvidas separadamente e depois montadas, desde que existam interfaces e padrões bem definidos e PI codificada com direitos precisos (Pisano 2006b). No entanto, a IA - e algumas outras tecnologias digitais emergentes - parece cada vez mais fugir à essa regra de segmentação estanque e de simples “montagem” *a posteriori*. O desenvolvimento e funcionamento em redes envolve uma parcela crescente de inovações colaborativas, já identificadas na fase da IA preditiva nos ecossistemas de plataformas digitais, que tende a se aprofundar, considerando as tendências relacionadas aos *chips*, a computação na nuvem e as inovações em indústrias *downstream*, na fase da implementação da IA.

Nesse sentido, a construção de capacidades dinâmicas continuam centrais para sustentar estratégias empresariais e apropriabilidade da inovação, mas não parece possível que isso ocorra da forma independente, isolada e estanque sugerida por Pisano (2017). A noção genérica de capacidades é incompleta para explicar a diversificação das empresas, inclusive porque a competição se dá cada vez mais na criação dessas mesmas capacidades. De fato, o estoque de capacidades acumuladas pelas firmas *BigTech* é resultado de investimentos passados e a decisão de mudar o portfólio futuro de capacidades a partir da construção e investimentos (em P&D, em adoção das tecnologias, com experimentação formal e informal) que envolvem incertezas pelo lado da demanda e da oferta. A dimensão crítica na estratégia de construção de capacidades pela empresa é escolher entre aprofundar capacidades existentes (como as capacidades do Google em IA para automatizar automóveis) ou ampliar capacidades fora do seu repertório

³³⁴ Em apenas cinco dias o ChatGPT3 alcançou 1 milhão de usuários, que o Facebook só conseguiu em 10 meses.

³³⁵ As recentes iniciativas regulatórias da economia digital em curso sobre usos da IA e, principalmente, as mudanças na abordagem antitruste poderão ter efeitos similares em termos da arquitetura da indústria.

(como investir em capacidades para projetar/produzir um automóvel), como mencionado pelo autor. O aprofundamento de capacidades tem como elemento crítico o aprendizado acumulado pela experiência, ao passo que a ampliação de capacidades envolve investimentos em P&D ou aquisições.

7.2. A apropriabilidade da inovação da IA Generativa

O êxito da inovação da IA (preditiva) em serviços de plataformas resultou de uma “combinação” de complementaridades (cooperação), bem como competição nos ecossistemas de plataformas, modelo de negócios crucial para a criação e captura de valor das inovações da IA preditiva, embora as inovações da IA generativa exijam uma abordagem distinta das plataformas, privilegiando aquelas de natureza de infraestrutural ou híbrida.

Na fase da IA preditiva o regime de apropriabilidade da inovação dependeu fortemente do modelo de negócios de plataformas digitais, comandado pelo conjunto reduzido de empresas líderes, as *BigTechs* americanas e chinesas, às quais se atrelavam empresas “auxiliares” que necessitavam entrar na plataforma para poder acessar o mercado, ao custo de compartilhamento de renda e de conhecimento com os controladores das plataformas, mediante o pagamento de taxas, aprisionamento e risco permanente da competição com outros devs. e com os próprios orquestradores.

A divulgação em artigos científicos sobre as novas arquiteturas, invenções algorítmicas, técnicas de treinamento e ajuste fino dos novos modelos e o padrão de disseminação *open source* pela indústria, foram vitais para o avanço do conhecimento e a explosão de modelos de IA generativa, muitos deles modelos derivados, menores e com custos imensamente inferiores, além de APIs, por concorrentes reais e potenciais.

Isso resultou na reorientação estratégica das empresas líderes. Primeiro, recuando na divulgação e “abertura” de detalhes técnicos dos modelos, com a exceção notável da Meta AI. Segundo, pela ampliação do tamanho dos modelos a escalas estratosféricas, forçando a dependência (computacional e financeira) pelas empresas emergentes das infraestruturas de computação em nuvem do núcleo infraestrutural das *BigTechs*, como forma para se manter na corrida da IA generativa. Terceiro, com empresas líderes abraçando o discurso regulatório, ditando, elas próprias, seu tom e “as regras do jogo”, com exceção da Meta AI, a menor das *BigTechs*, que em 2024 decidiu ingressar no núcleo infraestrutural. Quarto, pelas muitas incertezas que rondam os direitos proprietários da IA, relacionadas à elegibilidade de invenções que envolvem ideias abstratas (algoritmos) e aos requisitos formais de patenteamento. Por fim, criando relações intrincadas com as empresas emergentes.

À medida que os muitos pedidos de patentes começarem a ser concedidos, pode ser grande o potencial de litígio, com risco de direitos invalidados e uma insegurança jurídica que enfraquece o regime de proteção legal e, certamente, a apropriabilidade quando chegar o momento crucial de implementação da IA em setores *downstream*.

A ampla distribuição de modelos *open-source* parece representar uma mudança no paradigma de apropriabilidade, com empresas abrindo estrategicamente o código-fonte de forma a promover a rápida disseminação das novas tecnologias de IA, ampliando seus ecossistemas, obtendo contribuições da comunidade para aperfeiçoamento dos modelos e capturando efeitos de rede, que lembram as antigas estratégias de empresas no software de código aberto. As empresas líderes abraçaram o *open source*, que passou a constituir um importante instrumento de competição estratégica, um quasi-DPI, disponibilizando modelos para a comunidade de devs. nas bibliotecas de treinamento da IA preditiva e, posteriormente, para o público em geral com a IA generativa. O *open source* na IA, mais do que no software convencional, se transformou em um componente crucial das estratégias empresariais das empresas líderes, mesmo que na forma de “*open for business*”.

O modelo institucional híbrido decorre de estratégias empresariais e se ampara na complexidade da inovação em IA que contempla códigos de computação, mas também extensos e complicados detalhes técnicos que progressivamente se revestem de pseudo direitos proprietários conjugados a segredos comerciais, além de ativos complementares especializados progressivamente tornados proprietários, integrados ao *core business*. Parcela importante da tecnologia expressa nesses detalhes técnicos dos modelos de IA generativa, começaram a ser progressivamente omitidos nos artigos científicos que acompanham seu lançamento e em dispositivos incluídos em alguns acordos internacionais de comércio.

Ou seja, os avanços recentes da IA não se resumem apenas ao código de computação, mas aos amplos e complexos detalhes técnicos que sustentam a performance extraordinária dos modelos. Incluem arquiteturas, ajuste fino, parâmetros, *prompts*, dados de treinamento e hardware, que progressivamente se revestem de grande sigilo, inclusive de partes dos códigos. Alguns aspectos técnicos são divulgados em artigos científicos publicados concomitantemente ao lançamento dos modelos disponibilizados, alguns de forma gratuita, tendo por trás licenças de código aberto, mas com informações relevantes frequentemente omitidas, mantidas em sigilo.

Embora a maioria dos autores privilegie os aspectos técnicos e jurídicos, existe uma dimensão econômica insuficientemente tratada. O código aberto da IA parece se tornar a espinha dorsal das infraestruturas digitais modernas, com impactos potenciais na dinâmica do mercado. Prática usual do núcleo infraestrutural das *BigTechs* é dar acesso a parte substancial

de suas patentes como forma de atrair usuários para seus serviços na nuvem, como a Microsoft³³⁶.

Uma vez que o ponto nevrálgico do sucesso da IA incorporou, além das inovações algorítmicas e amplos conjuntos de dados, a escalabilidade e o hardware nos modelos generativos, mudando os determinantes da inovação, o futuro de emergentes notórias que acompanharam o avanço dos modelos em tamanho é incerto. Dependerá da capacidade de resistirem à voracidade das *BigTechs*, às quais se acoplaram computacional e financeiramente, de encontrarem uma aplicação “matadora” para a IA, como foi o sistema operacional Windows, e de uma inovação organizacional em um modelo de negócios que viabilize a monetização e apropriação efetiva da inovação tecnológica.

A perpetuação da liderança das empresas incumbentes, por seu turno, depende da manutenção das plataformas na nuvem como modelo de negócios para a implementação da IA, ainda incerta diante da crescente pressão de reguladores, que poderá resultar em desmembramento de empresas e do negócio e promover uma nova alteração da arquitetura da indústria e do padrão de competição. Mesmo que a concentração de ampla capacidade computacional em um punhado de empresas crie resistências da política antitruste pelos riscos de práticas anticompetitivas, as questões da oferta, reestruturação de cadeias de valor e desafios geopolíticos em semicondutores e chips parecem temas igualmente urgentes e cruciais nas agendas dos governos.

A literatura considera fundamental para a captura de valor da inovação a definição de um *design* tecnológico dominante. Este padrão não está estabelecido na IA generativa. Vale lembrar que foi definido na IA preditiva quando os modelos de redes neurais profundas foram desenvolvidos de forma “coopetitiva” pelos principais *players* do setor que, junto ao exitoso modelo de negócios de plataformas digitais, viabilizou a implementação dessa tecnologia em segmentos dos mercados digitais, com as empresas especializando-se e tornando-se dominantes em aplicações da IA em segmentos de mercado, alcançando meios para se expandirem em outros mercados. A Amazon domina o mercado de varejo *online* e a computação em nuvem, nesta última sob pressão permanente da Microsoft e Google; Meta e Google dominam mercados vinculados à receita de publicidade *online*; Apple dita o desenvolvimento de aplicações nos seus próprios dispositivos; e Microsoft opera grandes pacotes de produtos atrelados a sistemas operacionais. Há poucos pontos de atrito entre as *BigTechs* – como Apple Store e Google Play,

³³⁶ Disponível em <https://azure.microsoft.com/en-us/solutions/iot/ip-advantage-program/#advantage-faq>. Acesso em 05 set. 2024

o Google Search Engine e rivais menos relevantes dos mecanismos de busca online como Microsoft, mas que prometem grandes disputas no mercado da computação na nuvem.

Ferràs-Hernandéz, Nylung e Brem (2023) consideram que ainda não ocorreu a definição de um *design* dominante da IA, com o momento atual caracterizado pela disputa entre as principais empresas inovadoras. Aqui, entretanto, considera-se que a acirrada rivalidade atual se limita a IA generativa por meio da corrida pelo lançamento de novos e versáteis modelos. Com isso, a definição do *design* tecnológico dominante da IA generativa parece menos “coopetitiva” do que na IA preditiva ou talvez não haja nem mesmo um *design* dominante, como parece acreditar Mark Zuckerberg, CEO da Meta, para quem não deverá haver uma única IA, muito menos sob controle de poucas empresas (Perez, 2024).

A disputa pelo *design* tecnológico dominante ocorre com as empresas tentando se beneficiar das vantagens do pioneiro, por meio do lançamento cada vez mais rápido de modelos baseados em distintas arquiteturas e inovações algorítmicas, o que mostra o estágio ainda em evolução das tecnologias da IA generativa, com uma disputa acirrada entre incumbentes e um número não desprezível de empresas emergentes. Em mercados que dependem de efeitos de rede, as empresas buscam ser as primeiras a lançar um produto/serviço, com vistas a consolidação e fidelização da demanda, visando dificultar a entrada de potenciais concorrentes, com o *open source* como a ferramenta para sustentar tal estratégia, atraindo usuários, até mesmo evitando que concorrentes patenteiem a tecnologia.

Difere do padrão usual retratado na abordagem do PFI, em que a rivalidade pelo *design* dominante se dava de forma competitiva entre os *players* e precedia as escolhas estratégicas relacionadas à obtenção dos ativos complementares, na IA generativa a competição nesses dois níveis parece ocorrer de forma simultânea, envolvendo padrão tecnológico, hardware e modelo de negócios.

Os dois ativos complementares essenciais para a criação e captura de valor das inovações de redes neurais profundas são massivas bases de dados e elevada capacidade computacional, constituem as principais barreiras à entrada no setor. Na fase da IA preditiva, a maioria das empresas líderes estruturou amplos conjuntos de dados proprietários, alimentados pelo fluxo contínuo obtido por meio de seus produtos e serviços de plataformas digitais de transação, que continuam alimentando as bases da IA generativa.

As inovações algorítmicas se tornaram acessíveis a organizações nascentes, de menor porte, pela ampla divulgação em artigos científicos e em modelos disponibilizados em *open source* pelas incumbentes do setor, possibilitando o surgimento de *startups*, organizações acadêmicas e de pesquisa aberta, que ainda esbarravam nas barreiras à entrada típicas da era

dos grandes modelos. Empresas nacentes conseguiram contornar (temporariamente) as duas principais barreiras à entrada no setor. Primeiro pelo acesso a dados na Internet e pela criação de bases de dados públicas. Por outro, pelo acesso à ampla infraestrutura de computação disponibilizada como serviços na nuvem pelas empresas do núcleo infraestrutural do setor. No entanto, não resolveram a assimetria entre incumbentes e emergentes de forma definitiva.

A importância de amplos conjuntos de dados para o desenvolvimento e desempenho dos modelos de IA foi amplamente retratada na literatura, bem como a *BigData*, analisada em diversos estudos sobre a IA preditiva, em meio a temas como privacidade, proteção de dados, vieses algorítmicos e poder de mercado que conferem às empresas líderes.

A escala crescente da IA generativa precisa de conjuntos de dados cada vez mais extensos e domínios variados para o treinamento, desempenho/acuracidade dos modelos, desenvolvimento de maior conjunto de tarefas, habilidade para criar conteúdo e, em última instância, para o progresso técnico no campo da IA. A importância dessas massivas bases de dados é clara para aprendizado/treinamento de modelos, cuja essência é a identificação de padrões e capacidade de fazer classificações, recomendações, previsões ou criação de conteúdo (novos dados) a partir de inferências probabilísticas. Negar isso é desconhecer a essência dos sistemas de redes neurais profundas.

O acesso a amplos conjuntos de dados, passíveis de anonimização, é uma das principais fontes de assimetria entre empresas nascentes e grandes empresas do setor, uma das principais vantagens competitivas dos pioneiros que estruturaram amplos ecossistemas de plataformas digitais e o acesso a fluxos contínuos de dados de sua ampla base de clientes, possibilitando a estruturação de extensas bases de dados proprietárias. A *BigData* é um exemplo paradigmático de ativos complementares, intangíveis, proprietários das *BigTechs*, conservados em sigilo.

Bases proprietárias constituem uma típica vantagem do pioneiro e possibilitam avanços nos novos modelos generativos. As empresas nascentes que conseguiram se destacar no desenvolvimento de modelos notórios de IA enfrentaram essa barreira à entrada por meio do emprego de práticas de “raspagem” de dados de linguagem e imagem de toda a Internet, como OpenAI, Stability AI e Midjourney, entre outras.

Muitos dos dados utilizados pelas startups são propriedade de terceiros e estão levando a incontáveis ações judiciais por violação de direitos autorais pelo uso não autorizado, movidas por empresas proprietárias de bancos de dados, principalmente de imagem, por grandes gravadoras, romancistas, roteiristas de cinema e televisão e editores da grande mídia. Como mencionado, duas visões opostas permeiam o debate.

Por um lado, ainda que possam envolver direitos autorais legítimos de criadores –

embora inegavelmente interesses pecuniários oportunistas -, o acesso a grandes conjuntos de dados é essencial para o treinamento dos modelos de IA, que não é uma ferramenta digital de cópia e plágio, mas insumos de sistemas computacionais capazes de revolucionar muitos campos da economia e da sociedade.

Por outro lado, alguns defendem que esses casos sejam avaliados a luz da doutrina do *fair use* ou como exceções do direito autoral, em virtude da sua importância para o progresso técnico para esse valioso propósito social, direitos proprietários não podem se sobrepor ao benefício maior para a sociedade dessa tecnologia potencialmente revolucionária.

Os modelos convertem dados de entrada, na verdade pedaços de palavras ou imagens em representações numéricas vetoriais, codificados e anonimizados, com uso limitado ao treinamento, sem a reprodução do material de origem na inferência que define as saídas do modelo. A cópia ou reprodução, se ocorrer, decorre de falhas técnicas dos modelos.

Algumas empresas emergentes estão buscando alternativas, como a aquisição do Rockset, banco de dados analítico em tempo real, em junho de 2024, pela OpenAI (OpenAI, 2024b), muitas, entretanto, podem sucumbir em meio aos demorados custos financeiros e reputacionais dos processos judiciais, fechando as portas para exitosos modelos de IA.

Todavia, a importância dos dados para treinamento dos modelos de IA está levando a proliferação de bases públicas, na sua maioria de imagens em regime *open-source* (em geral com licenças *Creative Commons*), como LAION e Common Crawl, organizações sem fins lucrativos que rastreiam dados na Web e fornecem livremente ao público (a segunda inclui até obras protegidas por direitos autorais distribuídas nos EUA sob a doutrina do “*fair use*”, na forma de amostras de sites, sendo a mais usada para treinamento de LLMs) (Tabela 7.1).

Além da questão relacionada à propriedade dos dados, há projeções que indicam o esgotamento dos dados linguísticos de baixa qualidade entre 2030 e 2050 (dados linguísticos de alta qualidade se extinguiriam já em 2026) e os dados de visão/imagem, entre 2030 e 2060 (Villalobos et al., 2022). Com isso, os sistemas de IA enfrentarão gargalos potenciais para expandir conjuntos de dados de treinamento, incluindo os LLMs treinados em trilhões de palavras de texto gerado por humanos da Internet.

Villalobos et al. (2024) estimam que o estoque atual de dados de texto público gerados por humanos com qualidade para ser usados em treinamentos, é da ordem de 300 trilhões de tokens. A partir de dois cenários distintos – um deles baseado na mera extrapolação da taxa histórica de crescimento dos conjuntos de dados; o outro calculado como decorrência da projeção do crescimento da computação de treinamento -, os autores prevêm a partir de um modelo econométrico a utilização de todo o estoque disponível de dados públicos entre 2026 e

2032, se mantidas as atuais tendências de treinamento³³⁷, ou mesmo antes, se as escalas de treinamento aumentarem, exceto se treinados computacionalmente de forma otimizada³³⁸.

Tabela 7.1 - Vinte principais bases de dados de acesso aberto da IA generativa

Nome do conjunto de dados	Citação de documentos	Total de menções	Modalidade principal
ImageNet	2741	6823	imagem
MNIST	2533	9292	imagem
CIFAR-10	2160	7744	imagem
CelebA	1705	5713	imagem
COCO	1141	3390	imagem
Wikipedia	662	2599	texto
FFHQ dataset	596	1983	imagem
FASHIONMNIST	520	1375	imagem
CelebAHQ	474	1081	imagem
SVHN	398	1414	imagem
PubMed	393	1144	texto
GSM8K	350	1704	texto
CIFAR-100	338	849	imagem
HumanEval	322	1269	texto
LAION	314	731	imagem
CUB dataset	312	1118	imagem
LSUN	310	608	imagem
CommonCrawl	290	546	texto
Cityscapes	272	974	imagem
MMLU	270	746	texto

Fonte: WIPO (2024)

Diante dos desafios em termos de quantidade e qualidade de dados criados por humanos, o progresso dos modelos de IA dependerá da geração de dados sintéticos (Almeida, 2024), embora esses dados ainda pareçam confiáveis somente em domínios estreitos, como matemática e codificação, não em texto, principal modalidade usada no treinamento de modelos de fronteira³³⁹. O progresso dos modelos dependerá também da ampliação da utilização de aprendizado por transferência de domínios ricos em dados e das melhorias na eficiência de dados (Villalobos et al., 2024). Uma outra alternativa é o crescimento dos parâmetros dos modelos, permitindo manter o tamanho do conjunto de dados constante, estratégia chamada de "subtreinamento", que também alcançará um limite computacionalmente ótimo.

Até mesmo a maioria das *BigTechs*, que coletaram grandes conjuntos de dados, um ativo complementar essencial ao seu negócio, começam a apostar fortemente em dados sintéticos, gerados artificialmente, criados e curados por IA, a partir de técnicas como *label-efficient*

³³⁷ Modelos recentes, como o Llama 3, são sobre-treinados (menos parâmetros e mais dados) para que sejam mais eficientes computacionalmente durante a inferência. Com isso, são modelos treinados em mais dados do que o prescrito pelas leis de dimensionamento ideal de computação, ainda que com desempenho um pouco menor.

³³⁸ O termo "computacionalmente ótimo" refere-se à seleção do tamanho do modelo e do conjunto de dados de uma forma que resulte no melhor desempenho possível, com nível fixo de computação - para modelos densos isso é alcançado no treinamento de 20 tokens para cada parâmetro do modelo (Hoffmann et al., 2022).

³³⁹ Dados privados, não públicos, como conversas em mensagens instantâneas, são improváveis de serem usados em escala devido a questões legais e porque estão fragmentados em várias plataformas controladas por *players* com interesses concorrentes.

learning (Deng, 2023). Desde o AlphaZero, passando pelo LLaMA 3.1 (Meta), dados sintéticos têm sido usados no treinamento de modelos (Sevilla et al., 2024), embora um tópico de pesquisa ainda incipiente. As equipes da Microsoft também trabalham em dados sintéticos, treinando modelos menores com quantidades significativas desses dados em modelos maiores (Afonja et al., 2024) entre outras, campo promissor para a próxima geração de modelos (Ramos e Subramanyam, 2021).

Todavia, diversos autores têm apontado problemas no uso de dados sintéticos, como Shumailov et al. (2023) que constatam que o uso desses dados no treinamento causa defeitos irreversíveis, colapsando os modelos – processo degenerativo pelo qual, ao longo do tempo, os modelos “esquecem” a verdadeira distribuição de probabilidades dos dados subjacentes, aumentando erros em dados gerados, com danos e riscos imprevisíveis. Martinez et al. (2023) argumentam que mesmo o uso desses dados combinados com dados reais criados por humanos no treinamento de modelos causará risco de sua degradação.

Finalmente, a chegada da IA generativa transformou os ativos tangíveis em principal ativo/tecnologia complementar, repondendo cada vez mais pelo avanço dos poderosos modelos e abrindo uma nova etapa de reconhecimento irrefutável da importância da infraestrutura de hardware, dos chips, data centers, redes e computação na nuvem.

De acordo com a abordagem do PFI, muitas invenções importantes, passíveis de proteção por PI, não asseguraram vantagens econômicas duradouras, porque firmas pioneiras deixaram de exercer o controle sobre ativos e tecnologias complementares essenciais para o sucesso da inovação. Ainda que a simples contratação de terceiros tenha sido suficiente em alguns casos, em muitos outros fontes reais ou potenciais de gargalos e competição de fornecedores tornaram necessária a integração vertical do inovador pioneiro. Os ativos “tangíveis” da IA, parecem representar perfeitamente o conceito desses ativos/tecnologias complementares, essenciais para o sucesso a longo prazo da inovação em IA.

As notórias startups americanas – às quais Patel e Nishball (2023) agregam Databricks (MosaicLM), Together e HuggingFace - alcançaram protagonismo na IA generativa por disporem de talentos de classe mundial, mas principalmente pelas parcerias, investimentos e créditos de computação das *BigTechs* – o que é também provável nas *startups* chinesas. Comprar cargas de computação se tornou um requisito essencial para competir.

A gigantesca capacidade computacional cada vez mais requerida pelos modelos da IA generativa é controlada pelo núcleo infraestrutural das *BigTechs*, americanas e chinesas, com destaque para Amazon, Microsoft e Google. Estudo da Forrester sobre as principais provedoras de soluções de infraestrutura de IA classificou essas empresas em uma tipologia de 4 grupos:

líderes (Google, Microsoft, Amazon, Nvidia e Dell); desempenho forte (HP e IBM), concorrentes (Alibaba, Lenovo, Oracle) e desafiadoras (Cerebras Systems, Graphcore) (Gualtieri et al., 2024).

Esses ativos/tecnologias complementares, originalmente provido por fornecedores especializados, passaram a ser produto da colaboração e relações entrelaçadas com as líderes da IA, respondendo em parte pelo avanço espetacular dos novos modelos. Com isso, empresas como Nvidia começaram a desenvolver modelos próprios de IA, bem como apostando em empresas nascentes de IA, além de investimentos em infraestrutura de computação na nuvem. A construção dessa infraestrutura supera expectativas, com gastos incrementais de US\$ 60-80 bilhões acima do investimento regular em nuvem, ferramenta cada vez mais crítica para construir e treinar modelos de IA (Goldman Sachs, 2024).

A assimetria entre emergentes e incumbentes da IA é absoluta quando se trata dos ativos de hardware, cujo investimento requerido é totalmente incompatível com seu porte – o investimento da Microsoft (US\$ 100 bilhões) no desenvolvimento do supercomputador *Stargate* vale mais do que o valor de mercado da OpenAI, a maior das emergentes da IA generativa – da mesma forma que os custos de treinamento dos modelos e gastos com energia, não lhes restando outra alternativa além dos acordos contratuais e garantias de créditos de computação do núcleo infraestrutural pelas *BigTechs*.

Algumas *startups* europeias começam a despontar no cenário da IA generativa apoiadas por investimentos dos governos e União Europeia, grandes grupos locais e das próprias *BigTechs*, para infraestrutura computacional, com destaque para os supercomputadores que estão sendo instalados em países europeus pelo consórcio EuroHCP JU,³⁴⁰ iniciativa conjunta UE, países do bloco e setor privado, voltado para construção de um ecossistema regional de supercomputação de classe mundial, com orçamento €7 bilhões para 2021/27. No entanto, ainda está muito distante dos investimentos em infraestrutura das *BigTechs*, evidenciando o enorme *gap* da maioria dos países, desenvolvidos e em desenvolvimento, com exceção da China e alguns países do Oriente Médio que estão investindo pesadamente em infraestrutura em larga escala para IA.

Na estrutura do PFI, ativos genéricos são aqueles de uso geral, que não demandam qualquer tipo de adaptação à inovação focal, enquanto ativos especializados são aqueles em que existe dependência unilateral entre a inovação e o ativo complementar; além dos co-especializados, em que há dependência bilateral entre a tecnologia focal e o ativo. A

³⁴⁰ Disponível em: https://eurohpc-ju.europa.eu/index_en. Acesso em: 10 ago. 2024

computação inicialmente usada nos modelos de redes neurais profundas foram as GPUs, relativamente padronizadas adaptadas para aprendizado profundo, um claro ativo especializado, evoluindo na IA generativa para ativos efetivamente co-especializados, dependentes de complementaridades bidirecionais ou, mais provavelmente, ecossistêmicas, como no caso das TPUs do Google e dos chips para IA – o que não significa que não possam ser ativos genéricos no caso de devs. dedicados a aplicativos, que usam modelos pré-treinados e operam em GPUs padronizadas, sem necessidade de estruturação de clusters específicos.

O uso das tecnologias de IA ainda está limitado à indústria criativa ou incorporado aos produtos e serviços existentes e consolidados das *BigTechs*, como o Copilot incorporado aos produtos do Microsoft 365 para assinantes, o ChatGPT ao buscador Bing/Microsoft, o Gemini ao Google Search, ao navegador Chrome, entre outros produtos da empresa, podendo repetir a estratégia exitosa do Search Engine se firmar acordo com a Apple para acesso no iPhone. Os exemplos são muitos considerando outras empresas, como o Samsung Gauss, modelo de IA generativa anunciado que a empresa planeja integrar a celulares e eletrodomésticos inteligentes, o Hunyuan multimodal que a Tencent integrou ao WeChat, o Doubao AI da ByteDance incorporado aos seus produtos de plataforma, além dos chatbots de IA generativa acoplado aos produtos de plataforma da Baidu e Alibaba, voltados para setores como TI, Transporte ou Energia (WIPO, 2024). O LLaMA 3 70B já está disponível nos três produtos da Meta – Facebook, WhatsApp e Instagram – em 14 países, totalizando 3,24 bilhões de usuários ativos diários, que atende com um modelo pequeno e eficiente, como apontam Patel e Nishball (2024a), autores que compartilham da visão de que a maior utilização dos modelos generativos passa pelas plataformas digitais existentes, assegurando as empresas que as comandam renovado êxito das inovações por meio do consagrado modelo de negócios.

A computação na nuvem parece se sagrar como o ativo complementar essencial para a apropriabilidade da IA generativa, em que as *BigTechs* com robusta infraestrutura de computação na nuvem migram para uma nova geração de plataformas digitais como prestadores de serviços na nuvem (*Cloud Services Providers*), fornecendo uma gama crescente de serviços - infraestrutura, software, modelos e até dados -, com perspectivas de se tornar um modelo de negócios que abranja toda a cadeia de valor digital.

O fornecimento do “*Everything as a Service*” (*XaaS*) parece representar a transição para um novo modelo de negócios, uma segunda geração de plataformas digitais voltada para empresas e negócios, não somente pessoas físicas, baseados em instrumentos contratuais (Graux, 2021). Daí emerge um novo formato de apropriabilidade, regido por contratos, complementar ao amplo arsenal de direitos de PI, além do regime *open-source*.

O fornecimento desse amplo conjunto de serviços promete ser o embrião do novo modelo de negócios para a implementação da IA em setores *downstream*, respondendo às incertezas sobre o formato de monetização da IA generativa (Jacobides, 2023), ainda que ao custo da subordinação dos clientes setoriais e sua capacidade de inovação às estratégias de expansão setorial das próprias *BigTechs* (Rikap 2023d, Rikap, 2022), dependendo da resiliência de líderes *downstream* em lidar com ondas de tecnologias transversais como a IA.

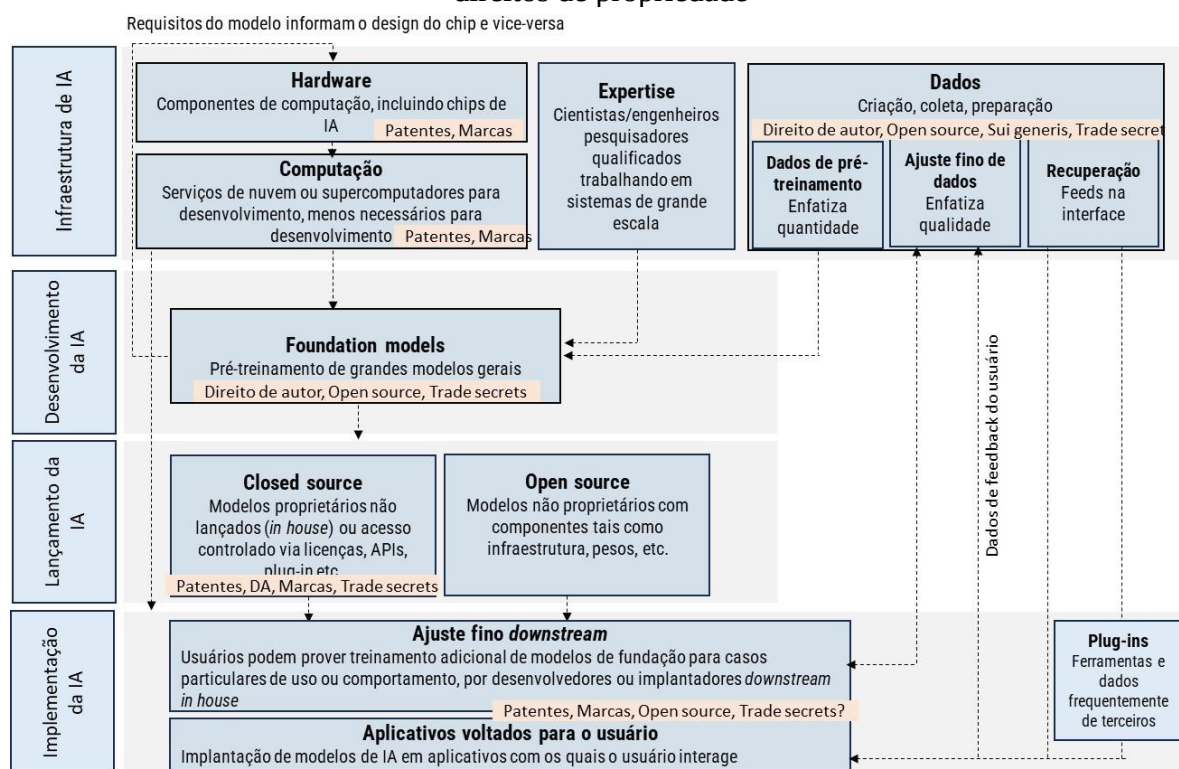
O caso das inovadoras emergentes é, entretanto, mais desalentador. Enfrentam sérias dificuldades para desenvolver aplicações reais no mercado e projetar um modelo de negócios que viabilize a inovação da IA generativa, uma vez que a disponibilização gratuita ou quase gratuita dos modelos criam incertezas sobre seu formato de monetização (Jacobides, 2023), com implicações inexoráveis sobre as condições de apropriabilidade. As exceções são pontuais e pouco exitosas, como a mencionada GPT Store e o protótipo do Search GPT. Na realidade, nada novo, apenas a tentativa de replicar produtos/serviços consagrados.

Além dos impactos mais amplos do regime de apropriabilidade, a IA generativa parece ampliar o escopo do regime legal de PI, com variados desses direitos permeando várias dimensões da estendida cadeia de valor que origina produtos da IA. Estudo recente da autoridade reguladora britânica detalhou a cadeia de valor para os modelos de fundação (CMA, 2024), que comporta os vários recursos exigidos para a criação e implementação dos modelos da IA, que nesta tese foram tratados por meio da noção de ecossistema da IA, conceito que melhor reflete a realidade da inovação das tecnologias emergentes do século XXI, numa espécie de ecologia de diferentes ações, interrelações e estratégias em permanente evolução. O estudo apresenta o diagrama reproduzido na Figura 7.1, que detalha a ampla cadeia de valor de um tipo de modelos da IA generativa, os complexos modelos de fundação, em estágios diferenciados, tais como a infraestrutura da IA (que contempla a hardware, computação e dados), o desenvolvimento da IA (os modelos propriamente ditos), o lançamento da IA (nas modalidades aberta ou fechada) e a implementação da IA, na forma de sistemas com ajuste fino para aplicações específicas ou aplicativos para usuários finais.

A figura foi adaptada com vistas a elencar os muitos instrumentos de PI relacionados aos insumos e invenções da IA, que envolvem patentes, direito de autor, direitos *sui generis*, *trade secrets*, além de marcas, e o próprio regime *open-source*, em vários elos da cadeia de valor, até mesmo na fase de aplicativos ao usuário, ainda que preferencialmente por organizações sem fins lucrativos. Vale lembrar que o *open source* na versão “open for business” pode estar sendo a escolha estratégica para muitas empresas, uma vez que existem outros mecanismos de proteção, como a omissão de informações técnicas sobre os modelos, que

progressivamente deixaram de ser reveladas, principalmente depois do impacto da disseminação desenfreada de novos modelos desde o vazamento do LLaMa. Além disso, as líderes do setor dispõem de amplos conjuntos de dados digitais proprietários e uso generalizado de medidas de restrição técnica sobre essas bases de dados, além da extraordinária infraestrutura computacional que lhes assegura a efetiva apropriabilidade da inovação – seja a partir de modelos próprios, seja por modelos de terceiros, desde que as líderes do núcleo infraestrutural empurraram o progresso técnico para modelos com escalas estratosféricas, possibilitado pelo controle que detêm sobre hardware e infraestrutura digital, fora do alcance das empresas emergentes.

Figura 7.1 - Representação visual da cadeia de valor dos modelos de fundação e possíveis direitos de propriedade



Fonte: Adaptado de CMA (2024)

7.3. Os verdadeiros vencedores da corrida da IA

A IA generativa, mesmo com a emergência de *startups* e unicórnios ao redor do mundo, parece reforçar a tendência da IA preditiva, desde meados da década passada, com a concentração do progresso técnico nas mãos de umas poucas empresas, sobretudo americanas e chinesas, suplantando de longe capacidades até mesmo da academia e do governo, com perspectivas de promover amplas mudanças estruturais nos setores onde vier a ser aplicada.

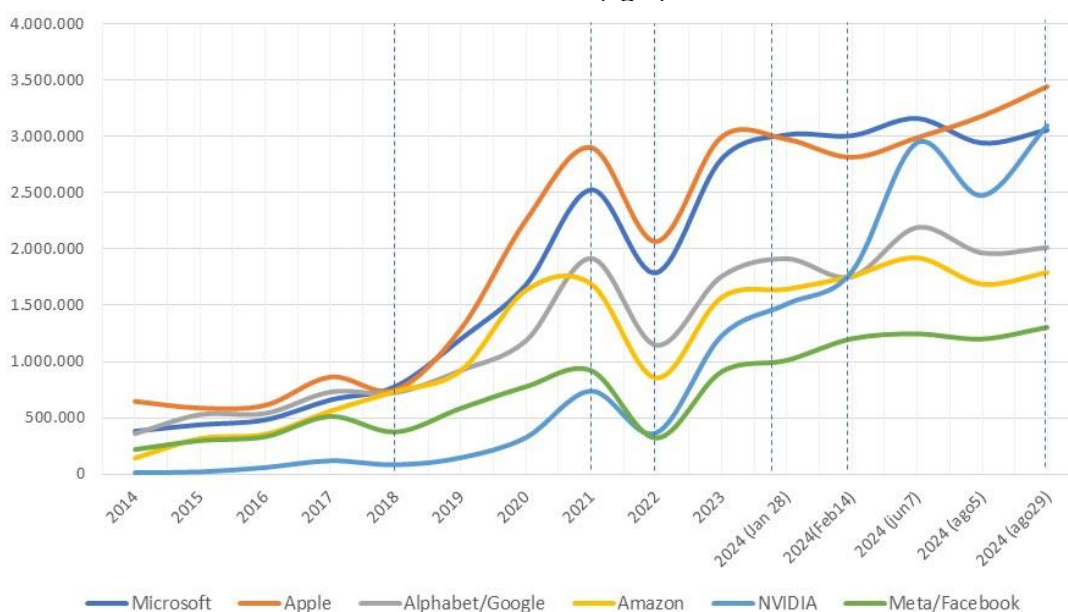
Mesmo que qualquer afirmação ainda seja prematura, os verdadeiros vencedores da IA generativa, principalmente com o avanço dos modelos intensivos em escala, parecem ser as mesmas empresas líderes da inovação da IA, em particular o núcleo infraestrutural das *BigTechs*,

americanas e chinesas, como Google, Microsoft, Amazon, Alibaba, Baidu e Tencent, aos poucos incorporando a Meta AI e a Apple, cujo tradicional pragmatismo silencioso sempre poderá surpreender.

A elas podem se juntar empresas de segmentos contíguos ao ecossistema da IA, como software e soluções empresariais (IBM, Salesforce, SAP e Oracle) ou empresas especializadas em bancos de dados (Databricks e Adobe) ou chips de hardware como Nvidia e rivais que parecem voltar ao jogo desde as medidas de incentivo e proteção do governo americano. Muitas não dispõem de modelos de negócios de plataformas digitais convencionais, mas começam a replicar a nova geração de plataformas na nuvem das *BigTechs*, o que abrirá novos horizontes para uma efetiva competição. Dificilmente, contudo, os unicórnios da IA poderão suportar a corrida dos modelos de grande escala, exceto se alcançarem uma aplicação matadora da IA, uma possível AGI ou alguma inovação organizacional que possibilite suplantá-las suas limitações da falta de integração vertical e de condições de apropriabilidade da inovação em IA.

A liderança na expansão da IA generativa fica evidente na avaliação do mercado sobre as principais empresas, as cinco *BigTechs* americanas e a inclusão da Nvidia, a partir da sua capitalização (Gráfico 7.4). Ainda que a capitalização de mercado nada mais seja do que a simples multiplicação do preço de mercado da empresa pelo número total de ações ordinárias ou ordinárias em circulação, que corresponde ao montante de recursos necessário para comprar todas as ações de uma empresa pelo atual preço de mercado, essa avaliação reflete a visão sobre seus recursos, realizações e perspectivas futuras.

Gráfico 7.4 – Evolução das maiores empresas em capitalização de mercado, US\$ bilhão
2014 a 2024 (ago.)

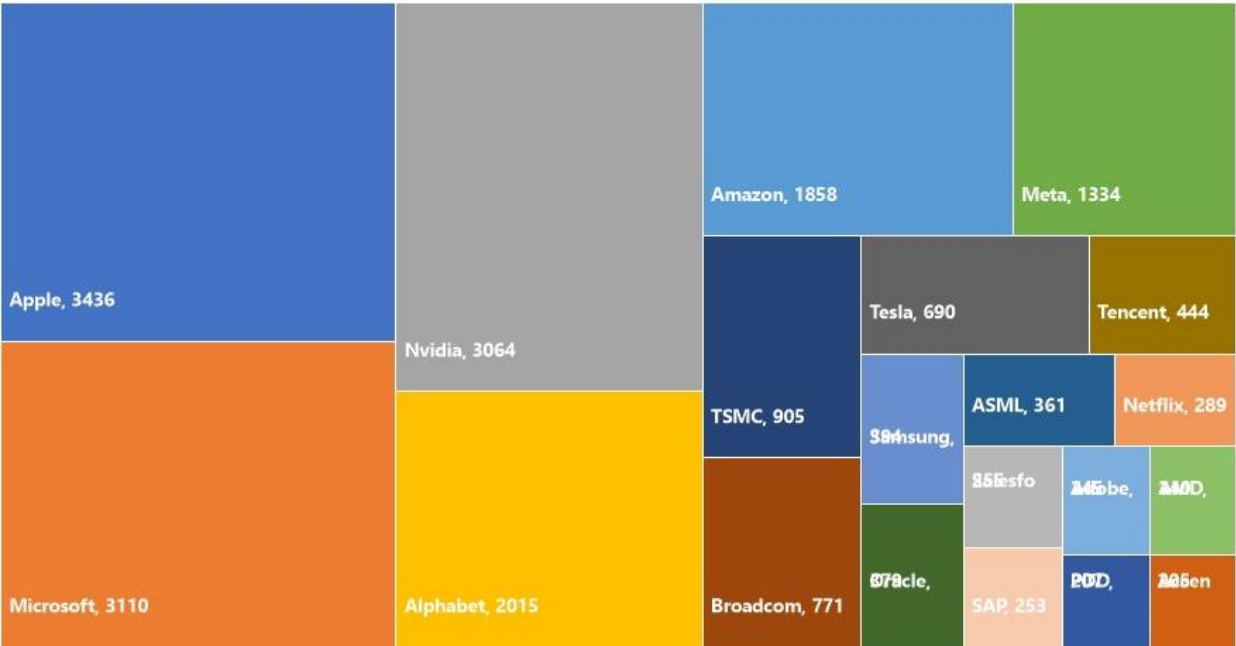


Fonte: Elaborado a partir de <https://companiesmarketcap.com/> (último acesso 29 ago. 2024)

A evolução desse indicador, na falta de outro melhor e a despeito das flutuações recentes das ações das empresas em meio a conjecturas sobre um possível estouro da bolha da IA generativa, evidencia quatro pontos importantes: a) um avanço acelerado das quatro principais *BigTechs* desde 2018 (que hoje superam a gigante do petróleo Saudi Aramco); b) a expansão progressiva da Nvidia, principalmente desde 2022 – a despeito da queda no preço das ações de todas as empresas de tecnologia nesse ano -, superando *BigTechs* notórias como a Meta (em 2022), a Amazon e a Alphabet/Google (em fev. 2024) e disputando com a Microsoft (em ago. 2024) a segunda posição no *ranking* das maiores empresas; c) a retomada do crescimento da Microsoft, revertendo antigas avaliações pessimistas, desde que apostou na OpenAI e incorporou os modelos aos seus produtos, chegando até a suplantar transitoriamente a liderança história da Apple; d) o avanço da Meta, a menor das *BigTechs*, desde 2023, que voltou ao clube das maiores empresas.

Essas seis empresas ultrapassaram o *benchmark* de US\$ 1 trilhão em capitalização de mercado e no seu conjunto têm uma avaliação agregada de mais de US\$ 15 trilhões, quase três quartos do somatório das 20 maiores empresas de tecnologia do mundo (Gráfico 7.5) e quatro vezes o valor total do conjunto dos unicórnios – empresas com valor de mercado superior a US\$ 1 bilhão. A Tabela 7.2, que discrimina os vinte unicórnios mais valiosos com atuação em IA, bases de dados (Databricks) e em chips (Celebras, SambaNova, Graphcore), resume as várias assimetrias que distinguem os dois conjuntos de inovadores da IA.

Gráfico 7.5 - Capitalização de mercado, vinte maiores empresas de tecnologia, US\$ bilhão



Fonte: Elaborado a partir de <https://companiesmarketcap.com/> (último acesso 17 ago. 2024)

Tabela 7.2 - Valor de mercado dos principais unicórnios de IA, chips e bases de dados em US\$ bilhão (nov. 2023)

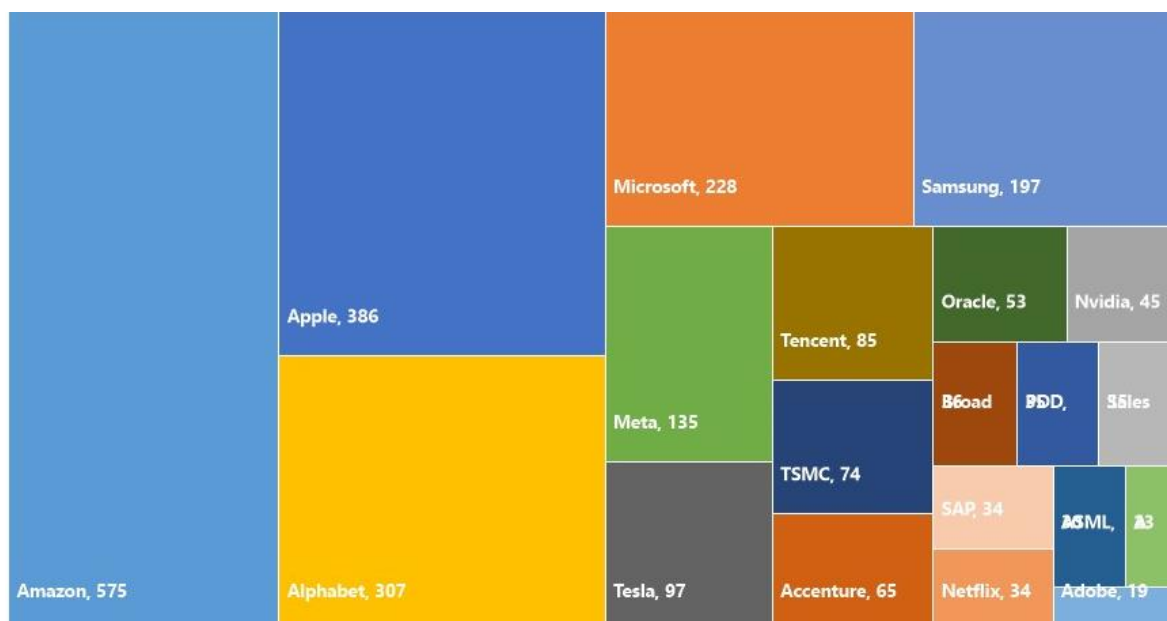
Empresa	País	Valor
OpenAI	US	80
xAI	US	24
Anthropic	US	16
Huggingface	US	4,5
Inflection AI	US	4
Cohere	CA	2
Stability AI	UK	1
Zhipu AI	CN	3
Baichuan AI	CN	1
01.AI	CN	1
AI.21	IL	1,4
Mistral AI	FR	6,2
Databricks	US	43
Synthesia	UK	1
Jasper	US	1,5
Cerebras	US	4
Discord	US	20
Grammarly	US	13
SambaNova	US	5
Graphcore	UK	2,7

Fonte: Elaborado a partir de [CBInsights \(2024\)](#) (último acesso 17 ago. 2024)

O maior dos unicórnios da IA, OpenAI, representa apenas 6% do valor da menor das *BigTechs* (Meta), com receita de cerca de US\$ 2 bilhões, em 2023 (e praticamente zero no ano anterior), embora na atual nova rodada de captação no mercado a empresa informe receitas de US\$ 300 milhões em agosto de 2024, anualizada em US\$ 3,7 bilhões, e uma reavaliação que quase dobra seu valor de mercado. O aumento previsto de receita parece ter origem no aumento dos valores da assinatura paga do ChatGPT, para US\$ 22/mês, e da versão premium, para US\$ 44/mês (Zitron, 2024b). No entanto, ainda terá resultado final de prejuízo de US\$ 5 bilhões, em parte pelo acordo leonino com a Microsoft, comentado no Capítulo 4. Um negócio estruturado de modo que parece incapaz de gerar lucros, repetindo a história da DeepMind, que terminou incorporada ao Google. Até mesmo a Nvidia, cujo avanço em capitalização de mercado não se reflete em receita, muito inferior a das *BigTechs* – apenas um terço da receita da Meta (Gráfico 7.6).

Os gastos com P&D das seis maiores empresas são expressivos, com destaque para a Amazon, em valor absoluto, e para a Meta, como percentual da receita (Tabela 7.3). O gasto com P&D da Nvidia é similar a todo o gasto com C&T do governo federal do Brasil, no mesmo ano. Há grandes discrepâncias em termos dos salários médios pagos – cabendo destacar que os salários médios mais baixos da Amazon carregam o alto número de empregados de varejo e logística, sem relação com atividades de tecnologia – e discrepâncias ainda mais evidentes em termos da remuneração dos CEOs.

Gráfico 7.6 - Receita anual, vinte maiores empresas de tecnologia em capitalização de mercado, US\$ bilhão, 2023



Fonte: Elaborado a partir de <https://companiesmarketcap.com/> (último acesso 17 ago. 2024)

Tabela 7.3 - Gasto em P&D (US\$ bilhões) e percentual da receita (2022); Salário médio (US\$, 2023); Salário CEO (US\$ milhão, 2023), empresas selecionadas

Empresa	Gasto P&D	P&D % da Receita	Salário médio	Salário CEO
Amazon	73,2	14%	36.274	1,4
Alphabet	39,5	14%	315.531	8,8
Meta	35,3	30%	379.050	24,4
Apple	27,7	7%	94.118	63,2
Microsoft	26,6	13%	193.770	48,5
Nvidia	7,3	27%	266.939	34,2

Fonte: VisualCapitalist (Disponível em: <https://www.visualcapitalist.com/cp/ranked-the-10-biggest-nasdaq-companies-by-rd-investment/>. Acesso 15 set. 2024)

As *BigTechs* parecem um conjunto dissonante em termos de estratégia, até mesmo seu núcleo infraestrutural, que empreendeu distintas estratégias financeiras, entre 2000 e 2019, com políticas diferentes de distribuição de dividendos - aparentemente relacionada a posição relativa de cada empresa no mercado – Amazon, Meta (Facebook) e Alphabet não distribuindo dividendos, optando pelo reinvestimento de lucros, enquanto se firmavam, enquanto as mais maduras/consolidadas, Microsoft, Apple, Alibaba e Tencent, distribuindo valores expressivos aos acionistas. A política de investimento em ativos fixos também foi diferenciada, com destaque para Alphabet e Amazon, cujos ativos fixos apresentavam valores muito superiores a Microsoft, Apple e Facebook - Amazon por apoiar sua atividade no varejo em operação de logística e armazenamento e expansão em infraestrutura em nuvem, enquanto a Alphabet pelos investimentos em data centers (Fernandez et al., 2020). Mas que parece ter mudado, pelo menos no caso da Microsoft.

Os números de capitalização de mercado parecem refletir o poder de mercado dos principais *players* e seus elevados investimentos em P&D e infraestrutura de computação em nuvem - mais de 30% das suas receitas na nuvem são revertidas em investimentos em IA, com destaque para o núcleo infraestrutural das *BigTechs*, fornecedoras de serviços de computação e armazenamento em escala, cada vez mais críticos para desenvolver/treinar modelos.

Entre os novos competidores em IA a empresa verdadeiramente promissora é a Nvidia, evidenciando a inevitável integração dos modelos à capacidade computacional na fase da IA generativa. Seu serviço na nuvem DGX Cloud mostra capacidades que não se restringem à supercomputação e pré-treinamento de modelos, disponibilizando estruturas de processamento de dados, bancos de dados vetoriais e personalização, mecanismos de inferência otimizados, APIs e suporte para ajuste de modelos de terceiros a usos personalizados. Seu amplo e diversificado leque de clientes, não limitado ao setor de TICs e empresas de IA, contempla empresas dos setores farmacêutico, biotecnologia, automotivo, seguros etc., como AstraZeneca, Amgen, Accenture, Adobe, Getty Images, ServiceNow, Insilico Medicine e muitas parcerias não reveladas (Patel e Nishball, 2023), incluindo empresas chinesas. De fato, todas as gigantes da IA, passando pelas *startups* e empresas de setores *downstream*, dependem da Nvidia e respondem por sua expansão recente. A exceção notória parece ser o Google, que investiu pesadamente na construção de infraestrutura específica para IA há quase vinte anos e, principalmente após 2013 – após Amazon –, com ampliação de data centers e a base para o início da produção de TPUs em 2016 (Sato e Young, 2017; Patel, Cozma e Wong, 2023).

O Google parece ser o efetivo pioneiro da inovação da IA, responsável pela revolucionária arquitetura Transformers, base da maioria dos atuais modelos de fronteira, pela construção de pontes com a comunidade de código aberto, abrindo o código-fonte de sua biblioteca de aprendizado de máquina TensorFlow, precursor da integração vertical dos ativos complementares da IA, como grandes conjuntos de dados, TPUs projetadas em chips personalizados para a IA, investindo em uma infraestrutura de classe mundial. Três empresas parecem liderar o avanço das invenções IA generativa, OpenAI, Google DeepMind e Meta AI, mas o sucesso a longo prazo dependerá do regime de apropriabilidade da inovação. Dos avanços inegáveis da OpenAI, rejuvenescendo a estratégia empresarial da Microsoft, aos modelos “open for business” da Meta e responsável por muitas invenções algorítmicas, seguidas de perto pela Amazon e, mais recentemente, Apple, não deixam dúvidas sobre os prováveis vencedores.

Embora as *startups* tenham explorado muitas possibilidades da IA generativa, o avanço de modelos gigantes e a importância da infraestrutura computacional parecem criar uma barreira intransponível a longo prazo, gerenciada no curto prazo por meio da computação na

nuvem provida pelas gigantes do setor, com quem passaram a ter vínculos viscerais, e sem conseguir inovar organizacionalmente no sentido de um modelo que assegure a apropriabilidade das suas inovações, têm poucas chances de efetivamente se juntar ao clube das gigantes. Isso se reflete em uma estrutura de custos que se desvia das *startups* do software convencional, passando a recair sobre chips e arquiteturas fundamentais para a escalabilidade dos modelos, com peso considerável sobre suas receitas e levando ao ceticismo sobre o futuro da IA ou, pelo menos de muitas empresas emergentes.

De qualquer forma, é inegável que a IA deixou de ser um fenômeno exclusivamente americano, e chinês, disseminando pelo mundo principalmente em função de políticas industriais ativas, políticas comerciais, dispositivos em acordos comerciais, paralelamente à ampla regulação em diversos países – e a nível global – cujo impacto provável recairá sobre os pioneiros, justamente as empresas americanas.

O (s) verdadeiro (s) vencedor (es) dessa disputa acirrada serão aquelas que conseguirem definir o *design* tecnológico dominante e a monetização das invenções da IA generativa, que até agora parece restrito às *BigTechs*, ainda que não exclua as gigantes chinesas de plataformas, que também dispõem de robustas infraestrutura na nuvem, desde que contornem as restrições americanas de hardware.

O avanço descomunal do *open source* na IA generativa, atrelada ao objetivo de definição do *design* dominante e as incertezas sobre as regras americanas para patenteamento *vis-à-vis* as regras chinesas, talvez respondam pela desaceleração da corrida por patentes pelas empresas americanas. O recente maior fechamento dos detalhes técnicos dos modelos, face ao acentuado surgimento de novos *players*, não significa uma desaceleração do *open for business*, sendo mais provável um menor patenteamento, de forma a evitar que parte dessas informações sejam divulgadas.

De todo modo, o mais lento crescimento das patentes americanas frente às chinesas precisa ser mais bem explorado, considerando aspectos institucionais e das políticas governamentais como, de um lado, os incentivos chineses ao patenteamento da China e, de outros, as incertezas sobre a patenteabilidade das invenções e dificuldade de reivindicações de patentes amplas nos Estados Unidos. Não obstante, o turbulento ambiente regulatório e escrutínio antitruste sobre a IA, que agora recai fortemente sobre o negócio de computação na nuvem, diante dos riscos da IA, do poder de mercado das empresas líderes americanas, poderá ter implicações ainda imprevisíveis sobre a arquitetura da indústria e, possivelmente desproporcionais sobre as empresas líderes, com efeitos sobre o regime de apropriabilidade da inovação e do próprio progresso técnico da IA.

CONCLUSÕES

A longa trajetória de evolução da IA, desde um ecossistema de orquestração pública (defesa) até um ecossistema de orquestração privada, em meio aos avanços da Internet, da digitalização de dados e do crescimento da capacidade computacional, possibilitou o desenvolvimento de sistemas computacionais baseados em complexos modelos matemáticos e estatísticos, da abordagem conexionista, progressivamente mais eficientes e com elevados níveis de precisão. São modelos desenvolvidos a partir de amplas redes globais de pesquisa colaborativa e conexões interempresariais, ainda que predominantemente levados ao mercado por um grupo seleto de firmas, as *BigTechs*, que estruturaram complexos ecossistemas de inovação e exploraram as potencialidades da inovação organizacional mais importante da era digital.

Desde a introdução do conceito, passando pela demonstração das capacidades e do desempenho dos modelos de redes neurais profundas em competições de jogos complexos, suplantando jogadores humanos, os modelos de IA avançaram gradativamente, impulsionados por pontos de inflexão, resultando em inovações de produtos e serviços inéditos atrelados às plataformas digitais.

As inovações da popularmente chamada IA preditiva resultaram de inovações algorítmicas e modelos construídos a partir dos massivos conjuntos de dados e de desenvolvimentos em capacidade computacional, abarcando conhecimentos internos e externos às empresas, vinculados ou não a estrutura organizacional das plataformas digitais, a partir da construção do mais extenso ecossistema de inovação aberta, que parece caracterizar o progresso técnico na organização empresarial moderna, mesmo que envolvendo muitas assimetrias e tensões em termos de apropriabilidade.

Corresponde a um fenômeno complexo em que o progresso técnico e o avanço do conhecimento evoluem trançados sob o comando de empresas privadas, as maiores do planeta, que não parece ter precedente em nenhum outro setor. Modelos de IA, que crescem em escala e desenvolvem funcionalidades cada vez mais variadas, até mesmo integradas em modelos multimodais que criam conteúdo diverso, embora distantes das capacidades do cérebro humano, que muitos duvidam se serão alcançadas.

As firmas líderes da inovação da IA ergueram um intrincado regime de apropriabilidade que reflete a própria complexidade da inovação em IA. Primeiro, contempla um regime legal de PI extenso e híbrido, cuja combinação de variados direitos, que cobrem as etapas da cadeia de valor e do ecossistema de inovação, decorrente de estratégias empresariais diferenciadas, mas convergentes, convivendo harmonicamente com o regime *open-source*. Segundo, parece

um regime ainda precário, marcado por subjetividade e apoiado em regulamentos administrativos, sem força de lei, com potencial de criar insegurança jurídica e riscos de direitos invalidados quando os muitos pedidos de patentes se converterem em direitos efetivos. Terceiro, as firmas mais dinâmicas em termos da inovação em IA são as que complementam direitos formais de propriedade intelectual, *open-source* e sigilo, evidenciando a precariedade de DPIs como indicador do conhecimento e da inovação em IA, e por meio do controle sobre ativos/tecnologias complementares cada vez mais cruciais para a criação e captura de valor da inovação.

Esse regime de apropriabilidade, que possibilitou a consolidação das principais empresas inovadoras da IA preditiva, passa por mudanças profundas desde o surgimento de modelos da IA generativa, baseados em novas arquiteturas de redes neurais artificiais e técnicas de aprendizado, promovendo uma disseminação incontrolável e uma expansão ilimitada na escala dos modelos.

A literatura sobre apropriabilidade mostrou que o sucesso da inovação a longo prazo depende de estratégias empresariais que definam o *design* tecnológico dominante e de seu posicionamento em ativos complementares. Os sistemas de aprendizado profundo em redes neurais parecem ter potencial para constituir o *design* dominante da IA e ensejar padrões técnicos subjacentes. No entanto, essas tecnologias têm sido consideradas caixas pretas, baseadas em técnicas de aprendizado e algoritmos herméticos. Às iniciativas de organizações governamentais e de pesquisa na busca de padrões de explicabilidade dos sistemas de aprendizado e dos conjuntos de dados, fontes das inovações e riscos associados aos sistemas de IA, se somam programas internos das próprias firmas destinados à explicabilidade, transparência e confiabilidade da IA que, por meio da sua participação nas organizações definidoras de padrões, deverão influenciar a definição de padrões globais, assegurando-lhes vantagens competitivas a longo prazo.

A proliferação de modelos generativos, que pareceu democratizar a IA, disseminados em *open-source* e levando ao surgimento de novos *players*, na realidade consolidou de forma mais profunda o poder (infraestrutural e econômico) das poucas empresas que mantêm o controle sobre os componentes-chaves para a construção de modelos de fronteira.

As empresas emergentes encontraram formas de enfrentar as duas principais barreiras à entrada na corrida tecnológica da IA. Por um lado, raspando tudo o que podiam encontrar na Internet, ao custo de incontáveis ações judiciais por violação de direitos autorais, cujo desfecho poderá inviabilizar a capacidade de muitas delas desenvolverem modelos de IA, quiçá sobreviver. Por outro, o acesso à capacidade computacional por meio de arranjos contratuais e

créditos de computação com as gigantes do setor em serviços prestados na nuvem, que implica em intrincado acoplamento em função de efeitos de rede, interoperabilidade e elevados custos de mudança.

Os avanços acelerados dos últimos dois anos incorporaram novos *players*, unicórnios promissores pioneiros de inovações importantes de modelos generativos, ainda incapazes de capturar o valor da inovação, cada vez mais enredados em inextricável dependência financeira e computacional ao núcleo infraestrutural das *BigTechs*, algumas ficando pelo caminho, “canibalizadas” pelas gigantes do setor. As estimativas da empresa de capital de risco Sequoia são de gastos em hardware Nvidia pelas empresas da ordem de US\$ 50 bilhões, em 2023, frente a receitas auferidas de apenas US\$ 3 bilhões (Romero, 2024), indicando que as startups de IA estão avaliadas irrealisticamente, como a promissora Inflection AI, que apostou alto e terminou desmantelada e parcialmente incorporada a Microsoft, ou da Character AI, parcialmente “absorvida” pela Alphabet/Google, e Adept, pela Amazon, entre outras *startups* menos notórias. Muitos questionam até mesmo as reais perspectivas de crescimento ou mesmo sobrevivência da OpenAI, que divide com Google DeepMind, Meta AI e algumas empresas chinesas o protagonismo da inovação da IA generativa.

As grandes consultorias começam a questionar o real potencial da IA e os mercados financeiros especulam sobre um possível estouro da “bolha”, diante da incapacidade das empresas emergentes desenvolverem um modelo de negócios que monetize sua inovação, e mesmo o futuro das *BigTechs*, cujos lucros expressivos parecem incompatíveis com seu elevadíssimo investimento infraestrutural para garantir a demanda por serviços na nuvem da IA generativa (Goldman Sachs, 2024). O mais provável é que o momento atual corresponda a fase de consolidação e extração de valor e soerguimento de barreiras à entrada e eliminação de concorrentes, que se seguem às ondas de inovação e surgimento de novas empresas que criam valor e auferem rendas extraordinárias temporárias (O’Reilly, 2024).

Os episódios recentes corroboram, mais uma vez, que nem sempre o pioneiro na inovação é quem captura seu valor. A Microsoft, mesmo sem mostrar ao mundo modelos tão disruptivos, conseguiu superar o atraso em IA – e sua ausência quase completa na economia de plataformas – por meio de uma agressiva reestruturação estratégica, apostando em aquisição de empresas e até de repositórios *open-source*, respaldada em seu poderio financeiro e computacional, que a recolocou entre as maiores e mais dinâmicas empresas do mundo.

A disponibilidade de recursos computacionais passou a ser um ativo crucial para a inovação e apropriação de valor da IA, sobretudo como serviços na nuvem, permitindo o acesso

a grandes capacidades de processamento e armazenamento de dados. Autores como Taddy (2018) acreditam que técnicas de aprendizado de máquina poderão se transformar em *commodity* da computação em nuvem no curto prazo, enquanto Rikap (2022) considera que isso aprofundará os monopólios intelectuais das empresas líderes. O mais provável, entretanto, parece ser a provisão de uma série de atividades como serviços em computação na nuvem. *Everything as a Service (XaaS)* pelas *BigTechs* como *cloud service providers (CSP)*. Graux (2021) considera que apesar dos inegáveis benefícios para usuários, há riscos em modelos de negócios que passariam a ser baseado em instrumentos menos padronizados e mais incertos do que DPis como os contratos entre as partes.

No entanto, a arquitetura da “indústria” de IA e seu padrão de competição passam por um momento de ebulição e incerteza. Primeiro, ações antitruste e regulatórias sobre as empresas de plataformas digitais, agora estendidas ao negócio de computação na nuvem, que podem resultar no fracionamento das empresas e interromper sua trajetória inovadora em IA. Segundo, empresas incumbentes de tecnologia, como Qualcomm, HP, Oracle, Dell e IBM, e até fabricantes de chips, como Nvidia, estão reposicionando estratégias apoiadas na expansão da computação na nuvem, com potencial de rivalizar no mercado de IA. Terceiro, incertezas na oferta mundial de chips e a retomada da importância do hardware para o avanço da IA. Quarto, questões geopolíticas ligadas as disputas entre EUA e China nas novas tecnologias e a dependência crescente da segurança cibernética de empresas privadas, poderão ter consequências imprevisíveis sobre a inovação e padrão de competição na IA.

Não obstante, a história ainda não parece ter chegado ao fim. Porque o momento é de efetiva disputa pelo *design* dominante e ainda não há, com efeito, uma tecnologia que garanta a prometida ubiquidade da IA enquanto tecnologia de propósito geral. A aparente facilidade e baixo custo de utilização na Internet de aplicativos como ChatGPT, já automatizaram inúmeras tarefas, sendo previsto alcançar 30% delas até 2030 nos EUA (Elingrud et al., 2023).

O amplo uso da IA em variados setores da economia ainda não ocorreu, seja porque os sistemas computacionais ainda não oferecem um número suficiente de *cases* exitosos de aplicações no setor empresarial; seja porque este mesmo setor empresarial ainda resiste a empregar tecnologias não completamente testadas e, principalmente, diante da tão propalada não explicabilidade, transparência e confiabilidade da IA, baseada em modelos que erram, viesam, “alucinam” e vazam dados; seja, ainda, porque talvez ainda não sejam essas as tecnologias prontas para o mercado.

De fato, os sistemas de IA ainda são fenômenos de entretenimento, com uso restrito e limitado em poucos setores, herméticos e onde muitos usos possíveis sequer são conhecidos

ex-ante, tal como ocorreu com tecnologias de propósito geral precedentes, onde a disseminação pela economia e apropriação efetiva do valor da inovação levaram décadas. Nem toda a pesquisa e desenvolvimento visando inovações complementares e usos potenciais setoriais será realizada pelas empresas que criam a IA, pois como outras tecnologias de propósito geral envolvem *spillovers* dinâmicos que drenarão parte dos lucros, criando incentivos insuficientes para o investimento privado ao desafiar a apropriabilidade da inovação. Mesmo que o ritmo do progresso técnico da IA pareça muito mais rápido do que qualquer padrão histórico, ainda existe um enorme campo de estudo multidisciplinar sobre usos setoriais potenciais dessas tecnologias, que deveriam ensejar pesquisas acadêmicas e corporativas, hoje majoritariamente centradas nos riscos e danos potenciais da IA.

A IA é uma tecnologia em rápida evolução, de modo que em pouco tempo tudo o que foi escrito nesta tese pode se tornar obsoleto. O desafio atual parece ser desenvolver um algoritmo geral que aprenda a resolver tarefas para um extenso conjunto de aplicações, sem exigir adaptação e que possibilite adoção direta para os mais diversos propósitos e ambientes (Rikap, 2023b). As novas apostas incluem agentes de IA e a própria AGI, que muitos duvidam ser possível. Agentes de IA, sequer abordados nesta tese, são sistemas que podem interagir com o ambiente, por meio de coleta e uso de dados para realizar tarefas autodeterminadas para atingir metas predeterminadas pelo programador, com o agente capaz de escolher de forma independente as melhores ações necessárias para atingir as metas. Esses agentes representam avanços em relação aos *chatbots* conhecidos que não conseguem planejar com antecedência, nem responder perguntas específicas e com dados do usuário, não retêm memória, nem aprendem com erros. Os algoritmos de aprendizado por reforço são capazes de aplicação a tarefas semelhantes para as quais foram desenvolvidos, capaz de generalizar a partir da experiência e experimentação. O DreamerV3 (Google), por exemplo, é um algoritmo geral que supera métodos especializados capaz de realizar 150 tarefas com uma única configuração (The State of AI Report Compute Index, 2024). O recente GPT-4o1, apelidado *Strawberry*, anunciado como um suposto modelo de “raciocínio” e mais um passo rumo a AGI, foi lançado pela OpenAI. Mas talvez o avanço *downstream* somente seja alcançado quando a IA puder ser embarcada em tudo.

São essas (ou futuras) promessas que possibilitarão a efetiva implementação da IA em setores a jusante e que abrirão novos e complexos desafios de apropriabilidade da inovação. Isso implica em que, mesmo uma tecnologia poderosa como a IA, não parece capaz de suplantiar a investigação setorial específica para sua implementação. A literatura mais recente se debruçava exatamente sobre a implementação da IA em outros setores, regimes de

apropriabilidade e potenciais modelos de negócios, quando foi atropelada pela enxurrada de novos modelos generativos.

Por fim, essas tecnologias, que contemplam regimes híbridos de PI e cada vez mais inovação aberta, demandarão futura revisão institucional, inclusive pelo caráter cada vez mais complementar em nível global. TRIPS foi desenvolvido com base em uma visão da economia, sobretudo industrial, que não existirá no formato padrão e que ameaça pilares básicos como a territorialidade dos direitos. O ecossistema de inovação que caracterizou o avanço da IA ainda não levou a ajustes em nenhum regime de direito de propriedade intelectual, ou mesmo de uma política da concorrência que reconheça a importância da inovação aberta e mecanismos de criação e captura de valor dentro e entre ecossistemas.

A tese demonstrou como em um setor como a IA a abordagem da estruturação de um regime de apropriabilidade que enfatize a integração de ativos/tecnologias complementares em ecossistemas cada vez mais extensos é definitivamente crucial para a captura de valor e lucratividade a longo prazo da inovação. Mostrou como a construção de capacidades dinâmicas para a inovação é central em tecnologias que avançam em ritmo inédito frente aos padrões históricos, mas que estratégias empresariais bem-posicionadas possibilitaram superar o atraso e se manter na corrida tecnológica. Corroborou que políticas regulatórias podem modificar a estrutura da indústria, seu padrão de competição e o próprio padrão de inovação, moldando o regime de apropriabilidade vis-à-vis estratégias empresariais. Nenhuma forma de intervenção pública é neutra e, conseqüentemente, alteram possíveis vencedores e perdedores da corrida da IA.

De fato, as principais economias começam a implementar políticas públicas com vistas à construção de ecossistemas que permitam alcançar o protagonismo das empresas americanas e chinesas. O que não parece uma tarefa fácil. O valor em capitalização de mercado das quatro maiores BigTechs e da Nvidia só é superado por 10 economias – e a menor das BigTechs supera nada menos que quinze maiores economias do planeta. A União Europeia, originalmente focada na dimensão puramente regulatória, avança de forma acelerada na definição de políticas ativas para se tornar inovadora e produtora de IA, respaldada no princípio de soberania digital e restauração da competitividade do bloco econômico. Países como Reino Unido, Canadá, Ásia e mesmo Oriente Médio estão buscando políticas ativas para desenvolver IA a partir da estruturação de amplos ecossistemas, a despeito das atuais disputas geopolíticas e geoestratégicas.

Nesse sentido, a principal lição para as economias em desenvolvimento é a necessidade de implementar a regulação do uso dessa poderosa tecnologia, conjugada a políticas públicas

que possibilitem a construção de ecossistemas locais que interajam com os ecossistemas globais de inovação. A efetiva implementação da IA ainda está nos seus primórdios e ainda existem perspectivas de inserção e aproveitamento das oportunidades que as revoluções tecnológicas oferecem, revertendo a tendência à ampliação das desigualdades, mas que dependerão de políticas públicas capazes de desenvolver capacidades e inovações organizacionais de natureza híbrida, da qual participem *players* do governo, da indústria e da academia de IA, mas incorporem setores *downstream* usuários potenciais da IA, capazes de desenvolver tecnologias complementares, buscando inspiração nas recentes iniciativas de economias europeias com vistas a reverter o atraso na IA. Embora economias em desenvolvimento dificilmente tenham condições de (re) definir regras de propriedade intelectual e políticas da concorrência, é importante ter clareza que qualquer forma de intervenção pública afeta preços relativos e da imprescindibilidade de participarmos das regras e instituições que definirão os ciclos de desenvolvimento e crescimento das próximas décadas. Não existe determinismo tecnológico, mas o que os homens farão com a IA, longe de uma panaceia ou de uma caixa de pandora.

Não deveremos nos resignar em ser meros usuários de IA, mas ambicionar um lugar entre efetivos produtores e inovadores. Ainda que recorrentemente seja repetido que a IA aumentará a desigualdade, mas a efetiva desigualdade será entre países que desenvolvem IA e aqueles que não avancem nessas tecnologias com enorme potencial de transformação e mudança estrutural profunda dos setores econômicos e melhoraria de serviços públicos.

Junto com Chile e Uruguai, o Brasil está entre as economias pioneiras em capacitação, P&D, adoção e governança em IA na América Latina (ILIA, 2024). O país ocupa atualmente a vigésima primeira posição mundial em publicações científicas e a trigésima segunda posição em citações (2023) (Scimago journal & country rank), revertendo a partir de 2020 a expansão da década passada. Não faltam talentos, startups nascentes e importantes conjuntos de dados, sobretudo no setor público, embora as maiores (e mais dispendiosas) deficiências sejam relacionadas à infraestrutura computacional e a ausência de grandes corporações capazes de fazer frente aos elevados investimentos exigidos para treinamento de modelos de fronteira.

No final de julho passado, o governo federal anunciou o Plano Brasileiro de IA (PBIA), com investimentos de R\$ 23 bilhões até 2028, ambicioso para os padrões brasileiros, mas irrisório frente ao investimento privado, carro-chefe da IA em quase todo o mundo. As iniciativas infraestruturais são dispersas, contemplando R\$ 1,8 bilhão para modernização do supercomputador Santos Dumont do Laboratório Nacional de Computação Científica (LNCC), risível se comparado ao investimento de mais de R\$ 550 bilhões da Microsoft no supercomputador Stargate. Investimentos de R\$ 125 milhões para compra de

supercomputadores para cinco centros de computação regionais (CENAPADS) e R\$ 250 milhões para 25 projetos de aquisição, instalação e modernização de infraestrutura de IA (capacidade computacional e datacenters) nas Instituições de Ciência e Tecnologia brasileiras – suficiente para custear clusters de 45 GPUs em cada centro, inexpressivos comparado aos usados no treinamento dos modelos citados na Tabela 6.1.1.

O objetivo maior de participar do desenvolvimento da IA irá requerer um esforço concentrado e duradouro, baseado em estratégias híbridas que envolvam empresas potenciais usuárias de IA em setores *downstream*, diante dos inevitáveis *spillovers* dinâmicos característicos de tecnologias de propósito geral e da necessidade de investimentos complementares em áreas adjacentes. Parcerias com empresas estrangeiras parecem inevitáveis, mas não a qualquer preço.

O anúncio de investimentos de US\$ 2,7 bilhões da Microsoft e de US\$ 1,8 bilhão da Amazon em datacenters no Brasil são bem-vindos, mas não sem condicionalidades. Antes de mergulhar no entusiasmo de que esses investimentos representem nosso passaporte para a revolução tecnológica do século XXI, é preciso lembrar que a busca por espaço físico, energia renovável e água para resfriamento dos equipamentos é hoje prioridade para a expansão de grandes datacenters das BigTechs, mas as necessidades de energia e água deverão ser maiores em localidades com clima quente. O mais recente escrutínio público da IA relaciona-se aos seus impactos ambientais, tornando atraentes investimentos em países com matriz energética composta por fontes renováveis. Nesse sentido, exigir o uso de sistemas mais eficientes e técnicas alternativas para resfriamento de equipamentos, em conformidade com as recentes regulamentações internacionais, serão fundamentais para evitar que nosso ingresso na era da IA traga novos danos ambientais.

O país dispõe de energia renovável, bases de dados setoriais importantes, em setores prioritários para implementação da IA pelas grandes corporações, como educação, saúde e agricultura, cuja disponibilização deverá contemplar também condicionalidades, como a definição das áreas prioritárias para a pesquisa colaborativa do país.

Os prêmios Nobel de química e física de 2024 para pesquisadores em IA, como Geoffrey Hinton e Demis Hassabis, pareceram o reconhecimento definitivo da área e de sua importância para o avanço da ciência e do conhecimento, em uma área onde as fronteiras entre academia e empresas são tênues e cujas contribuições em termos dos modelos AlexNet e das várias versões do AlphaFold transcendem os limites corporativos do Google.

Referências

- ABDIN, M. et al. Phi Technical Report: A Highly Capable Language Model Locally on Your Phone. *ArXiv*. 2024. <https://arxiv.org/abs/2404.14219>
- ABBOTT, R.B. *The Reasonable Robot: Artificial Intelligence and the Law*. Cambridge University Press. 2020
- ABBOTT, R.B. I Think, Therefore I Invent: Creative Computers and the Future of Patent Law. *Boston College Law Review* 57(4), 2016. <http://dx.doi.org/10.2139/ssrn.2727884>. Acesso em: 19 ago. 2021. (2016a)
- ABBOTT, R.B. Hal the Inventor: Big Data and Its Use by Artificial Intelligence. In Sugimoto, C. R., Ekbis, H.R., Mattioli, M. (eds), *Big Data Is Not a Monolith*, MIT Press. 2016 <https://ssrn.com/abstract=2565950>. Acesso em: 19 ago. 2021. (2016b)
- ABRAMSON, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 2024, pp. 493–500. <https://www.nature.com/articles/s41586-024-07487-w>. Acesso em: 20 mai 2024
- ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: Survey on XAI. *IEEE Access*, Vol. 6, 2018, pp. 52138–52160. doi: 10.1109/ACCESS.2018.2870052
- ADIWARDANA D. et al. Towards a Human-like Open-Domain Chatbot. *ArXiv* 2020. <https://arxiv.org/abs/2001.09977>
- ADNER, R. Ecosystem as Structure: An Actionable Construct for Strategy. *Journal of Management* 43 (1), 2017, pp. 39–58. <https://doi.org/10.1177/0149206316678451>
- AFONJA, G. et al. The crossroads of innovation and privacy: private synthetic data for generative AI. May 29, 2024. <https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/>. Acesso em: 06 ago. 2024
- AGRAWAL, A.; GANS, J. e GOLDFARB, A. What to Expect from Artificial Intelligence. *MIT Sloan Management Review*. February 7, 2017. Spring. <https://sloanreview.mit.edu/article/what-to-expect-from-artificial-intelligence/>. Acesso em: 20 mai 2020
- ABDEL-HAMID, O. et al. Convolutional Neural Networks for Speech Recognition in *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (10), 2014, pp.1533–1545. doi: 10.1109/TASLP.2014.2339736.
- AFFELDT, P. e KESLER, R. Big Tech Acquisitions Towards Empirical Evidence. *Journal of European Competition Law & Practice* 12 (6), 2021, pp. 471–478. <https://doi.org/10.1093/jeclap/lpab025>
- AGARWAL, S. Google's next virtual assistant could chat your ear off on just about any topic. *Digital Trends* Jan 29, 2020. <https://www.digitaltrends.com/cool-tech/google-meena-chatbot-research/>. Acesso em: 13 jun.2024
- AHMED, N.; WAHED, M. e THOMPSON, N.C. The growing influence of industry in AI research. *Science* 379 (6635), 2023, pp.884–886. <https://science.org/doi/10.1126/science.ade2420>. Acesso em: 15 ma.2023
- AKHTAR, S.I.; SCHWARZENBERG, A.B. e TRACHTENBERG, D.M. U.S.-EU Trade and Technology Council: Background and Issues. In *Focus* March 27, 2024. Congressional Research Service <https://crsreports.congress.gov/product/pdf/IF/IF12575>
- ALBERGOTTI, R. OpenAI has received just a fraction of Microsoft's \$10 billion investment. *Semafor* Nov 18, 2023. <https://www.semafor.com/article/11/18/2023/openai-has-received-just-a-fraction-of-microsofts-10-billion-investment?ref=wheresyoured.at>. Acesso em: 20 nov.2023
- ALIZADEH, K. et al. LLM in a flash: Efficient Large Language Model Inference with Limited Memory. *ArXiv* 2024. <https://arxiv.org/abs/2312.11514>
- ALLISON, J. R. e LEMLEY, M. A. Empirical Evidence on the Validity of Litigated Patents. *American Intellectual Property Law Association (AIPLA) Quarterly Journal* V. 26, 1998, p. 185. <https://ssrn.com/abstract=118149>
- ALMEIDA, D.R. Synthetic data generation. *OpenAI* Apr 9, 2024. <https://cookbook.openai.com/examples/sdg1>. Acesso em: 11 abr.2024
- ALT, R. Electronic markets and current general research. *Electronic Markets* 28(2), 2018, pp. 123–128. <https://doi.org/10.1007/s12525-018-0299-0>
- ALLYN, B. 'New York Times' considers legal action against OpenAI as copyright tensions swirl. *NPR* Aug 16, 2023. <https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl>. Acesso em: 18 ago.2023
- AMADEO, R. Google's iron grip on Android: Controlling open source by any means necessary. *Ars technica* July 21, 2018. <https://arstechnica.com/gadgets/2018/07/googles-iron-grip-on-android-controlling-open-source-by-any-means-necessary/>. Acesso em: 20 jan. 2023
- AMARO, S. e KHARPAL, S. The U.S. imposed semiconductor export controls on China. Now a key EU nation is set to follow suit. *CNBC* March 9, 2023. <https://www.cnbc.com/2023/03/09/asml-netherlands-to-restrict-semiconductor-machine-exports-after-us-pressure.html>
- AMAZON. Amazon and Anthropic Announce Strategic Collaboration to Advance Generative AI. *Amazon* Sept 25, 2023. <https://press.aboutamazon.com/2023/9/amazon-and-anthropic-announce-strategic-collaboration-to-advance-generative-ai>. Acesso em: 02 mar 2024
- ANDERLJUNG, M. et al. Frontier AI Regulation: Managing Emerging Risks to Public Safety. *ArXiv*. 2023. <https://arxiv.org/abs/2307.03718>. Acesso em: 08 nov.2023

- ANDERSEN, B. If 'intellectual property rights' is the answer, what is the question? Revisiting the patent controversies. *Economics of Innovation and New Technology* 13(5), 2004, pp.417-442.
- ANDREONI, A.; CHANG H.J. e LABRUNIE, L. Natura non facit saltus: Challenges and opportunities for digital industrialisation across developing countries. *The European Journal of Development Research* 33, 2022, pp. 330-370. <https://doi.org/10.1057/s41287-020-00355-z>.
- ANTON, J.J.; GREENE, H. e YAO, D.A. (2006) Policy Implications of Weak Patent Rights. *Innovation Policy and the Economy* 6, 2006, pp. 1-26. DOI 10.1086/ipe.6.25056178
- APNEWS. Microsoft reports higher profits and revenue powered by cloud computing and AI Investments. *AP News* Oct. 24, 2023. <https://apnews.com/article/microsoft-earnings-artificial-intelligence-cloud-cce0a0de6080696bfc9f1b66eee3456f>. Acesso em: 12 nov. 2023
- ARCHIBUGI, D. e FILIPPETTI, A. Knowledge as Global Public Good in Archibugi, D. e Filippetti, A. (eds.): *Handbook of Global Science, Technology and Innovation*, Wiley, Oxford. 2015. <https://ssrn.com/abstract=2557339>. Acesso em: 10 set. 2022
- ARROW, K. Economic Welfare and the Allocation of Resources for Invention. Princeton University Press. 1962. <http://www.nber.org/chapters/c2144>
- AUTOR, D. et al. The Fall of the Labor Share and the Rise of Superstar Firms, *The Quarterly Journal of Economics* 135 (2), 2020, pp. 645–709, <https://doi.org/10.1093/qje/qjaa004>
- AUTORITÉ DE LA CONCURRENCE. On competition in the cloud sector. *Autorité de la concurrence* June 29, 2023. https://www.autoritedelaconcurrence.fr/sites/default/files/attachments/2023-09/23a08_EN.pdf
- AWS AWS global infrastructure. AWS, 2023. <https://aws.amazon.com/pt/about-aws/global-infrastructure/>. Acesso em: 23 ago. 2024
- AZHAR, A. et al. Generative AI's Impact on Computing Power: Lessons from a Bullish Model of Global Demand. *Exponential View*, Oct 1, 2024. <https://www.exponentialview.co/p/genai-impact-on-computing-power>. Acesso em: 20 out.2024.
- BA, J. M.; KIROS, J. R. e HINTON, G. E. Layer Normalization. *ArXiv* 2016. <https://arxiv.org/abs/1607.06450>
- BALDWIN, C.Y. e CLARK, K.B. Design Rules, Volume 1: The Power of Modularity. The MIT Press. 2000. <https://doi.org/10.7551/mitpress/2366.001.000>
- BALTRUSAITIS, T. AHUJA, C. e MORENCY, L-P Multimodal Machine Learning: A Survey and Taxonomy. *ArXiv* 2017. <https://arxiv.org/abs/1705.09406>
- BAPTISTA, E. Exclusive: China's ByteDance working with Broadcom to develop advanced AI chip. *Reuters* June 24, 2024. https://www.reuters.com/technology/artificial-intelligence/chinas-bytedance-working-with-broadcom-develop-advanced-ai-chip-sources-say-2024-06-24/?utm_source=opentoolsai-newsletter&utm_medium=newsletter&utm_campaign=new-ai-chip-collab
- Barcellos, P. C.A. e Bastos, V.D. Regulação da Inteligência Artificial: reflexões destacadas sob as perspectivas nacional e global. *Revista Compliance Rio* 6 (6), 2024, pp. 110-121. <https://icrio.org/wp-content/uploads/2024/03/Revista-ComplianceRio-2024-A.pdf>.
- BARON, J. e SPULBER, D. F. Technology Standards and Standard Setting Organizations: Introduction to the Searle Center Database. *Northwestern Law & Econ Research Paper* No. 17-16, 2018. <http://dx.doi.org/10.2139/ssrn.3073165>
- BARUFFALDI, S., et al. Identifying and measuring developments in artificial intelligence: Making the impossible possible. *OECD Science, Technology and Industry Working Papers*, No. 2020/05, OECD Publishing, Paris, 2020. <https://doi.org/10.1787/5f65ff7e-en>.
- BASS, D. Amazon enters chatbot wars with AI startup partnership. *Los Angeles Times*. Feb 21, 2023. <https://www.latimes.com/business/story/2023-02-21/amazons-aws-hugging-face-ai-deals-chatgpt>
- BASTOS, V.D. Artificial Intelligence: Intellectual Property Rights Dilemma. Paper presented at the 18th *International Joseph A. Schumpeter Society Conference*, Rome/Italy, July 8 – 10, 2021. <https://bit.ly/36bMjqn>
- BAWCOM, A. et al. Getting the most from cloud services and containers. *McKinsey Digital* July 19, 2022. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/getting-the-most-from-cloud-services-and-containers>? Acesso em: 12 jan. 2022
- BEKKERS, R. e UPDEGROVE, A. A study of IPR policies and practices of a representative group of Standards Setting Organizations worldwide. Prepared by Commissioned by the US National Academies of Science, Board of Science, Technology, and Economic Policy/Project on Intellectual Property Management in Standard Setting Processes. 2012. <https://dx.doi.org/10.2139/ssrn.2333445>
- BELTON, P. The computer chip industry has a dirty climate secret. *The Guardian* Sept. 18, 2021. <https://www.theguardian.com/environment/2021/sep/18/semiconductor-silicon-chips-carbon-footprint-climate>
- BENAICH, N. *State of AI Report* 2024. https://stateofai.substack.com/p/the-state-of-state-of-ai-report?utm_source=post-email-title&publication_id=1203391&post_id=141422815&utm_campaign=email-post-title&isFreemail=true&r=1ljt6w&utm_medium=email. Acesso em: 17 mar. 2024
- BENAICH, N. *State of AI Report* 2023. <https://www.stateof.ai/>. Acesso em: 17 out. 2023
- BENAICH, N. e HOGARTH, I. *State of AI Report* 2022. <https://www.stateof.ai/>. Acesso em: 13 mai. 2023.

- BENDER, E.M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT '21*, March 3–10, 2021. Virtual Event, Canada ACM ISBN 978-1-4503-8309-7/21/03. <https://doi.org/10.1145/3442188.3445922>
- BENGIO, Y. et al. Managing extreme AI risks amid rapid progress. *ArXiv* 2023. <https://arxiv.org/abs/2310.17688>. Acesso em: 30 out. 2023
- BERNIER, A. e THOROGOOD, A. Sharing Bioinformatic Data for Machine Learning: Maximizing Interoperability through License Selection. *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)* V. 3, 2020, pp.226-232. Valletta, Malta. <https://www.scitepress.org/Link.aspx?doi=10.5220/0009179502260232>. Acesso em: 31 jul. 2023.
- BESIROGLU, T. et al. The Compute Divide in Machine Learning: A Thread to Academic Contribution and Scrutiny. *ArXiv* 2024. <https://arxiv.org/abs/2401.02452>
- BESSY, C. The transformations of conventions for patent use and the role of legal intermediaries *ENS Paris Saclay IDHES*. 2019 <https://shs.hal.science/halshs-01872163/>
- BHATIA, R. So why did Google acquire Kaggle? *AIM* March 10, 2017. <https://analyticsindiamag.com/ai-origins-evolution/google-acquire-kaggle/>. Acesso: 03 mar.2023
- BICHICCHI, S. Intelligenza artificiale, fondi per 1 miliardo ma niente fondazione. Cosa c'è (e cosa no) nel disegno di legge del governo. *Milano Finanza* 24/4/2024. <https://www.milanofinanza.it/amp/news/intelligenza-artificiale-fondi-per-1-miliardo-ma-niente-fondazione-cosa-c-e-e-cosa-no-nel-disegno-di-202404241821487731>
- BIRCH, K., COCHRANE, D. e WARD, C. Data as asset? The measurement, governance, and valuation of digital personal data by Big Tech. *Big Data & Society* 8(1), 2021. <https://doi.org/10.1177/20539517211017308>
- BLIND, K. et al. The impact of Open Source Software and Hardware on technological independence, competitiveness and innovation in the EU economy, Final Study Report. Brussels. 2021. https://www.ospi.es/export/sites/ospi/documents/documentos/CNECT_OpenSourceStudy_EN_28_6_2021_LMBhSihnCeC7JEDsHXkK1JIZ0_79021_compressed.pdf
- BOGERS, M. The Open Innovation Paradox: Knowledge Sharing and Protection in R&D Collaborations. *European Journal of Innovation Management* 14 (1), 2011, pp. 93-117. <https://ssrn.com/abstract=1699380>
- BOMMASANI, R. et al. The Foundation Model Transparency Index. *ArXiv* 2023. <https://arxiv.org/abs/2310.12941> (2023a)
- BOMMASANI, R. et al. Decoding the White House AI Executive Order's Achievements. HAI Stanford University. 2023. <https://hai.stanford.edu/news/decoding-white-house-ai-executive-orders-achievements> (2023b)
- BOMMASANI, R. et al. On the Opportunities and Risks of Foundation Models. *ArXiv* 2022. <https://arxiv.org/abs/2108.07258>
- BOSTROM, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. May. 2016
- BOSTROM, N. How long before superintelligence? *Int. Jour. of Future Studies*, vol. 2, 1998. <https://nickbostrom.com/superintelligence>. Acesso em: 09 jun.2022
- BRADFORD, A. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. 2020
- BRADFORD, A. The Brussels Effect. *Northwestern University Law Review* 107 (1), Columbia Law and Economics Working Paper No. 533, 2012. <https://ssrn.com/abstract=2770634>
- BRESNAHAN, T. F. e TRAJTENBERG, M. General Purpose Technologies: Engines of Growth? *Journal of Econometrics* 65 (1), 1995, pp. 83-108. [https://doi.org/10.1016/0304-4076\(94\)01598-T](https://doi.org/10.1016/0304-4076(94)01598-T)
- BRESNICK, S.; LUONG, N. e CURLEE, K. Which Ties Will Bind? Big Tech, Lessons from Ukraine, and Implications for Taiwan. *Center for Security and Emerging Technology (CSET)*. Feb, 2024. Acesso em: 3 mar.2024
- BRICS. Final Report IP BRICS Offices. 2022. Disponível em: <http://www.ipbrics.net/secondpage/project/Patent%20Processes%20and%20Procedure%20-%20AI%20Study%20Report.pdf>. Acesso em: 12 nov. 2024
- BROWN, T. B. et al. Language Models are Few-Shot Learners. *ArXiv* 2020. <https://arxiv.org/abs/2005.14165>
- BRYNJOLFSSON, E. ROCK, D. e SYVERSON, C. Artificial Intelligence and the Modern Productivity Paradox A Clash of Expectations and Statistics. NBER. 2017. <http://www.nber.org/books/agra-1>
- BUAINAIN, A.M., PAULINO DE CARVALHO, S. M.; PAULINO, S.R. e YAMAMURA, S. Propriedade Intelectual e Inovação Tecnológica: Algumas questões para o debate atual. 2004
- BUBECK, S. et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *ArXiv*. 2023. <https://arxiv.org/abs/2303.12712>
- BUKHT, R. e HEEKS, R. Defining, Conceptualising and Measuring the Digital Economy. *Development Informatics Working Paper* no. 68, August 3, 2017. <https://ssrn.com/abstract=3431732>
- BUNDESREGIERUNG. Microsoft invests billions in Germany as an AI hub. Bundesregierung. Feb 15, 2024. <https://www.bundesregierung.de/breg-en/news/investment-worth-billions-in-ai-2260096>

- BURKACKY, O. et al. Generative AI: The next S-curve for the semiconductor industry? *McKinsey & Co* March 29, 2024. <https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry>. Acesso em: 21 mai.2024
- BURRI, M. Towards a New Treaty on Digital Trade. *Journal of World Trade* 55, 2021, p.1. <https://ssrn.com/abstract=3623734>
- BURT, J. Google Accused of Patent Infringement With Its TPU AI Chips. *Techstrong.ai*. Jan 16, 2024. <https://techstrong.ai/generative-ai/google-accused-of-patent-infringement-with-its-tpu-ai-chips/>
- CALVIN, N. e LEUNG, J. Who owns artificial intelligence? A preliminary analysis of corporate intellectual property strategies and why they matter. 2020. https://www.fhi.ox.ac.uk/wp-content/uploads/Patents_-FHI-Working-Paper-Final-.pdf. Acesso em: 10 ago. 2024
- CAMBURU, O. Explaining Deep Neural Networks. *ArXiv* 2020. <https://arxiv.org/abs/2010.01496>
- CAPPELLEN, F. Intellectual property strategies in the age of artificial intelligence. Master Thesis. Utrecht University. 2021. <https://studenttheses.uu.nl/handle/20.500.12932/41112>. Acesso em: 10 fev.2023
- CARABANTES, M. Black-box artificial intelligence: an epistemological and critical analysis. *AI & Soc* 35, 2020, pp. 309–317. <https://doi.org/10.1007/s00146-019-00888-w>
- CARMAN, A. e METZ, R. Major Record Labels Sue AI Music Generators. *Times* June 24, 2024. <https://time.com/6991466/record-labels-sue-ai-music-generator-startups/>
- CARVALHO, S.M.P. Propriedade Intelectual na Agricultura. Tese de Doutorado. IG/UNICAMP. Campinas/SP. Dez 2003. <https://repositorio.unicamp.br/acervo/detalhe/297436>. Acesso em: 30 mar. 2020
- CHANDLER Jr., A.D. *Strategy and Structure: Chapters in the History of the Industrial Enterprise*, The MIT Press. 1962
- CHANDLER Jr., A.D. *The Visible Hand: The Managerial Revolution in American Business*, Harvard University Press. 1977
- CHARNOVITZ, S. Patent Harmonization under World Trade Rules. *Journal of World Intellectual Property* 127, 1998. https://charnovitz.org/publications/Patent_Harmonization_Under_World_Trade_Rules.pdf
- CHAU, F. et al. Protecting Inventions Relating to Artificial Intelligence: Best Practices. 2022. https://ipo.org/wp-content/uploads/2022/02/AI-Patenting_white-paper_final.pdf
- CHAUDHURI, S. et al. Neurosymbolic Programming, Foundations and Trends in Programming Languages 7 (3), 2021, pp.158-243. <http://dx.doi.org/10.1561/25000000049>
- CHEE, F.Y. Microsoft's OpenAI investment could face EU merger probe, EU regulators say. *Reuters* Jan 9, 2024. <https://www.reuters.com/technology/microsofts-openai-investment-risks-eu-merger-probe-eu-regulators-say-2024-01-09/>. Acesso em: 09 jan.2024
- CHEN, M. et al. Evaluating Large Language Models Trained on Code. *ArXiv* 2021. <https://arxiv.org/abs/2107.03374>
- CHESBROUGH et al. *The Oxford Handbook of Open Innovation*. Oxford University Press. 2024
- CHESBROUGH, H.; LETTL, C. e RITTER, T. Value Creation and Value Capture in Open Innovation. *Journal of Product Innovation Management* 35 (6), 2018, pp. 930-938. <https://doi.org/10.1111/jpim.12471>
- CHESBROUGH, H. e APPLEYARD, M.M. Open Innovation and Strategy. *California Management Review*, 50(1), 2007, pp. 57-76. <https://doi.org/10.2307/41166416>
- CHESBROUGH, H. *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Review Press. 2003
- CHO, W. Authors Sue OpenAI Claiming Mass Copyright Infringement of Hundreds of Thousands of Novels. *The Hollywood Reporter*. June 29, 2023. <https://www-hollywoodreporter-com.cdn.ampproject.org/c/www.hollywoodreporter.com/business/business-news/authors-sue-openai-novels-1235526462/amp/>. Acesso em: 10 ago.2023
- CHO, K. et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv* 2014. <https://arxiv.org/abs/1406.1078>
- CHUVPILO, G. Who's Ahead in AI Research at NeurIPS 2020? Insights and AI Research Rankings at the Leading AI Conference. *Medium* Nov 17, 2020. <https://chuvpilo.medium.com/whos-ahead-in-ai-research-in-2020-2009da5cd799>. Acesso em: 10 jan. 2022
- CIMOLI, M.; DOSI, G. MAZZOLENI, R. e SAMPAT B. Innovation, technical change and patents in the development process: A long term view. *LEM Working Paper Series* 2011/06. March 7, 2011. ISSN (online) 2284-0400. <https://doi.org/10.1093/acprof:oso/9780199660759.003.0002>
- CINGARI, P. Who in Europe is investing the most in artificial intelligence? *Europe News*, April 29, 2024. <https://www.euronews.com/business/2024/04/29/who-in-europe-is-investing-the-most-in-artificial-intelligence>
- CLARK, K. e WEINBERG, C. SoftBak to invest US\$ 500 million in OpenAI. *The Information* Sept 30, 2024. <https://www.theinformation.com/articles/softbank-to-invest-500-million-in-openai>. Acesso 30 set.2024
- CMA. AI Foundation Models Technical update report. Competition and Markets Authority. Apr. London. UK. 2024. https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf. Acesso em: 02 jun.2024

- CMA. Online platforms and digital advertising. Market study final report. *Competition and Markets Authority* July 1, 2020. London. https://assets.publishing.service.gov.uk/media/5fa557668fa8f5788db46efc/Final_report_Digital_ALT_TEXT.pdf. Acesso em: 02 ago.2021
- CAN. Baidu says AI chatbot 'Ernie Bot' has amassed 200 million users. *Channel New Asia* April 16, 2024. <https://www.channelnewsasia.com/business/baidu-says-ai-chatbot-ernie-bot-has-amassed-200-million-users-4268301>. Acesso em: 30 jun. 2024
- CHIAO, J. Localization of Chip Manufacturing Rising. Taiwan to Control 48% of Global Foundry Capacity in 2022, Says TrendForce. *Trend Force* April 25, 2022. <https://www.trendforce.com/presscenter/news/20220425-11204.html>. Acesso em: 14 ago.2024
- CLOUD, P. G., CHANKONG, D. e HOPFINGER, T.M. AI Inventorship: Navigating Patent Rights Around the Globe. *Sterne, Kessler, Goldstein & Fox*. Aug 21, 2024. Disponível em: <https://www.sterneessler.com/news-insights/publications/ai-inventorship-navigating-patent-rights-around-the-globe/>. Acesso em: 12 set.2024
- COCKBURN, I. M.; HENDERSON, R. e STERN, S. The Impact of Artificial Intelligence on Innovation. *NBER Working Paper N. 24449*. Cambridge, MA. March, 2018. DOI 10.3386/w24449
- COHEN W. M.; NELSON, R.R. e WALSH, J.P. Protecting their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). *Working Paper 7552*. NBER. 2000. DOI 10.3386/w7552
- COLDEWEY, D. After raising \$1.3B, Inflection is eaten alive by its biggest investor, Microsoft. *TechCrunch* March 19, 2024. <https://techcrunch.com/2024/03/19/after-raising-1-3b-inflection-got-eaten-alive-by-its-biggest-investor-microsoft/>. Acesso em: 19 mar.2024
- COMPETITION POLICY BRIEF. Standard-essential patents. *European Commission/EU*. Issue 8, June, 2014. https://ec.europa.eu/competition/publications/cpb/2014/008_en.pdf. Acesso em: 20 mar.2021
- CONTRERAS, J. L. National disparities and standards-essential patents: considerations for India. University of Utah College of Law Research Paper No. 221, 2017. <https://ssrn.com/abstract=2995900>
- CONTRERAS, J. L. A Tale of Two Layers: Patents, Standardization, and the Internet. *Denver University Law Review* 93(4), 2016. <https://ssrn.com/abstract=2832538>
- COONS, C. Sens. Coons and Tillis and Reps. Collins, Johnson, and Stivers Release Section 101 Patent Reform Framework. 2019. <https://www.coons.senate.gov/news/press-releases/sens-coons-and-tillis-and-reps-collins-johnson-and-stivers-release-section-101-patent-reform-framework>
- COOPER, D. e SOMAINI, L. Italy Proposes New Artificial Intelligence Law. *Inside Global Tech* May 30, 2024. <https://www.insideglobaltech.com/2024/05/30/italy-proposes-new-artificial-intelligence-law/>
- COOPER, T. Google and other tech giants are quietly buying up the most important part of the internet. *Venture Beat* April 6, 2019. <https://venturebeat.com/commerce/google-and-other-tech-giants-are-quietly-buying-up-the-most-important-part-of-the-internet/>
- COOPER, D. et al. European Commission Announces New Package of AI Measures. *Inside Global Tech*. Feb 14, 2024. <https://www.insideglobaltech.com/2024/02/14/european-commission-announces-new-package-of-ai-measures/>
- COPPLE, B. Can you copyright the content you make with generative AI? *Descript* May 17, 2023. <https://www.descript.com/blog/article/what-copyright-law-says-about-ai-generated-content>. Acesso em: 27 out.2023.
- CORIAT, B. e WEINSTEIN, O. Intellectual Property Right Regimes, Firms and the Commodification of Knowledge. *CLPE Research Paper* 17/2009, V. 05 (3). <https://digitalcommons.osgoode.yorku.ca/cgi/viewcontent.cgi?article=1130&context=clpe>
- CORIAT, B. e ORSI, F. Establishing a new intellectual property rights regime in the United States: Origins, content and problems. December. *Research Policy* 31(8-9), 2002, pp. 491-1507. doi: 10.1016/S0048-7333(02)00078-1
- CORSO, J. Is Open Source AI Bull? Medium, Feb 1, 2024. Disponível em: <https://medium.com/@jasoncorso/is-open-source-ai-bull-9da010411658>. Acesso em: 25 OCT.2024
- CORUMBA, D. Intel Gaudi 3 custará \$15 mil, metade da NVIDIA Blackwell B100. *Adrenaline* June 7, 2024. <https://www.adrenaline.com.br/intel/intel-gaudi-3-custara-15-mil-metade-da-Nvidia-blackwell-b100/>. Acesso em: 14 ago.2024
- COTTIER, B.; BESIROGLU, T. e OWEN, D. Who is leading in AI? An analysis of industry AI research. *ArXiv*. 2023. <https://arxiv.org/abs/2312.00043>. Acesso em: 02 dez. 2023
- COTTIER, B. Trends in the Dollar Training Cost of Machine Learning Systems. *Epoch AI*. 2023. <https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems>. online
- COTTIER, B. The rising costs of training frontier AI models. *ArXiv* 2024. <https://arxiv.org/abs/2405.21015>.
- COYLE, D.; MANLEY, A. What is the Value of Data? A review of empirical methods. *Policy Brief*. Bennett Institute for Public Policy. Cambridge. July, 2022. https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/07/policy-brief_what-is-the-value-of-data.pdf

- COYLE, D.; DIEPEVEEN, S.; WDOWN, J.; KAY, L. e TENNISON, J. The Value of Data summary report 2020. Bennett Institute for Public Policy. Cambridge. 2020. https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2020/12/Value_of_data_summary_report_26_Feb.pdf
- CHRISTENSEN, C.M. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business Review Press. 1997
- CROITORU, F.A., HONDRU, V., IONESCU, R.T. E SHAH, M. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 45 (9): 10850- 10869, 2023, doi: 10.1109/TPAMI.2023.3261988
- CROVITZ, G. Who Really Invented the Internet? *The Wall Street Journal* July 22, 2012. <https://www.wsj.com/articles/SB10000872396390444464304577539063008406518>. Acesso em: 07 mai.2022
- CRAWFORD, K. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press. 2022
- CUNNINGHAM, C. et al. Killer acquisitions. *Journal of Political Economy* 129 (3), 2021, pp. 649–702. <http://dx.doi.org/10.2139/ssrn.3241707>
- CUSUMANO, M.A.; GAWER, A. e YOFFIE, D.B. The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power. Harper Business. 2019
- CUTOLO, D. e KENNEY, M. Platform-dependent entrepreneurs: Power asymmetries, risks, and strategies in the platform economy. *Academy of Management Perspectives* 35(4), 2021, pp. 584-605. <https://doi.org/10.5465/amp.2019.0103>. Acesso em: 07 ago. 2023
- CYTERA, C. Gallium, Germanium, and China — The Minerals Inflaming the Global Chip War. *CEPA* Aug 8, 2023. <https://cepa.org/article/china-gallium-and-germanium-the-minerals-inflaming-the-global-chip-war/>. Acesso em: 11 jul. 2024
- DAVE, P. OpenAI Agreed to Buy \$51 Million of AI Chips From a Startup Backed by CEO Sam Altman. *Wired* Dec. 3, 2023. <https://www.wired.com/story/openai-buy-ai-chips-startup-sam-altman/>. Acesso em: 04 dez. 2023
- DAVID, P.A. Heroes, Herds and Hysteresis in Technological History: Thomas Edison and 'The Battle of the Systems' Reconsidered. *Industrial and Corporate Change* 1(1), 1992, pp. 129-180 DOI:10.1093/icc/1.1.129
- DAVIES, E. Mathematics, word problems, common sense, and artificial intelligence. *ArXiv* 2023. <https://arxiv.org/pdf/2301.09723>
- DEBUSSCHE, J. e CÉSAR, J. Big Data & Issues & Opportunities: Intellectual Property Rights. Bird&Bird. News Centre. 2019. <https://www.twobirds.com/en/news/articles/2019/global/big-data-and-issues-and-opportunities-ip-rights>. Acesso 22/06/2020
- DEEPEARNING BOOK. Capítulos 85 a 90. Transformers. *Data Science Academy*. 2022. Disponível em: <https://www.deeplearningbook.com.br/transformadores-o-estado-da-arte-em-processamento-de-linguagem-natural/>. Acesso em: 08 ago 2024.
- DEN UIJL, S.; BEKKERS, R. e DE VRIES, H. J. Managing Intellectual Property Using Patent Pools: Lessons from Three Generations of Pools in the Optical Disc Industry. *California Management Review*, 55(4), 2013, pp. 31–50. <https://doi.org/10.1525/cmr.2013.55.4.31>
- DENG, H. Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer, Geneva, Switzerland: UNIDIR. 2023. https://unidir.org/wp-content/uploads/2023/11/UNIDIR_Exploring_Synthetic_Data_for_Artificial_Intelligence_and_Autonomous_Systems_A_Primer.pdf. Acesso em: 15 mar.2024
- DICKSON, B. DeepMind's big losses, and the questions around running an AI lab. *Venture Beat* Dec. 27, 2020. <https://venturebeat.com/ai/deepminds-big-losses-and-the-questions-around-running-an-ai-lab/>. Acesso em: 30 dez. 2020
- DIGITAL WATCH OBSERVATORY. Germany announces €1.75 Billion in funding to help high-tech growth. Feb 9, 2024. <https://dig.watch/updates/germany-announces-e1-75-billion-in-funding-to-help-high-tech-growth>.
- DILLET, R. Kyutai is a French AI research lab with a \$330 million budget that will make everything open source. *TechCrunch* Nov 17, 2023. <https://techcrunch.com/2023/11/17/kyutai-is-an-french-ai-research-lab-with-a-330-million-budget-that-will-make-everything-open-source/>. Acesso em : 18 nov.2023
- DOSI, G. et al. Varieties of deindustrialization and patterns of diversification: Why microchips are not potatoes chips. *LEM Working Paper Series, No. 2020/11*, Scuola Superiore Sant'Anna, Laboratory of Economics and Management (LEM), Pisa. 2021. <https://www.econstor.eu/bitstream/10419/228150/1/1698304846.pdf>
- DOSI, G. Technological paradigms and technological trajectories. *Research Policy* 11 (3), 1982, pp. 147-162. [https://doi.org/10.1016/0048-7333\(82\)90016-6](https://doi.org/10.1016/0048-7333(82)90016-6)
- DOSI, G. e STIGLITZ, J. The role of intellectual property rights in the development process, with some lessons from developed countries: An introduction. *LEM Working Paper Series N. 2013/23*. <https://www.econstor.eu/bitstream/10419/89516/1/771928769.pdf>
- DOSI, G.; MARENGO, L. e PASQUALI, C. How much should society fuel the greed of innovators? On the relations between appropriability, opportunities and rates of innovation. *LEM Working Paper Series, N. 2006/17*. <https://www.econstor.eu/bitstream/10419/89527/1/51846122X.pdf> (2006a)

- DOSI, G.; MALERBA, F.; RAMELLO, G.B. e SILVA, F. Information, appropriability, and the generation of innovative knowledge four decades after Arrow and Nelson: an introduction. *Industrial and Corporate Change* 15(6), 2006, pp. 891–901. <https://doi.org/10.1093/icc/dtl028> (2006b)
- DOSI, G. e NELSON, R.R. Technical Change and Industrial Dynamics, in Hall, B.; Rosenberg, N. (eds.), *Handbook of the Economics of Innovation*, North-Holland, Volume 1, Chapter 3, pp. 51-127. 2010. [https://doi.org/10.1016/S0169-7218\(10\)01003-8](https://doi.org/10.1016/S0169-7218(10)01003-8)
- DOSOVITSKIY, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* 2021. <https://arxiv.org/abs/2010.11929>
- DRAGHI, M. The future of european competitiveness. Part A. Sept. 2024. https://commission.europa.eu/document/download/97e481fd-2dc3-412d-be4c-f152a8232961_en?filename=The%20future%20of%20European%20competitiveness%20_%20A%20competitiveness%20strategy%20for%20Europe.pdf. Acesso em: 26 set.2024 (2024a)
- DRAGHI, M. The future of european competitiveness. Part B. Sept. 2024 https://commission.europa.eu/document/download/ec1409c1-d4b4-4882-8bdd-3519f86bbb92_en?filename=The%20future%20of%20European%20competitiveness_%20In-depth%20analysis%20and%20recommendations_0.pdf (2024b)
- DRAHOS, P. Global Property Rights in Information: The Story of TRIPS at the GATT. *Prometheus* 13 (1), 1995, pp. 6-19. <https://doi.org/10.1080/08109029508629187>
- DREXL J. et al. Comments of the Max Planck Institute for Innovation and Competition on the Draft Issues Paper of the World Intellectual Property Organization on Intellectual Property Policy and Artificial Intelligence. *Max Planck Institute for Innovation and Competition*. February 20, 2020. https://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/2020-02-11_WIPO_AI_Draft_Issue_Paper_Comments_Max_Planck.pdf. Acesso em: 07 ago. 2023.
- DREXL, J. et al. Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Lay Perspective. *Max Planck Institute for Innovation and Competition Research Paper Series*. Max Planck Institute for Innovation and Competition Research Paper No. 19-13, October, 2019. https://www.zora.uzh.ch/id/eprint/179880/1/Drexl_Hilty_et_al_MPI_Working_Paper_19-13_2019.pdf. Acesso em: 10 set. 2020.
- DU, N. et al. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. *ArXiv* 2022. <https://arxiv.org/abs/2112.06905>
- DUHIGG, C. e LOHR, S. The patent used as a sword. *The New York Times*. Oct 7, 2012. <https://www.nytimes.com/2012/10/08/technology/patent-wars-among-tech-giants-can-stifle-competition.html>. Acesso em: 20 ago. 2023
- DYER-WITHEFORD, N. et al. *Inhuman Power: Artificial Intelligence and the Future of Capitalism*. Pluto Press. June, 2019..
- EBRAHIM, T. Artificial Intelligence Inventions and Patent Disclosure. *Penn State Law Review* 125 (1). *U Iowa Legal Studies Research Paper* No. 2021-48. SSRN 2021. <https://ssrn.com/abstract=3722720>
- EFI Report. Report on Research, Innovation and Technological Performance in Germany. 2022. https://www.efi.de/fileadmin/Assets/Gutachten/2022/EFI_Report_2022.pdf. Acesso em 25 set. 2024
- EDLER, J. et al. Technology sovereignty as an emerging frame for innovation policy. Defining rationales, ends and means, *Research Policy* 52 (6), 2023. <https://doi.org/10.1016/j.respol.2023.104765>.
- EDLER, J. et al. Technology Sovereignty as an Emerging Frame for Innovation Policy – Defining Rationales, Ends and Means. *Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis* No. 70. July 2021. https://www.isi.fraunhofer.de/content/dam/isi/dokumente/cc/i/innovation-systems-policy-analysis/2021/discussionpaper_70_2021.pdf. Acesso em: 22 ago.2022
- EISENHARDT, K.M. Building Theories from Case Study Research. *The Academy of Management Review* 14 (4), 1989, pp. 532-550. <https://doi.org/10.2307/258557>
- EISENHARDT, K.M. e MARTIN, J.A. Dynamic Capabilities: What are they? *Strategic Management Journal* 21, 2000, pp. 1105–1121. [https://doi.org/10.1002/1097-0266\(200010/11\)21:10<1105::AID-SMJ133>3.0.CO;2-E](https://doi.org/10.1002/1097-0266(200010/11)21:10<1105::AID-SMJ133>3.0.CO;2-E)
- EPO. Artificial intelligence. *European Patent Office*. 2022. <https://www.epo.org/news-events/in-focus/ict/artificial-intelligence.html>. Acesso em: 28 jan 2023
- ERIXON, F. e BAUER, M. Standard Essential Patents and the Quest for Faster Diffusion of Technology. *Elipe Policy Brief* 2. European Centre for International Political Economy. 2017. ecipe.org/wp-content/uploads/2017/10/Standard-Essential-Patents-and-the-Quest-for-Faster-Diffusion-of-Technology.pdf
- ERNST, D. Competing in Artificial Intelligence Chips: China’s Challenge amid Technology War. *Centre for International Governance Innovation*. Special Report. March 30, 2020. <https://ssrn.com/abstract=3564669>
- EU AI ACT (21/05/2024) Regulation of The European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations. European Union. Disponível em: <https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf>. Acesso em: 22 mai 2024

- EU COMMISSION EPSC. European Commission: European Political Strategy Centre. *Rethinking strategic autonomy in the digital age*. 2019. <https://data.europa.eu/doi/10.2872/231231>
- EUROPEAN COMMISSION. European Commission: Joint Research Centre, *Mapping EU level funding instruments to Digital Decade targets – Application to main digital instruments in 2014-2027*, LÓPEZ COBO, M.(editor) and DE PRATO, G.(editor), Publications Office of the European Union, 2023. <https://data.europa.eu/doi/10.2760/986930>
- EUROPEAN COMMISSION. Trends and Developments in Artificial Intelligence: Challenges to the Intellectual Property Rights Framework. Final report. Joint Institute for Innovation Policy. Disponível em: https://www.ivir.nl/publicaties/download/Trends_and_Developments_in_Artificial_Intelligence-1.pdf. Acesso em: 12 jul. 2021.
- EVANSON, N. Report claims that OpenAI has burned through \$8.5 billion on AI training and staffing, and could be on track to make a \$5 billion loss. *PCGamer* July 29, 2024. <https://www.pcgamer.com/software/ai/report-claims-that-openai-has-burned-through-dollar85-billion-on-ai-training-and-staffing-and-could-be-on-track-to-make-a-dollar5-billion-loss/>. Acesso em: 12 set.2024
- EXAME. Google compra empresa brasileira de busca.Out 9, 2008. <https://exame.com/tecnologia/google-compra-empresa-brasileira-de-busca-m0061876/>. Acesso em: 29 set.2024
- FEFER, R.F. EU Digital Policy and International Trade. *Congressional Research Service R46732*. March 25, 2021. <https://crsreports.congress.gov/product/pdf/R/R46732>
- FERRANDIS, C.M. e LIZARRALDE, M.D. Open Sourcing AI: Intellectual Property at the Service of Platform Leadership. SSRN 2022. <https://ssrn.com/abstract=4018413>
- FERNANDEZ, R. et al. The financialisation of Big Tech. Dec 17, 2020. <https://www.somo.nl/the-financialisation-of-big-tech/>. Acesso 21 ago. 2024.
- FERRARI, F. Neural production networks: AI's infrastructural geographies. *Environment and Planning F* 2(4), 2023, pp. 459-476. <https://doi.org/10.1177/26349825231193226>. Acesso: 03 set.2023
- FERRÁS-HERNÁNDEZ, X., NYLUND, P. A. e BREM, A. The Emergence of Dominant Designs in Artificial Intelligence. *California Management Review*, 65(3), 2023, pp. 73-91. <https://doi.org/10.1177/00081256231164362>
- FERREIRA, T. Microsoft acusa Google de vantagem competitiva em IA generativa. *Olhar Digital* 15/03/2024. <https://olhardigital.com.br/2024/03/15/pro/microsoft-acusa-google-de-vantagem-competitiva-em-ia-generativa/>
- FINK, C. et al. Submarine trademarks. *Journal of Economics & Management Strategy* 31(4), 2022, pp. 818-840. <https://doi.org/10.1111/jems.12480>. Acesso em: 05 jul. 2023
- FISCHER, M. AI, Standards and Patents: Time for a Gold Rush. *Venner Shipley*. November 27, 2020.
- FLAHERTY, N. ASML sees export controls on latest DUV lithography equipment. *EE News Europe* June 30, 2023. <https://www.eenewseurope.com/en/asml-sees-export-controls-on-latest-duv-lithography-equipment/>. Acesso em: 18 ago.2024
- FLORIDI, L. AI and Its New Winter: from Myths to Realities. *Philos. Technol* 33, 2020, pp. 1-3. <https://doi.org/10.1007/s13347-020-00396-6>
- FOOTE, K.D. A brief history of cloud computing. *Dataversity*. 2021. <https://www.dataversity.net/brief-history-cloud-computing/>. Acesso em: 10 set.2024
- FORCE DISTANCE TIMES. The US Is Vulnerable in Critical Minerals. But There Is a Solution. Oct 24, 2022. <https://forcedistancetimes.com/us-critical-mineral-dependence-china-gallium-production/>. Acesso em: 12 jul. 2024
- FORD, M. Rule of the Robots: How Artificial Intelligence Will Transform Everything. Basic Books. 2021
- FOSS, N.J.; SCHMIDT, J. e TEECE, D.J. Ecosystem leadership as a dynamic capability, *Long Range Planning* 56 (1), 2023. <https://doi.org/10.1016/j.lrp.2022.102270>.
- FRANEK, K. What Companies Google & Alphabet Own: Visuals & Full List. Oct 16, 2021. <https://www.kamilfranek.com/what-companies-alphabet-google-owns/>. Acesso em: 15 jul.2022
- FRANK, R. H. e COOK, P. J. Winner-Take-All Markets. *Studies in Microeconomics* 1(2), 2013, pp. 131-154. <https://doi.org/10.1177/2321022213501254>
- FRIEDER, S. et al. Mathematical Capabilities of ChatGPT. *ArXiv* 2023. <https://arxiv.org/abs/2301.13867>
- FTC. FTC Launches Inquiry into Generative AI Investments and Partnerships. *FTC*. Jan 25, 2024. <https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships>
- FTC. Generative AI Raises Competition Concerns. *FTC Technology blog*. June 29, 2023. <https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns>. (2023a)
- FTC. Chatbots, deepfakes, and voice clones: AI deception for sale. *FTC Business blog* Mar 20, 2023. <https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale> (2023b)

- FULCRUM. Nortel Bankruptcy Patent Auction Is Biggest Technology Patent Sale Ever. Financial Complexity Made Clear. *Free Initial Consultation* (213) 787-4100. 2011. https://www.fulcrum.com/nortel_bankruptcy_patent_auction/. Acesso em: 10 out. 2022
- GAMBARDELLA, A.; HEATON, S.; NOVELLI, E. e TEECE, D.J. Profiting from Enabling Technologies? *Strategy Science* 6(1), 2021, pp. 75-90. <https://doi.org/10.1287/stsc.2020.0119>
- GANGULI, D. et al. Predictability and Surprise in Large Generative Models. *FACCT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. pp. 1747 – 1764. <https://doi.org/10.1145/3531146.3533229>
- GANT, B. Patents Are Not Probabilities: Refuting the Probabilistic Patent Theory. *Chicago-Kent Journal of Intellectual Property* 20. 2021. <https://scholarship.kentlaw.iit.edu/ckjip/vol20/iss2/3>. Acesso em: 29 set.2024
- GAO, H. S. Data Sovereignty and Trade Agreements: Three Digital Kingdoms. In Anupam Chander & Sun Haochen (ed.), *Data sovereignty: From the digital Silk Road to the return of the state* (pp. 213-239) Oxford: Oxford University Press. 2021. <http://dx.doi.org/10.2139/ssrn.3940508>
- GARCEZ, A.D.; LAMB, L. e GABBAY, D. *Neural-Symbolic Cognitive Reasoning*. New York: Springer. 2009
- GATES, B. The Age of AI has begun. *Gates Notes* March 21, 2023. <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun>. Acesso em: 28 ago.2023
- GAWER, A. Digital platforms and ecosystems: remarks on the dominant organizational forms of the digital age. *Innovation* 24(1), 2021, pp. 110–124. <https://doi.org/10.1080/14479338.2021.1965888>
- GAWER, A. Bridging differing perspectives on technological platforms: Toward an integrative framework. *Research Policy* 43(7), 2014, pp. 1239-1249. <https://doi.org/10.1016/j.respol.2014.03.006>
- GEIGER, C., FROSIO, G. e BULAYENKO, O. The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects. Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02. <http://dx.doi.org/10.2139/ssrn.3160586>. Acesso em: 07 jun. 2023.
- GEMINI TECHNICAL REPORT. 2024 Gemini: A Family of Highly Capable Multimodal Models. Google DeepMind. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.
- GERVAIS, D.J. Is Intellectual Property Law Ready for Artificial Intelligence? *GRUR International* 69 (2), 2020, pp. 117-118 . <https://doi.org/10.1093/grurint/ikz025>
- GERVAIS, D.J. Exploring the Interfaces Between Big Data and Intellectual Property Law, *10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law* 3, 2019. Disponível em: <https://scholarship.law.vanderbilt.edu/faculty-publications/1095>. Acesso em: 10 mar 2021.
- GERVAIS, D.J. The Protection of Databases, *82 Chicago-Kent Law Review*. 1109, 2007. Disponível em: <https://scholarship.law.vanderbilt.edu/faculty-publications/839>. Acesso em: 10 mar 2021.
- GERVAIS, D.J. e NAY, J.J. Artificial intelligence and interspecific law. *Science* 382 (6669), 2023, pp. 376-378. <https://doi.org/10.1126/science.adi8678>
- GHAFALE, R. T. (2021) The Role of Standards and Patents in Artificial Intelligence. *Ox First* March 24. <https://www.linkedin.com/pulse/role-standards-patents-artificial-intelligence-roya-ghafele>
- GIBSON DUNN In Three New Rules, BIS Continues Efforts to Reshape Global Semiconductor Supply Chains and the Future of Artificial Intelligence Training and Supercomputing. 2023. <https://www.gibsondunn.com/in-three-new-rules-bis-continues-efforts-to-reshape-global-semiconductor-supply-chains-and-future-of-artificial-intelligence-training-and-supercomputing/>. Acesso em: 02 jul. 2024
- GIL, M.A. “Precisamos parar de falar do ChatGPT”, diz Amy Webb. *Época Negócios* Março 7, 2023. https://epocanegocios.globo.com/tecnologia/noticia/2023/03/precisamos-parar-de-falar-do-chatgpt-diz-amy-webb.ghtml?fbclid=IwAR18nYwoluMTKVLjXJx2Pfm4vGbp0AIsD3irkxOWGLx2jsMqrDqSfW_xmd8. Acesso em: 07 mar.2023
- GODOY, A. S. (1995) Pesquisa qualitativa. *Revista de Administração de Empresas* 35 (3):20–29. <https://doi.org/10.1590/S0034-75901995000300004>
- GOERTZEL, B. How coined the term AGI. 2011. <http://goertzel.org/who-coined-the-term-agi/>. Acesso em: 19 abr 2024
- GOERTZEL, B. e PENNACHIN, C. Artificial General Intelligence. Springer. 2007. <https://link.springer.com/book/10.1007/978-3-540-68677-4>
- GOLDMAN, S. Sam Altman wants up to \$7 trillion for AI chips. The natural resources required would be ‘mind boggling’. *Venture Beat* Feb 9, 2024. <https://venturebeat.com/ai/sam-altman-wants-up-to-7-trillion-for-ai-chips-the-natural-resources-required-would-be-mind-boggling/>. Acesso em: 09 fev.2024
- GOLDMAN SACHS. Gen AI: Too much spend, too little benefit? *Top of Mind* Issue 129, June 25, 2024. https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend,-too-little-benefit-TOM_AI%202.0_ForRedaction.pdf. Acesso em: 18 jul.2024
- GOLDMAN SACHS. AI investment forecast to approach \$ 200 billion globally by 2025. Goldman Sachs 2023. <https://www.goldmansachs.com/insights/articles/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html>

- GOLLER, C. e KUCHLER, A. Learning task-dependent distributed representations by backpropagation through structure. *Proceedings of International Conference on Neural Networks (ICNN'96)*, Washington, DC, USA, 1996, vol 1, pp. 347-352. doi: 10.1109/ICNN.1996.548916.
- GOODFELLOW, I.J. et al. Generative Adversarial Networks. *ArXiv* 2014. <https://arxiv.org/abs/1406.2661>
- GOOGLE DEEPMIND. Gemini: A Family of Highly Capable Multimodal Models. *Google DeepMind*. 2024. https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf. Acesso em: 22 ago.2024 (2024a)
- GOOGLE DEEPMIND. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv* 2024. <https://arxiv.org/abs/2403.05530>. Acesso em: 10 set.2024 (2024b)
- GOOGLE. Google Announces Expansion of AI Partnership with Anthropic. *Google Cloud*. Nov 8, 2023. <https://www.googlecloudpresscorner.com/2023-11-08-Google-Announces-Expansion-of-AI-Partnership-with-Anthropic>
- GOOGLE DEEPMIND. Announcing Google DeepMind. April 20, 2023. <https://deepmind.google/discover/blog/announcing-google-deepmind/>
- GOTZEN, F. The industrial property right protection of the producer of a database - Some reflections on the future of European and Japanese protection schemes. *IIP Bulletin* 2007. Disponível em: https://www.iip.or.jp/e/summary/pdf/detail2006/e18_15.pdf. Acesso em: 10 dez 2023
- GRAEF, I. The Opportunities and Limits of Data Portability for Stimulating Competition and Innovation. *Competition Policy International, Antitrust Chronicle*, November. SSRN 2020 <https://ssrn.com/abstract=3740185>
- GRAHAM, S. e VISHNUBHAKAT, S. Of Smart Phone Wars and Software Patents. *Journal of Economic Perspectives* 27 (1), 2013, pp. 67–86. DOI: 10.1257/jep.27.1.67
- GRANSTRAND, O. Corporate Innovation Systems: A Comparative Study of Multi-Technology Corporations in Japan, Sweden and the USA. *Paper submitted to the Dynacom project* 2020. https://www.lem.sssup.it/Dynacom/files/D21_0.pdf
- GRANSTRAND, O. e HOLGERSSON, M. Innovation ecosystems: A conceptual review and a new definition. *Technovation*, 90-91, 2020. <https://doi.org/10.1016/j.technovation.2019.102098>
- GRAUX, H. Data Sharing as a Service: Will Data services remove intellectual property rights from the picture, and at what cost? 2021. <https://data.europa.eu>
- GRISHIKASHVILI, K.; DIBB, S. e MEADOWS, M. Investigation into BigData Impact on Digital Marketing. *Online Journal of Communication and Media Technologies Special Issue*. October 2014. <https://www.ojcm.net/download/investigation-into-big-data-impact-on-digital-marketing-5702.pdf>
- GROSSMAN, A.B.; BLEVINS, E.G. e SUTTER, K.M. Semiconductors and the Semiconductor Industry. *CRS Reports R47508*. April 19, 2023. <https://crsreports.congress.gov/product/pdf/R/R47508>. Acesso em: 02 out. 2023
- GUALTIERI, M. et al. The Forrester Wave: AI Infrastructure Solutions, Q1 2024. *Forrester* March 17, 2024. <https://reprints2.forrester.com/#/assets/2/157/RES180430/report>
- GUNASEKAR, S. et al. (2023) Textbooks Are All You Need. *ArXiv*. <https://arxiv.org/abs/2306.11644>
- HACKER, P.; ENGEL, A. e MAUER, M. Regulating ChatGPT and other Large Generative AI Models. *FACCT '23*, June 12–15, 2023, Chicago, IL, USA. <https://doi.org/10.1145/3593013.3594067>
- HALL, B.H. e ZIEDONIS, R.M. The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979-1995. *The RAND Journal of Economics* 32 (1), 2001, pp. 101-128
- HALL, A.J. Open-Source Licensing and Business Models: Making Money by Giving it Away, *Santa Clara High Tech* 33 L.J. 427, 2017. <https://digitalcommons.law.scu.edu/chtj/vol33/iss3/3>
- HAMMOND, G. Big Tech outspends venture capital firms in AI investment frenzy. *Financial Times* Dec 29, 2023. <https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532>. Acesso em: 30 dez.2023
- HANDLER, S. AI-Assisted Inventions Could Spur New Patent Litigation Wave. *Bloomberg Law*. August 9, 2022. <https://news.bloomberglaw.com/ip-law/ai-patent-ruling-could-spur-new-wave-of-litigation>. Acesso em: 10 set.2022
- HARRIS, L.A. e JAIKARAN, C. Highlights of the 2023 Executive Order on Artificial Intelligence for Congress. Congressional Research Service. R47843. April 3, 2024. <https://crsreports.congress.gov/product/pdf/R/R47843>
- HARRIS, L.A. Semiconductors and Artificial Intelligence. Congressional Research Service. *In Focus* Sept 18, 2023. <https://crsreports.congress.gov/product/pdf/IF/IF12497>.
- HARTMANN, P. e HENKEL, J. The Rise of Corporate Science in AI: Data as a Strategic Resource. *Academy of Management Discoveries* 6 (3), 2020. <https://doi.org/10.5465/amd.2019.0043>
- HAWKINS, M. House Republicans demand full Huawei sanctions after chip breakthrough. *Bloomberg* Sept 14, 2023. <https://www.bloomberg.com/news/articles/2023-09-14/republicans-demand-full-huawei-sanctions-after-chip-breakthrough>
- HEATH, A. Meta is putting AI chatbots everywhere. *The Verge* Sept. 27, 2023. <https://www.theverge.com/2023/9/27/23891128/meta-ai-assistant-characters-whatsapp-instagram-connect>. Acesso em: 12 out.2023

- HELFAT, C.E. e QUINN, J.B. Open Innovation: The New Imperative for Creating and Profiting from Technology. *Academy of Management Perspectives* 20 (2), 2006. <https://doi.org/10.5465/amp.2006.20591014>
- HELMOND, A. The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society* 1(2), 2015. <https://doi.org/10.1177/2056305115603080>
- HENDERSON, P. et al. Foundation Models and Fair Use. *ArXiv* 2023. <https://arxiv.org/abs/2303.15715>
- HEMPHILL, C. S. e WU, T. Nascent Competitors. *University of Pennsylvania Law Review*, 1879. 2020. https://scholarship.law.columbia.edu/faculty_scholarship/2661
- HERMAN, M. Sensible Open Source Licenses For Standards Development Organizations. SSRN 2020. <http://dx.doi.org/10.2139/ssrn.3717031>
- HERVEY, M. AI and Copyright in 2022. Kluwer Copyright Blog. Feb 2, 2023. Disponível em: <https://copyrightblog.kluweriplaw.com/2023/02/02/ai-and-copyright-in-2022/>. Acesso em:
- HESSLOW, D. et al. RITA: a Study on Scaling Up Generative Protein Sequence Models. *ArXiv* 2022. <https://arxiv.org/abs/2205.05789>
- HILTY, R. M.; HOFFMANN, J. e SCHEUERER, S. Intellectual Property Justification for Artificial Intelligence. *Max Planck Institute for Innovation and Competition Research Paper No. 20-02*. 2020. Max Planck Institute for Innovation and Competition Research Paper Series.
- HINTON, G. The Deep Learning Revolution. Turing Award Lecture. 2021. Disponível em: <https://www.youtube.com/watch?v=-cv0ddgcImk>. Acesso em 10 jan. 2022
- HINTON, G. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups in *IEEE Signal Processing Magazine* 29 (6), 2012, pp. 82-97. doi: 10.1109/MSP.2012.2205597.
- HINTON, G. Deep belief networks. *Scholarpedia* 4(5):5947. 2009. <http://dx.doi.org/10.4249/scholarpedia.5947>
- HO, J.; JAIN, A. e ABBEEL, P. Denoising Diffusion Probabilistic Models. *ArXiv* 2020. <https://arxiv.org/abs/2006.11239>
- HOANG, L. Big Tech warns Vietnam data rules risk 'damage' to digital economy. *Nikkei Asia* Nov 17, 2021. <https://asia.nikkei.com/Business/Technology/Big-Tech-warns-Vietnam-data-rules-risk-damage-to-digital-economy>
- HOBBHAHN, M.; HEIM, L. e AVDOS, G. Trends in Machine Learning Hardware. *Epoch* 2023. <https://epochai.org/blog/trends-in-machine-learning-hardware>
- HOCHREITER, S. e SCHMIDHUBER, J. Long Short-Term Memory. *Neural Comput* 9 (8), 1997, pp. 1735–1780. doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- HOFFMAN, J. et al. Training Compute-Optimal Large Language Models. *ArXiv* 2022. <https://arxiv.org/abs/2203.15556>
- HOLGERSSON, M.; GRANSTRAND, O. Value capture in open innovation markets: the role of patent rights for innovation appropriation. *European Journal of Innovation Management* 25(6), 2022, pp. 320-339. <https://doi.org/10.1108/EJIM-02-2021-0114>
- HOLGERSSON, M.; GRANSTRAND, O. e BOGERS, M. The evolution of intellectual property strategy in innovation ecosystems: Uncovering complementary and substitute appropriability regimes. *Long Range Planning* 51 (2), 2018, pp. 303-319. <https://doi.org/10.1016/j.lrp.2017.08.007>
- HOUSE DIGITAL MARKETS REPORT. Investigation of Competition in Digital Markets. Report Subcommittee on Antitrust, Commercial and Administrative Law of the Committee of the Judiciary, US House of Representatives. 2020. https://democrats-judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf
- HRADEC, J. et al. Multipurpose synthetic population for policy applications. EUR 31116 EM. Publications Office of the European Union, Luxembourg. 2022. doi:10.2760/50072, JRC128595
- HU, K. e VARGHESE, H.M. Microsoft pays Inflection \$650 mln in licensing deal while poaching top talents, source says. *Reuters* March 21, 2024. <https://www.reuters.com/technology/microsoft-agreed-pay-inflection-650-mln-while-hiring-its-staff-information-2024-03-21/>
- HURLER, K. Amazon Warns Employees to Beware of ChatGPT. *Gizmodo*. Jan 16, 2023. <https://gizmodo.com/amazon-chatgpt-ai-software-job-coding-1850034383>. Acesso em: 02 fev.2023
- HURMELINNA-LAUKKANEN, P. e YANG, J. Distinguishing between appropriability and appropriation: A systematic review and a renewed conceptual framing. *Research Policy* 51 (1), 2022. <https://doi.org/10.1016/j.respol.2021.104417>
- ILIA. Índice Latinoamericano de Inteligencia Artificial. CNIA/CL. 2024 <https://indicelatam.cl/wp-content/uploads/2024/10/ILIA2024.pdf>
- INDIA DISCUSSION PAPER. Discussion Paper on Standard Essential Patents and their Availability on FRAND Terms. Government of India Department of Industrial Policy and Promotion Ministry of Commerce & Industry. March 1, 2016. https://www.ipindia.gov.in/writereaddata/Portal/News/196_1_standardEssentialPaper_01March2016_1_.pdf

- INFOMONEY Google aceitou pagar US\$360 mi à Activision para eliminar competição, diz Epic Games. Nov 17, 2022. <https://www.infomoney.com.br/negocios/google-aceitou-pagar-us360-mi-a-activision-para-eliminar-competicao-diz-epic-games/>
- INTERESSE, G. China to Regulate Deep Synthesis (Deepfake) Technology Starting 2023. China Briefing Dec 20, 2022. <https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/>. Acesso 03 mai. 2024
- IPA Venture team/IP Australia. Provocations: Generative AI and the IP Rights. April 2023. Postado Alex Roberts no LinkedIn. https://media.licdn.com/dms/document/media/D561FAQEmrvQca_eymQ/feedshare-document-pdf-analyzed/0/1684466014078?e=1696464000&v=beta&t=rcmq12Ety8uS-Og1iic7od3dyEqlxTw_SEAeSbGcN0. Acesso em 05 jul. 2023
- JACOBIDES, M. G. GenAI will change the world. But there are still things it can't do. *Forbes* Oct 2, 2023. <https://www.forbes.com/sites/lbsbusinessstrategyreview/2023/09/29/genai-will-change-the-world-but-there-are-still-things-it-cant-do/?sh=6f89a344f9b8>. Acesso em: 02 Out.2023
- JACOBIDES, M. G.; BRUSONI, S. e CANDELON, F. The Evolutionary Dynamics of the Artificial Intelligence Ecosystem. *Strategy Science* 6 (4), 2021, pp; 412-435. <https://doi.org/10.1287/stsc.2021.0148>
- JACOBIDES, M.G.; CENNAMO, C. e GAWER, A. Towards a theory of ecosystems. *Strategic Management Journal* 39 (8), 2018, pp. 2255-2276. <https://doi.org/10.1002/smj.2904>.
- JAIN, P.; GYANCHANDANI, M. e KHARE, N. BigData privacy: a technological perspective and review. *Journal of Big Data* 3 (25), 2016, pp. 3:25 <https://doi.org/10.1186/s40537-016-0059-y>
- JAOKAR, A. Artificial Intelligence #43: How do AI chips/neural chips work. LinkedIn 2022. <https://www.linkedin.com/pulse/artificial-intelligence-43-how-do-ai-chips-neural-work-ajit-jaokar>. Acesso em: 19 jul.2024
- JOHNSON, K. Microsoft acquire startup de IA conversacional Semantic Machines. *VentureBeat*. May 20, 2018. <https://venturebeat.com/ai/microsoft-acquires-conversational-ai-startup-semantic-machines/>
- JONES, C. I. e TONETTI, C. Nonrivalry and the Economics of Data. *NBER Working Paper No. 26260*. Cambridge, MA. April. 2020. DOI 10.3386/w26260
- KAK, A. WEST, S.M. e WHITTAKER, M. Make no mistake: AI is owned by BigTech. *MIT Technology Review* Dec. 5, 2023. <https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/>. Acesso em: 08 dez. 2023
- KAPLAN, A. e HAENLEIN, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons* 62 (1), 2019, pp. 15-25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- KARIMI, A-H. et al. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *ArXiv* 2020. <https://arxiv.org/abs/2006.06831>
- KERR, S. et al. Saudi-China collaboration raises concerns about access to AI chips. *Financial Times* Oct 9, 2023. <https://www.ft.com/content/2a636cee-b0d2-45c2-a815-11ca32371763>. Acesso em: 10 set.2024
- KESERIS, D.V. e KOVARIK, R. ChatGPT: IP Strategy in the AI Space. *Lexology*. 2023. <https://www.lexology.com/library/detail.aspx?g=bed7b5f6-5e7a-49c5-a8fc-accb59f95819>
- KHAITAN S. K. e MCCALLEY J. D. Design Techniques and Applications of Cyber Physical Systems: A Survey. *IEEE Systems Journal*, July, 2014. DOI: 10.1109/JSYST.2014.2322503
- KHAN, L.M. We Must Regulate A.I. Here's How. *The New York Times* May 3, 2023. <https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html>. 02 mai.2024
- KHAN, L.M. The New Brandeis Movement: America's Antimonopoly Debate, *Journal of European Competition Law & Practice* 9 (3), 2018, pp. 131-132. <https://doi.org/10.1093/jeclap/lpy020>
- KHAN, L.M. Amazon's Antitrust Paradox. *The Yale Law Journal* 126 (3), 2017, pp. 654-907. https://scholarship.law.columbia.edu/faculty_scholarship/2808
- KHAN, S.M. e MANN, A. AI Chips: What they are and why they matter. *Center for Security and Emerging Technology*. 2020. <https://cset.georgetown.edu/wp-content/uploads/AI-Chips%E2%80%94What-They-Are-and-Why-They-Matter.pdf>
- KHAN, S.M.; MANN, A. e PETERSON, D. The Semiconductor Supply Chain: Assessing National Competitiveness. *CSET Issue Brief* 2021. <https://cset.georgetown.edu/wp-content/uploads/The-Semiconductor-Supply-Chain-Issue-Brief.pdf>
- KHARPAL, A. A globally critical chip firm is driving a wedge between the U.S. and Netherlands over China tech policy. *CNBC* Dec 4, 2022. <https://www.cnbc.com/2022/12/05/us-ramps-up-pressure-on-netherlands-to-fall-in-line-with-china-chip-policy.html>. Acesso em: 20 abr. 2024
- KILAMO, T., HAMMOUDA, I., KAIRAMO, V., RÄSÄNEN, P. e SAARINEN, J.P. Open Source, Open Innovation and Intellectual Property Rights – A Lightning Talk. In: Hammouda, I., Lundell, B., Mikkonen, T., Scacchi, W. (eds) *Open Source Systems: Long-Term Sustainability*. OSS. 2012. *IFIP Advances in Information and Communication Technology*, vol 378. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33442-9_25

- KILKENNY, K. e CHO, W. Attack of the Chatbots: Screenwriters' Friends or Foe? *The Hollywood Reporter* Jan 12, 2023. <https://www.hollywoodreporter.com/business/business-news/chatgpt-hollywood-screenwriters-film-tv-1235296724/>. Acesso em: 14 fev. 2023.
- KIM, Y. Convolutional Neural Networks for Sentence Classification. *ArXiv* 2014. <https://arxiv.org/abs/1408.5882>
- KIM, D. 'AI-Generated Inventions': Time to Get the Record Straight?, *GRUR International* 69 (5), 2020, pp. 443–456. <https://doi.org/10.1093/grurint/ikaa061>
- KIM, D.; DREXL, J.; HILTY, R. M. e SLOWISKI, P.R. Artificial Intelligence Systems as Inventors? A Position Statement of 7 September 2021 in view of the evolving case-law worldwide. Max Planck Institute for Innovation and Competition https://www.ip.mpg.de/fileadmin/ipmpg/content/stellungnahmen/MPI_Position_statement_AI_Inventor_2021-08-09.pdf
- KINGMA, D.P. e WELLING, M. Auto-Encoding Variational Bayes. *ArXiv* 2013. <https://arxiv.org/abs/1312.6114>
- KLINGE, T. J. et al. Augmenting digital monopolies: A corporate financialization perspective on the rise of Big Tech. *Competition & Change* 27(2), 2023, pp. 332-353. <https://doi.org/10.1177/10245294221105573>
- KLINGER J.; MATEOS-GARCIA, J. e STATHOULOPOULOS, K. Deep learning, deep change? Mapping the development of the Artificial Intelligence General Purpose Technology. Cornell University. 2018
- KLOBUCHAR, A. Klobuchar Reintroduces Bill to Promote Competition and Improve Antitrust Enforcement. 2024. <https://www.klobuchar.senate.gov/public/index.cfm/2024/5/klobuchar-reintroduces-bill-to-promote-competition-and-improve-antitrust-enforcement>
- KNECHTEL, J.; PATNAIK, S. e SINANOGLU, O. Protect Your Chip Design Intellectual Property: An Overview. *COINS '19: Proceedings of the International Conference on Omni-Layer Intelligent Systems*, 2019, pp. 211 – 216. <https://doi.org/10.1145/3312614.3312657>
- KORINEK, A. e STIGLITZ, J. Artificial Intelligence, Globalization, and Strategies for Economic Development. *Working Paper* 28453. NBER. 2021. <http://www.nber.org/papers/w28453>
- KOSTADINOVA, E. Intellectual Property for the EU in a World of Challenges, Intellectual Property Unit, DG Grow, European Commission, Zagreb, February, 2020. <https://www.dziv.hr/files/file/eu2020hr/d1/big2.pdf>
- KOUTROUMPIS, P.; LEIPONEN, A. e THOMAS, L. D. W. Markets for data. *Industrial and Corporate Change* 29 (3), 2020, pp. 645–660. March. <https://doi.org/10.1093/icc/dtaa002>
- KRAEMER, J. Personal Data Portability in The Platform Economy: Economic Implications and Policy Recommendations. *Journal of Competition Law & Economics*. 2020 <https://ssrn.com/abstract=3742771>
- KRASIN, I. e DUERIG, T. Introducing the Open Images Dataset. *Google Research*. Sept. 30, 2016. <https://research.google/blog/introducing-the-open-images-dataset/>. Acesso em: 02 out. 2022
- KRIZHEVSKY, A.; SUTSKEVER, I. e HINTON, G. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS 2012 Papers*. <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>. Acesso em: 02 jul. 2021
- KÜSPERT, S., MOËS, N. e DUNLOP, C. The value chain of general-purpose AI: A closer look at the implications of API and open-source accessible GPAI for the EU AI Act. *Adalovelace Inst*, 2023 <https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/#fn-1>. Acesso em: 16 mar. 2023
- LANGSTON, J. Microsoft announces new supercomputer, lays out vision for future AI work. *Microsoft* May 19, 2020. <https://news.microsoft.com/source/features/innovation/openai-azure-supercomputer/>. Acesso em: 11 mar. 2021
- LANGUS, G. et al. Standard Essential Patents: Who is Really Holding Up (and when)? *WIPO* Feb, 2013. https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_econ_ge_2_13/wipo_ip_econ_ge_2_13_neven.pdf Acesso 20 out. 2022
- LARDINOIS, F. e LUNDEN, I. Microsoft has acquired GitHub for \$7.5B in stock. *TechCrunch* June 4, 2018. <https://techcrunch.com/2018/06/04/microsoft-has-acquired-github-for-7-5b-in-microsoft-stock/>
- LAURSEN, K. E SALTER, A. J. The paradox of openness: Appropriability, external search and collaboration. *Research Policy* 43(5), 2014, pp. 867-878 <https://doi.org/10.1016/j.respol.2013.10.004>.
- LAURSEN, K. e SALTER, A. J. The Role of Appropriability Strategies in Shaping Innovative Performance. *DRUID Working Paper* N. 05-02. 2005. <https://cdi.mecon.gob.ar/bases/doc/druid/2005-2.pdf>
- LE SCAO, T. et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv* 2023. <https://arxiv.org/abs/2211.05100>
- LECUN, Y.; BENGIO, Y. e HINTON, G. Deep learning. *Nature* 521, 2015, pp. 436-444. Macmillan Publishers Limited. <https://doi.org/10.1038/nature14539>.
- LECUN, Y. e BENGIO, Y. Convolutional Networks for Images Speech and Time Series. 1998 <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e26cc4a1c717653f323715d751c8dea7461aa105>. Acesso 12 ago.2024
- LEE, K.F. Project AI 2.0. 2023. <https://www.chuangxin.com/ai2>. Acesso 20 jun.2023
- LEE, K.F. e QUIUFAN, C. AI 2041: Ten Visions for Our Future. September. Currency Ed. 2021.

- LEE, K. et al. Hallucinations in Neural Machine Translation. Interpretability and Robustness in Audio, Speech, and Language Workshop Conference on Neural Information Processing Systems (*NeurIPS 2018*) Montreal, Canada. https://clarafy.github.io/research/neurips_irasl_2018.pdf. Acesso em: 25 ago.2024
- LEFKOWITZ, M. Professor's perceptron paved the way for AI – 60 years too soon. *Cornell Chronicle* Sept 25, 2019. <https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon>. Acesso: 20 jun. 2021
- LEI, Z.; SUN Z. e WRIGHT, B. Patent subsidy and patent filing in China. 2012 https://funginstitute.berkeley.edu/wp-content/uploads/2013/12/patent_subsidy_Zhen.pdf
- LEMLEY, M. A. Ten Things to Do About Patent Holdup of Standards (And One Not to), *Boston College Law Review*, 48, 2007. <https://ssrn.com/abstract=923470>
- LEMLEY, M. A. Intellectual Property Rights and Standard-Setting Organizations, *California Law Review* 90, 2002, pp. 1896-97. <https://dx.doi.org/10.2139/ssrn.310122>
- LEMLEY, M. A. e CASEY, B. Fair Learning. SSRN. Jan 30, 2020. <http://dx.doi.org/10.2139/ssrn.3528447>. Acesso em: 08 jun. 2023
- LEMLEY, M. A. e SIMCOE, T. How Essential Are Standard-Essential Patents? *Cornell Law Review* 104 (3), 2019, pp. 607-642 <https://scholarship.law.cornell.edu/clr/vol104/iss3/2>.
- LEMLEY, M. A. e SHAPIRO, C. A Simple Approach to Setting Reasonable Royalties for Standard-Essential Patents, *Berkeley Technology Law Journal* 28 (1135), 2013, pp. 1164-66.
- LEMLEY, M. A. e SHAPIRO, C. Patent Holdup and Royalty Stacking. *Texas Law Review* 85, 2007, pp. 1991-2049. <https://ssrn.com/abstract=923468>
- LEMLEY, M.A. e SHAPIRO, K. Probabilistic Patents. *Journal of Economic Perspectives*, 19 (2), 2005, pp. 75–98. DOI: 10.1257/0895330054048650
- LÉVÊQUE, F. e MÉNIÈRE, Y. The Economics of Patents and Copyright. *The Berkeley Electronic Press*. 2004. http://memoires.scd.univ-tours.fr/Droit/Ebooks/The_economics_of_patents_and_copyright.pdf
- LEVIN, R. C.; KLEVORICK, A.K.; NELSON, R.R. e WINTER, S.G. Appropriating the Returns from Industrial Research and Development. *Brookings Papers on Economic Activity* 3, 1987, pp.783-831
- LEWKOWYCZ, A. et al. Solving Quantitative Reasoning Problems with Language Models. *ArXiv* 2022. <https://arxiv.org/abs/2206.14858>
- LI, W.Y. Regulatory capture's third face of power, *Socio-Economic Review* 21 (2), 2023, pp. 1217–1245. <https://doi.org/10.1093/ser/mwad002>
- LI, Y. et al. Competition-level code generation with AlphaCode. *Science* Dec. 8, 2022. <https://www.science.org/doi/10.1126/science.abq1158>. Acesso em: 10 dez. 2022
- LI, P. et al. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *ArXiv* 2023. <https://arxiv.org/abs/2304.03271>
- LI, F.F. Welcome Kaggle to Google Cloud. *Google Cloud*. March 8, 2017. <https://cloud.google.com/blog/products/gcp/welcome-kaggle-to-google-cloud/>
- LI, X. e WU, X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, pp. 4520-4524. doi: 10.1109/ICASSP.2015.7178826.
- LIANOS, I. e DREYFUSS, R.C. New Challenges in the Intersection of Intellectual Property Rights with Competition Law - A View from Europe and the United States. April, 2013. *CLES Working Paper Series 4/2013*. <https://discovery.ucl.ac.uk/id/eprint/10045063/>
- LIN, B. In Race for AI Chips, Google DeepMind Uses AI to Design Specialized Semiconductors. *The Wall Street Journal* July 20, 2023. <https://www.wsj.com/articles/in-race-for-ai-chips-google-deepmind-uses-ai-to-design-specialized-semiconductors-dcd78967>. Acesso em: 11 ago. 2023
- LIN, J. et al. M6-10T: A Sharing-Delinking Paradigm for Efficient Multi-Trillion Parameter Pretraining. *ArXiv* 2021. <https://arxiv.org/abs/2110.03888>
- LIN, S.; HILTON, J. e EVANS, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *ArXiv* 2022. <https://arxiv.org/abs/2109.07958>
- LOHR, S. 5 lessons from Microsoft's antitrust woes, by people who lived it. *New York Times* June 23, 2019. <https://www.nytimes.com/2019/06/23/technology/antitrust-tech-microsoft-lessons.html>.
- LOPES, A. Microsoft mira expansão de IA e investe US\$ 3,2 bilhões em infraestrutura de nuvem na Suécia. *Exame* 3 junho, 2024. <https://exame.com/inteligencia-artificial/microsoft-mira-expansao-de-ia-e-investe-us-32-bilhoes-em-infraestrutura-de-nuvem-na-suecia/>. Acesso em: 03 jun. 2024
- LORENZ, P.; PERSET, K. e BERRYHILL, J. Initial policy considerations for generative artificial intelligence", *OECD Artificial Intelligence Papers*, No. 1, OECD Publishing, Paris. 2023 <https://doi.org/10.1787/fae2d1e6-en>
- LUCAS JR, H.C. E GOH, J.M. Disruptive technology: How Kodak missed the digital photography revolution. *The Journal of Strategic Information Systems* 18 (1), 2009, pp. 46-55. <https://doi.org/10.1016/j.jsis.2009.01.002>.

- LUDLOW, E. e VANCE, A. e BLOOMBERG. Sam Altman was fundraising in the Middle East for a new chip venture to rival Nvidia before OpenAI's board ousted him. *Fortune* Nov 19, 2023. <https://fortune.com/2023/11/19/sam-altman-was-fundraising-for-chip-venture-to-rival-Nvidia-before-openai-ouster/>. Acesso em: 25 abr.2024
- LUITSE, D. e DENKENA, W. The great Transformer: Examining the role of large language models in the political economy of AI. *Big Data and Society* 1-14, 2021. DOI: 10.1177/20539517211047734
- LUO, Y., DAN, X.e SHEPHERD, N. China Proposes Draft Measures to Regulate Generative AI. *Convington*. April 12, 2023. <https://www.insideprivacy.com/artificial-intelligence/china-proposes-draft-measures-to-regulate-generative-ai/>. Acesso 03 mai. 2023
- LYU, Y.G. (2020) Artificial Intelligence: Enabling Technology to Empower Society. *Engineering* 6 (3), 2020, pp. 205-206. <https://doi.org/10.1016/j.eng.2020.01.005>
- MA, W. China's approach to data and AI is changing. Here's what that means. *World Economic Forum* Jan 11, 2024. <https://www.weforum.org/agenda/2024/01/chinas-data-and-ai-approach-is-changing-heres-what-that-means/>. Acesso em: 21 mar. 2024
- MACEDO M. F. e BARBOSA, A. L. Patentes, Pesquisa & Desenvolvimento: um manual de propriedade intelectual. Rio de Janeiro, Editora Fiocruz, 2000.
- MCCULLOCH, W. e PITTS, P. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology* 5, 1943, pp.115–133. <https://www.cs.cmu.edu/~/epxing/Class/10715/reading/McCulloch.and.Pitts.pdf>
- MACHLUP, F. e PENROSE, E. The Patent Controversy in the Nineteenth Century. *The Journal of Economic History*. New York, *Economic History Association* 10 (1), 1950, pp.1-29
- MACHLUP, F. An Economic Review of the Patent System. Study N° 15, Study of the Subcommittee on Patents, Trademarks, and Copyright of the Committee on the Judiciary, US Senate. US Government Printing Office. Washington. 1958
- MADIEGA, T. Digital sovereignty for Europe. European Parliamentary Research Service. PE 651.992. *EU Briefing*. July. 2022 [https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651992/EPRS_BRI\(2020\)651992_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651992/EPRS_BRI(2020)651992_EN.pdf). Acesso em: 15 set.2021
- MADISON, A. National Science Foundation rolls out NAIRR pilot with industry, agency support. *Feedscoop* Jan 24, 2024. <https://fedscoop.com/nsf-launches-nairr-pilot/>
- MAFFULLI, S. Meta's LLaMa 2 license is not Open Source. *Open Source Initiative* July 20, 2023. <https://opensource.org/blog/metals-llama-2-license-is-not-open-source>. Acesso em: 07 ago.2024
- MANGELSON, M. et al. Trademarks and patents in China: The impact of non-market factors on filing trends and IP systems. *USPTO*. January, 2021. <https://www.uspto.gov/sites/default/files/documents/USPTO-TrademarkPatentsInChina.pdf>
- MANSFIELD, E. R&D and Innovation: Some Empirical Findings in Griliches, V. (ed.). *R & D, Patents, and Productivity*. 1984. <http://www.nber.org/chapters/c10047>
- MANSFIELD, E.; SCHWARTZ, M. e WAGNER, S. Imitation costs and patents: an empirical study. *The Economic Journal* 91 (364), 1981, pp. 907-918. <https://doi.org/10.2307/2232499>
- MANSFIELD, E.; RAPOPORT, J.; ROMEO, A.; WAGNER, S. e BEARDSLEY, G. Social and Private Rates of Return from Industrial Innovations, *The Quarterly Journal of Economics* 91 (2), 1977, pp 221-240. <https://doi.org/10.2307/1885415>
- MARCUS, G. Deep Learning Is Hitting a Wall. *Nautilus*. March 10, 2022. <https://nautil.us/deep-learning-is-hitting-a-wall-238440/>. Acesso em: 03 jan. 2023
- MARGONI, T. e KRETSCHMER, M. A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, *GRUR International* 71 (8), 2022 pp. 685–701, <https://doi.org/10.1093/grurint/ikac054>
- MARSHALL, A. My Strange Day With Bing's New AI Chatbot. *Wired*. Feb 9, 2023. <https://www.wired.com/story/my-strange-day-with-bings-new-ai-chatbot/>. Acesso em: 18 feb. 2023
- MARTINEZ, G. et al. Towards understanding the interplay of generative artificial intelligence and the Internet. *Arxiv* 2023. <https://arxiv.org/abs/2306.06130>
- MASLEJ, N. et al. The AI Index 2023 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April, 2023. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf Acesso em: 30 mai. 2023
- MASLEJ, N. et al. The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April, 2024. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf. Acesso em: 22 mai. 2024
- MAZZUCATO, M. *Economy: A moonshot guide to changing capitalism*. Harper Business. 2021. ISBN-13: 978-0063046238
- MAZZUCATO, M. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. *Public Affairs*. 2013. ISBN-13: 978-0857282521

- McCARTHY, J., MINSKY, M.L., ROCHESTER, N. e SHANNON, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine* 27 (4), August 31 (2006 [1955]). Disponível em: <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>. Acesso em: 01 jul. 2020
- McCARTHY, J. *Some Expert Systems Need Common Sense*, *Computer Culture: The Scientific, Intellectual and Social Impact of the Computer*, Heinz Pagels, ed. vol. 426, Annals of the New York Academy of Sciences. 1984. <http://jmc.stanford.edu/articles/someneed.html>
- McCABE, D. e SWANSON, A. U.S. Using Trade Deals to Shield Tech Giants From Foreign Regulators. *The New York Times*. Oct 7, 2019. <https://www.nytimes.com/2019/10/07/business/tech-shield-trade-deals.html>. Acesso em: 29 mai. 2024
- McDERMOTT, E. e BRACHMANN, S. IP Organizations Want More on USPTO's AI Patent Eligibility Guidance. *Ipwatchdog* Oct 20, 2024. Disponível em: <https://ipwatchdog.com/2024/10/20/ip-organizations-want-usptos-ai-patent-eligibility-guidance/id=182378/Acesso> em: 25 out. 2024
- ELINGRUD, K. et al. Generative AI and the future of work in America. *MGI* July 26, 2023. <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america>. Acesso em: 10 out.2023
- McKINZIE, B. et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *Arxiv* 2024. <https://arxiv.org/pdf/2403.09611>
- McLAUGHLIN, K. E GARDIZY, A. After Years of Resistance, AWS Opens Checkbook for Open-Source Providers. By Kevin McLaughlin and Anissa Gardizy. May 24, 2023. <https://www.theinformation.com/articles/after-years-of-resistance-aws-opens-checkbook-for-open-source-providers>. Acesso em: 30 mai.2023
- McMAHANM, B. e RAMAGE, D. Federated Learning: Collaborative Machine Learning without Centralized Training Data. *Google Research* April 6, 2017. <https://research.google/blog/federated-learning-collaborative-machine-learning-without-centralized-training-data/>. Acesso em: 12 jan. 2022
- MEDEIROS, H.G. e WACHOWICZ, M. A sobreposição de direitos de propriedade intelectual no software *Revista Jurídica Luso-Brasileira* 5 (5), 2019, pp. 953-986. https://www.cidp.pt/revistas/rjlb/2019/4/2019_04_0953_0986.pdf
- MEDEIROS, H.G. A sobreposição de direitos de propriedade intelectual no software: Coexistência entre direito de autor e patente na sociedade informacional. Tese de doutorado UFSC/CCJ. Florianópolis. 2017 <https://repositorio.ufsc.br/bitstream/handle/123456789/178710/347865.pdf?sequence=1&isAllowed=y>
- MENDONÇA, S. et al. The Rise of 5G Technologies and Systems: a Quantitative Analysis of Knowledge Production, *Telecommunications policy* 46 (4), 2022. <https://doi.org/10.1016/j.telpol.2022.102327>
- MESKE, C. et al. Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management* 39 (1), 2022, pp. 53-63. <https://doi.org/10.1080/10580530.2020.1849465>
- MEHDI, Y. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. *Microsoft Official Blog*. Feb.7, 2023. <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>. Acesso em: 21 jul. 2024
- MEHTA, I. Character.AI CEO Noam Shazeer returns to Google. *TechCrunch*. Aug 2, 2024. <https://techcrunch.com/2024/08/02/character-ai-ceo-noam-shazeer-returns-to-google/>. Acesso em: 02.ago.2024
- MELL, P. e GRANCE, T. The NIST Definition of Cloud Computing . National Institute of Standards and Technology. US Department of Commerce. Special Publication 800-145. September, 2011. <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf>. Acesso em: 03 mar.2023
- METZ, C. Google Just Open Sourced TensorFlow, Its Artificial Intelligence Engine. *Wired* Nov 9, 2015. <https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/>. Acesso em: 2 mar.2023
- MICROSOFT. Microsoft and OpenAI extend partnership. *Official Microsoft Blog*. Jan 23, 2023. <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/>. Acesso 02 mar. 2023
- MINARI, G. Meta anuncia supercomputador de IA mais rápido do mundo. *CanalTech* Jan. 25, 2022. <https://canaltech.com.br/inteligencia-artificial/meta-anuncia-supercomputador-de-ia-mais-rapido-do-mundo-207457/>. Acesso em: 05 fev. 2022
- MOORE, S.K. Nvidia's Next GPU Shows That Transformers Are Transforming AI. *IEEE Spectrum* April 8, 2022. <https://spectrum.ieee.org/Nvidias-next-gpu-shows-that-transformers-are-transforming-ai>. Acesso em: 20 ago.2024
- MOORE, J.F. *The Death of Competition: Leadership and Strategy in the Age of Business Ecosystems*. Harper Business. 1996
- MORENO, J. 70% Of Generative AI Startups Rely On Google Cloud, AI Capabilities. *Forbes* July 25, 2023. <https://www.forbes.com/sites/johannmoreno/2023/07/25/70-of-generative-ai-startups-rely-on-google-cloud-ai-capabilities-says-alphabet-ceo-sundar-pichai/>. Acesso em: 12 jul.2024
- MOYER, M. Yes, Government Researchers Really Did Invent the Internet. July 23. *Scientific American*. 2012. <https://blogs.scientificamerican.com/observations/yes-government-researchers-really-did-invent-the-internet/>

- MUVIJA, M.; MEHTA, C. e SONI, A. Microsoft, OpenAI tie-up comes under antitrust scrutiny. *Reuters*. Dec 8, 2023. <https://www.reuters.com/world/uk/uk-antitrust-regulator-considering-microsoft-openai-partnership-2023-12-08/>. Acesso em: 08 dez.2023
- MUSSER M. et al.. The Main Resource is the Human'. *Center for Security and Emerging Technology*. April., 2023. <https://doi.org/10.51593/20210071>
- NARANG, S. e CHOWDHERRY, A. Pathways Language Model (PaLM): Scaling to 540 billion parameters for breakthrough performance. *Google Research*. April 4, 2022. <https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance/>. Acesso em: 16 jun.2023
- NARDIN, A. Artificial Intelligence as a General Purpose Technology: An exploratory analysis of PCT patents. Università Ca'Foscari. Venezia. 2021. <http://dspace.unive.it/handle/10579/19003>
- NATIONAL RESEARCH COUNCIL. Developments in Artificial Intelligence in Funding a Revolution: Government Support for Computing Research (Chapter 9). Washington, DC: The National Academies Press. Pp. 198-225. 1999. <https://doi.org/10.17226/6323>.
- NELSON, R.R. et al. Modern Evolutionary Economics, Cambridge University Press. 2018
- NELSON, R. R. The market economy, and the scientific commons, LEM Working Paper Series, N. 2003/24, Scuola Superiore Sant'Anna, Laboratory of Economics and Management (LEM), Pisa
- NELSON, R.R. Why do Firms Differ and How Does It Matters? A Revisitation, *Strategic Management Journal*, v, 12, Special Issue S2, 1991, pp. 61-74.
- NELSON, R.R. (1959) The Simple Economics of Basic Scientific Research. *The Journal of Political Economy* 67 (3), 1959, pp. 297-306. http://www.mba.intercol.edu/Entrepreneurship/UT%20Computer%20Science%20Course/Simple_Economics_Basic_Research.pdf
- NELSON, R. R. e WINTER, S. An evolutionary theory of economic change. The Belknap Press of Harvard University Press. 1982
- NILSSON, N.J. The quest for artificial intelligence a history of ideas and achievements for artificial intelligence a history of ideas and achievements. Cambridge University Press. 2009. <http://www.cambridge.org/us/0521122937>
- NIX, N. ; ZAKRZEWSKI, C. e DE VYNCK, G. Silicon Valley is pricing academics out of AI research. *Washington Post*. March 10, 2024. <https://www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research/>
- NORVIG, P. e RUSSEL, S. Artificial Intelligence: A Modern Approach. Pearson Education Limited. (2021 [2010]). ISBN-13: 978-1292401133
- NSCAI. National Security Commission on Artificial Intelligence. Final Report. 2021. <https://www.nsc.ai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf>. Acesso em: 20 ago. 2022
- NVIDIA. Nvidia. March 18, 2024. <https://Nvidianews.Nvidia.com/news/Nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing>. Acesso 17 abr. 2024
- OECD. A Blueprint for building compute capacity for Artificial Intelligence. OECD Digital Economy Papers N°350. Feb 2023. <https://www.oecd-ilibrary.org/docserver/876367e3-en.pdf?expires=1727928044&id=id&accname=guest&checksum=ACE6E12EFB1CEF08B449BC68AA864A8D>. Acesso em: 23 jun. 2023
- OECD. Artificial Intelligence in Society. OECD Publishing. Paris. 2019
- OPENAI. Video generation models as world simulators. Feb 15, 2024. <https://openai.com/index/video-generation-models-as-world-simulators/>. Acesso em: 02 mar.2024 (2024a)
- OPENAI. OpenAI acquires Rockset. June 21, 2024. <https://openai.com/index/openai-acquires-rockset/>. Acesso em: 02 ago.2024 (2024b)
- OPENAI. SearchGPT Prototype. July 25, 2024. <https://openai.com/index/searchgpt-prototype/>. Acesso em: 02 ago.2024 (2024c)
- OPENAI. GPT-4. March 14, 2023. <https://openai.com/index/gpt-4-research/>. Acesso em: 20 mar.2023
- ORATZ, L.T. et al. First Lawsuits Arrive Addressing Generative AI. *Perkins Coie* April 23, 2023. <https://perkinscoie.com/insights/update/first-lawsuits-arrive-addressing-generative-ai>. Acesso em: 10 out. 2023.
- O'REILLY, T. Rising tide rents and robber baron rents: How innovators lose their edge and their ideals. UCL Institute for Innovation and Public Purpose, *Working Paper Series (IIPP WP 2024-04)*. Disponível em: <https://www.ucl.ac.uk/bartlett/public-purpose/wp2024-04>
- OREMUS, W. Big Tobacco. Big Pharma. Big Tech? *Slate*. Nov 17, 2017. <https://slate.com/technology/2017/11/how-silicon-valley-became-big-tech.html>. Acesso em: 02 mai. 2020
- OSAWA, J. e HOLMES, A. Google, Microsoft Help Chinese Firms Skirt Ban on Nvidia Chips. *The Information*. 2024 https://www.theinformation.com/articles/google-microsoft-help-chinese-firms-skirt-ban-on-nvidia-chips?offer=rtsu-engagement-24&utm_campaign=RTSU%3A+Google+Microso&utm_content=4695&utm_medium=email&utm_source=cio&utm_term=3172
- OSTROM, E. e HESS, C. Private and Common Property Rights. Encyclopedia of Law & Economics. Northampton, MA: Edward Elgar. 2007. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1304699

- OUYANG, L. et al. Training language models to follow instructions with human feedback. *Arxiv* 2022 <https://arxiv.org/abs/2203.02155>
- OVIDE, S. We tested Amazon's new chatbot. It's not good. *The Washington Post* March 5, 2024. <https://www.washingtonpost.com/technology/2024/03/05/amazon-ai-chatbot-rufus-review/>. Acesso em: 05 mar.2024
- OWENS, J.D. et al. GPU Computing in *Proceedings of the IEEE* 96 (5), 2008, pp. 879-899. doi: 10.1109/JPROC.2008.917757.
- PALMA, R. et al. The Growing Challenge of Semiconductor Design Leadership. *BCG/SIA*. Nov, 2022. <https://web-assets.bcg.com/3f/b4/fd384ccd46dc8a381bd61a648105/bcg-the-growing-challenge-of-semiconductor-design-leadership-nov-2022-r.pdf>. Acesso em: 02 jul.2024
- PAN, X. et al. Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models. *Arxiv* 2022. <https://arxiv.org/abs/2211.10950>
- PAREDES-FRIGOLETT, H. e PYKA, A. The global stakeholder capitalism model of digital platforms and its implications for strategy and innovation from a Schumpeterian perspective. *Journal of Evolutionary Economics* 32, 2022, pp. 463–500. <https://doi.org/10.1007/s00191-022-00760-z>
- PARK, K. Google, Apple face fines in South Korea for breaching in-app billing rules. *TechCrunch* Oct 6, 2023. <https://techcrunch.com/2023/10/06/google-apple-face-fines-in-south-korea-for-breaching-in-app-billing-rules>
- PARKER, G.G.; VAN ALSTYNE, M.W. e CHOUDARY, S.P. (2016) Platform Revolution: How Networked Markets Are Transforming the Economy—and How to Make Them Work for. W. W. Norton & Company
- PATEL, D. e NISHBALL, D. Google Gemini eats the world. *Semianalysis* Ago, 2023. <https://www.semianalysis.com/p/google-gemini-eats-the-world-gemini>. Acesso em: 19 set.2023
- PATEL, D. e NISHBALL, D. OpenAI is doomed et tu Microsoft. *Semianalysis* May 7, 2024. <https://www.semianalysis.com/p/openai-is-doomed-et-tu-microsoft>. Acesso em: 10 mai. 2024 (2024a)
- PATEL, D. e NISHBALL, D. 100,000 H100 Custers: Power, Network, Tipology Ethernet vs. Infiliband, Failures, CheckPointing. *Semianalysis* Jun 17, 2024. <https://www.semianalysis.com/p/100000-h100-clusters-power-network>. Acesso em: 28 jul. 2024 (2024b)
- PATEL, D. e AHMAD, A. Google “We have no moat, and neither does OpenAI”. *Semianalysis*. May 4, 2023. <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>. Acesso em: 20 mai.2023
- PATEL, D. COZMA, J. e WONG, G. Google AI Infrastructure Supremacy: Systems matters more than microstructures. *Semianalysis* April 12, 2023. <https://www.semianalysis.com/p/google-ai-infrastructure-supremacy>. Acesso em: 20 set. 2023
- PATEL, D. e KOSTOVIC, A. AMD AI software solved – MI300X Pricing, Performance. *Semianalysis* June 30, 2023. <https://www.semianalysis.com/p/amd-ai-software-solved-mi300x-pricing>. Acesso em: 20 set. 2023
- PATEL, D. e XIE, M. Apple's AI strategy: Apple datacenters, on-device, cloud and more. *Semianalysis* May 17, 2024. <https://www.semianalysis.com/p/apples-ai-strategy-apple-datacenters>
- PATEL, D. Accelerator Industry Model. *Semianalysis* Feb 1, 2024. <https://www.semianalysis.com/p/accelerator-model>
- PATEL, D. Amazon's cloud crisis: how AWS will loose the future of computing. *Semianalysis* Mar. 23, 2023. <https://www.semianalysis.com/p/amazons-cloud-crisis-how-aws-will-loose-the-future-of-computing>
- PENG, T. Meet 100 Chinese LLMs. *Recode China AI* July 17, 2023. <https://recodechinaai.substack.com/p/friendlier-chinas-regulation-on-generative>. Acesso em: 12 ago.2023
- PÉNIN, J. e NEICU, D. Patents and Open Innovation: Bad Fences Do Not Make Good Neighbors. *Journal of Innovation Economics & Management* 1 (25), 2018, pp. 57-85. <https://shs.cairn.info/revue-journal-of-innovation-economics-2018-1-page-57?lang=fr>
- PENN ENGINEERING TODAY. A chip that can classify nearly two billion images per second. *Penn Today* June 6, 2022. <https://penntoday.upenn.edu/news/penn-engineers-chip-can-classify-nearly-two-billion-images-second>. Acesso em: 10 nov.2023
- PENROSE, E. The Theory of The Growth of the Firm, Oxford University Press. 1959
- PENROSE, E. The Economics of the International Patent System. Johns Hopkins Press. 1951
- PETERS, J. Apple and Nvidia may invest in OpenAI. *The Verge*, Aug 29, 2024. <https://www.theverge.com/2024/8/29/24231626/apple-Nvidia-openai-invest-microsoft>. Acesso em: 29 ago. 2024
- PETIT, N. e TECCE, D.J. Taking Ecosystems Competition Seriously in the Digital Economy. OECD. 134th meeting of the Competition Committee on 1-3 December, 2020. [https://one.oecd.org/document/DAF/COMP/WD\(2020\)90/en/pdf](https://one.oecd.org/document/DAF/COMP/WD(2020)90/en/pdf). Acesso 04/12/2020.
- PEREZ, C. Technological revolutions and techno-e economic paradigms. *Technology Governance. Working Paper in Technology Governance and Economic Dynamics* n°20, 2009. <http://hum.ttu.ee/wp/paper20.pdf>.
- PEREZ, C. Structural Change and Assimilation of New Technologies in the Economic and Social Systems. 1983 https://carlotaperez.org/wp-content/downloads/publications/theoretical-framework/scass_v04.pdf
- PEREZ, S. Zuckerberg *TechCrunch*. Jun 27, 2024. <https://techcrunch.com/2024/06/27/zuckerberg-disses-closed->

- source-ai-competitors-as-trying-to-create-god/. Acesso em: 02 ago. 2024
- PIKE, D. United States - Export control rules on semiconductors and advanced ICs updated and expanded. *BDO*. January 9, 2024. <https://www.bdo.global/en-gb/insights/tax/indirect-tax/united-states-export-control-rules-on-semiconductors-and-advanced-ics-updated-and-expanded>. Acesso em : 02 fev. 2024
- PINTO, K. R. V. F. Integração entre propriedade intelectual e defesa da concorrência: o licenciamento de patentes no Brasil. Tese de doutorado. Junho. 2009. Instituto de Economia/UFRJ
- PISANO, G.P. Toward a prescriptive theory of dynamic capabilities: connecting strategic choice, learning, and competition. *Industrial and Corporate Change*. 26(5), 2017, pp. 747–762. <https://doi.org/10.1093/icc/dtx026>
- PISANO, G.P. Profiting from Innovation and Intellectual Property Revolution. *Research Policy* 35 (8), 2006, pp. 1122–1130. <https://doi.org/10.1016/j.respol.2006.09.008> (2006a)
- PISANO, G.P. Can Science Be a Business?: Lessons from Biotech. *HBR*. October, 2006. <https://hbr.org/2006/10/can-science-be-a-business-lessons-from-biotech> (2006b)
- PISANO, G.P. e TEECE, D.J. How to Capture Value from Innovation: Shaping Intellectual Property and Industry Architecture. *California Management Review* 50(1), 2007, pp. 278–296. <https://doi.org/10.2307/41166428>
- PLANTIN, J.-C. et al. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society* 20(1), 2018, pp. 293–310. <https://doi.org/10.1177/1461444816661553>
- POSNER, R.A. *Frontiers of Legal Theory*. Harvard University Press. 2001
- PWONLYIAS Regulation On Big Tech Companies in India: Recommendations by Committee on Digital Competition Law. 2023. <https://pwnonlyias.com/current-affairs/regulation-of-big-tech-firms/>. Acesso em: 05 ago. 2024
- PUJOL, N. Freemium: Attributes of an Emerging Business Model. SSRN Dec 1, 2010. <https://ssrn.com/abstract=1718663>
- POUYANFAR, S. et al. A Survey on DeepLearning: Algorithms, Techniques, and Applications. *ACM Computing Surveys* 51(5), 2018. <https://courses.cs.duke.edu/spring20/compsci527/papers/Pouyanfar.pdf>
- POWELL, A.J.S. et al. BIS Strengthens Export Controls on Advanced Computing Items, Semiconductor Manufacturing Equipment, and Supercomputing End-Uses to China and Other Countries of Concern. *Backer McKenzie* Oct 25, 2023. <https://sanctionsnews.bakermckenzie.com/bis-strengthens-export-controls-on-advanced-computing-items-semiconductor-manufacturing-equipment-and-supercomputing-end-uses-to-china-and-other-countries-of-concern/>
- PUNDZIENE, A., GUTMANN, T., SCHLICHTNER, M. e TEECE, D. J. Value Impedance and Dynamic Capabilities: The Case of MedTech Incumbent-Born Digital Healthcare Platforms. *California Management Review*, 64(4), 2022, pp. 108–134. <https://doi.org/10.1177/00081256221099326>
- RADFORD, A. et al. Improving Language Understanding by Generative Pre-Training. *OpenAI*. June 11, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: 12 mar. 2024
- RADFORD, A. et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *ArXiv* 2015. <https://arxiv.org/abs/1511.06434>
- RAHMAN, K.S. e THELEN, K. *Politics and Society* 47 (2), 2019, pp. 177–204. <https://doi.org/10.1177/0032329219838932>. Acesso: 12 ago. 2024
- RAHMAN, R.; OWEN, D. e YOU, J. (2024) Tracking Large-Scale AI Models". *Published online at epochai.org*. <https://epochai.org/blog/tracking-large-scale-ai-models>
- RAJBAHADUR, G.K. et al. Can I use this publicly available dataset to build commercial AI software? – A Case Study on Publicly Available Image Datasets. *ArXiv* 2022. <https://arxiv.org/abs/2111.02374>. Acesso em: 06 jun. 2023.
- RAMOS, G. e MAZZUCATO, M. AI in the Common Interest. Project Syndicate. December 22, 2022. <https://www.project-syndicate.org/>
- RAS, G.; VAN GERVEN M. e HASELAGER, P. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. (o artigo é um capítulo de Explainable and Interpretable Models in Computer Vision and Machine Learning, da Springer series on Challenges in Machine Learning). *ArXiv* 2018. <https://arxiv.org/pdf/1803.07517.pdf>. Acesso em: 03 mar. 2022
- RAMOS, L. e SUBRAMANYAM, J. Maverick Research: Forget About Your Real Data — Synthetic Data Is the Future of AI. *Gartner Research*. 2021. <https://www.gartner.com/en/documents/4002912>
- RAMALHO, A. Patentability of AI-Generated Inventions: Is a Reform of the Patent System Needed? SSRN 2018. <https://ssrn.com/abstract=3168703>. Acesso em: 20 ago. 2021.
- RAMPONI, M. The Full Story of Large Language Models and RLHF. *AssemblyAI*. May 3, 2023. https://www.assemblyai.com/blog/the-full-story-of-large-language-models-and-rlhf/?utm_source=substack&utm_medium=email
- RANGEL, D. e WALLACH, L. Rethink Trade Report. March. 2023. <https://rethinktrade.org/wp-content/uploads/2023/03/International-Preemption-by-Trade-Agreement.pdf>. Acesso em: 28 set. 2023

- REUTERS. Microsoft's \$3.2 bln UK investment to drive AI growth. Nov 30, 2024. <https://www.reuters.com/technology/microsofts-32-bln-uk-investment-drive-ai-growth-2023-11-30/> (2024a)
- REUTERS. Microsoft to make \$2.7 billion cloud, AI investments in Brazil. Sept.26, 2024. <https://www.reuters.com/technology/microsoft-make-27-billion-cloud-ai-investments-brazil-2024-09-26/> (2024b)
- REUTERS. US compels Saudi fund to exit Altman-backed AI chip startup - Bloomberg News. Nov 30, 2023. <https://www.reuters.com/technology/us-compels-saudi-fund-exit-altman-backed-ai-chip-startup-bloomberg-news-2023-11-30/>
- REUTERS. Putin signs law forcing foreign social media giants to open Russian offices. July 1, 2021. <https://www.reuters.com/technology/putin-signs-law-forcing-foreign-it-firms-open-offices-russia-2021-07-01/>
- REUTERS. Google announces patent agreement with Tencent amid China push. Jan. 19, 2018. <https://www.reuters.com/article/us-china-google-tencent/google-announces-patent-agreement-with-tencent-amid-china-push-idUSKBN1F80DF/>
- RICCHIUTI, A. Training artificial intelligence: how the DSM directive's Text and Data Mining exception shaped the rules of AI Training and possible future outcomes. *Lawels*, June 19, 2023. <https://lawels.com/en/2023/06/19/training-artificial-intelligence-how-the-dsm-directives-text-and-data-mining-exception-shaped-the-rules-of-ai-training-and-possible-future-outcomes/>
- RICHTER, F. Amazon Maintains Cloud Lead as Microsoft Edges Closer. *Statista*. May 2, 2024. <https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/>. Acesso em: 28 set. 2024
- RIKAP, C. Varieties of corporate innovation systems and their interplay with global and national systems: Amazon, Facebook, Google and Microsoft's strategies to produce and appropriate artificial intelligence. *Review of International Political Economy*, 1–29, 2024. <https://doi.org/10.1080/09692290.2024.2365757>. Acesso em: 30 jun. 2024
- RIKAP, C. Capitalism as Usual? *New Left Review*, 139. Jan-Feb 2023. <https://newleftreview.org/issues/ii139/articles/capitalism-as-usual>. Acesso em: 24 jul. 2024 (2023a)
- RIKAP, C. Who Learns and Who Profits in the Era of Artificial Intelligence? *BotPopuli*. July 20. <https://botpopuli.net/who-learns-and-who-profits-in-the-era-of-artificial-intelligence/>. Acesso em: 23 jul.2023 (2023b)
- RIKAP, C. Same End by Different Means: Google, Amazon, Microsoft and Facebook's Strategies to Dominate Artificial Intelligence. SSRN 2023. <http://dx.doi.org/10.2139/ssrn.4472222>. Acesso em: 10 jun.2023 (2023c)
- RIKAP, C. The expansionary strategies of intellectual monopolies: Google and the digitalization of healthcare. *Economy and Society* 52 (1), 2023. <https://doi.org/10.1080/03085147.2022.2131271>. (2023d)
- RIKAP, C. Enclosing the internet - Big Tech's cloud cover. *Social Europe*. Sept 6, 2022. <https://www.socialeurope.eu/enclosing-the-internet-big-techs-cloud-cover>. Acesso em: 15 set.2022
- RIKAP, C. Microsoft: ganadora de las acusaciones contra las gigantes digitales. *Ambito* 2020 <https://www.ambito.com/opiniones/microsoft/ganadora-las-acusaciones-contra-las-gigantes-digitales-n5146339>. Acesso em: 08 dez.2021
- RIKAP, C. Capitalism, Power and Innovation. Routledge London. March. 2021. <https://doi.org/10.4324/9780429341489>
- RIKAP, C. e LUNDVALL, B.A The Digital Innovation Race: Conceptualizing the Emerging New World Order. Palgrave Macmillan. December. 2021
- ROACH, J. How Microsoft's bet on Azure unlocked an AI Revolution. March 13, 2023. <https://news.microsoft.com/source/features/ai/how-microsofts-bet-on-azure-unlocked-an-ai-revolution>. Acesso em: 30 jun.2023
- ROBINS-EARLY, N. Google broke law to maintain online search monopoly, US judge rules. *The Guardian*. Aug 6, 2024. <https://www.theguardian.com/technology/article/2024/aug/05/google-loses-antitrust-lawsuit> Acesso em: 07 ago. 2024
- ROBINSON, D. European Commission checks AI chip market for stifled competition. *The Register*. Oct 02, 2023. https://www.theregister.com/2023/10/02/ec_competition_ai/. Acesso em: 09 ago.2024
- RODRIK, D. Washington's New Narrative for the Global Economy. *Project Syndicate*. May 5, 2023. <https://www.project-syndicate.org/commentary/biden-administration-international-economic-agenda-must-not-undermine-global-economy-by-dani-rodrik-2023-05>. Acesso em: 04 jun.2023
- ROMBACH, R. et al. High-Resolution Image Synthesis with Latent Diffusion Models. *ArXiv* 2021. <https://arxiv.org/abs/2112.10752>
- ROMERA-PAREDES, B. et al. Mathematical discoveries from program search with large language models. *Nature* 625, 2024, pp. 468–475. <https://doi.org/10.1038/s41586-023-06924-6>
- ROMERO, A. The State of Generative AI, 2024. The Algorithmic Bridge. Apr 3, 2024. https://www.thealgorithmicbridge.com/p/the-state-of-generative-ai-2024?r=2rchy2&utm_campaign=post&utm_medium=email&triedRedirect=true . Acesso em: 05 abr.2024

- ROSENBERG, N. *Studies on Science and the Innovation Process: Selected Works of Nathan Rosenberg*. World Scientific Publishing Co. Pte. Ltd. 2010.
<https://www.worldscientific.com/worldscibooks/10.1142/7306#t=aboutBook>.
- ROSENBERG, N. *America's Entrepreneurial Universities* in HART, D. M. (ed.) *The Emergence of Entrepreneurial Policy*. Cambridge University Press (Chapter 7). 2003
- ROSENBERG, N. *Uncertainty and Technological Change*. Routledge. 1998
- ROSENBERG, N. Why do firms do basic research (with their own money)? *Research Policy* 19, 1990, pp.165-174
- RUBIERA, C.O. CASP14: what Google DeepMind's AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics. Oxford Protein Informatics Group. Dec 3, 2020.
<https://www.bloig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/> Acesso em: 06 dez.2022
- RUBINGER, B. A case study comparison of the AI chips patent landscape. *IAM*. Feb 19, 2020.
<https://www.iam-media.com/article/case-study-comparison-of-the-ai-chips-patent-landscape>. Acesso em: 04 ago.2024
- RUBINSTEIN, D. OSI officially releases its definition for Open Source AI. *SD Times*. Oct 28, 2024. Disponível em: <https://sdtimes.com/ai/osi-officially-releases-its-definition-for-open-source-ai/>. Acesso em: 28 out. 2024
- RUDIN, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *ArXiv*. 2019. <https://arxiv.org/pdf/1811.10154.pdf>
- RUDIN, C. e RADIN, J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition. *Harvard Data Science Review*, 1(2). 2019
<https://doi.org/10.1162/99608f92.5a8a3a3d>
- SALAKHUTDINOV, R. e HINTON, G.E. An Efficient Learning Procedure for Deep Boltzmann Machines. *Neural Computation* 24, 2012, pp. 1967–2006. https://www.cs.cmu.edu/~rsalakhu/papers/neco_DBM.pdf
- SALAKHUTDINOV, R. e HINTON, G.E. Deep Boltzmann Machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, Clearwater Beach, Florida, USA. Volume 5 of *JMLR:W&CP5*. <https://proceedings.mlr.press/v5/salakhutdinov09a/salakhutdinov09a.pdf>
- SAMUELSON, P.A. The Pure Theory of Public Expenditure. *The Review of Economics and Statistics* 36 (4), 1954, pp.387-389. https://courses.cit.cornell.edu/econ335/out/samuelson_pure.pdf
- SAMUELSON, P. Allocating Ownership Rights in Computer-Generated Works. *University of Pittsburgh Law Review* 47, pp. 1185-1208, 1985. Disponível em: <https://www.law.berkeley.edu/wp-content/uploads/2024/01/Pam-Samuelson-Allocating-Ownership-Rights-in-Computer-Generated-Works.pdf>. Acesso em: 12 out. 2024
- SANDELOWSKI et al. Mapping the Mixed Methods-Mixed Research Synthesis Terrain. *Journal of Mixed Methods Research* 6(4), 2012, pp. 317-331. doi: 10.1177/1558689811427913.
- SATARIANO, A. People think that data is in the cloud, but it's not. It's in the ocean. *The New York Times*. March 19, 2019. <https://www.nytimes.com/interactive/2019/03/10/technology/internet-cables-oceans.html>. Acesso em: 03 set.2024
- SATO, K. e YOUNG, C. An in-depth look at Google's first Tensor Processing Unit (TPU). *Google Cloud*. May 12, 2017. <https://cloud.google.com/blog/products/ai-machine-learning/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu>. Acesso em: 12 ago. 2024
- SAVONA, M. The Value of Data: Towards a Framework to Redistribute It. *SWPS* 21. October, 2019. <https://ssrn.com/abstract=3476668>
- SCHIRRU, L. Direito Autoral e inteligência Artificial: Autoria e titularidade nos produtos da IA, Tese de Doutorado, PPED/UFRJ. RJ. 2020. https://www.researchgate.net/publication/344677489_Direito_Autoral_e_Inteligencia_Artificial_Autoria_e_Titularidade_nos_Produtos_da_IA
- SCHROEDER, W. Germany's Industry 4.0 strategy Rhine capitalism in the age of digitalization. *FES London*. November, 2016. <https://uk.fes.de>
- SCHUMANN, C. e RICCO, S. A Step Toward More Inclusive People Annotations in the Open Images Extended Dataset. *Google Research* June 15, 2021. <https://research.google/blog/a-step-toward-more-inclusive-people-annotations-in-the-open-images-extended-dataset/>. Acesso em: 04 jul.2023
- SCHUMPETER, J.A. *Capitalism, Socialism and Democracy*. Routledge London & New York. 1942
- SCHWAB, K. The Fourth Industrial Revolution: what it means, how to respond. *World Economic Forum*. 2016. <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>
- SCOTT, K. Microsoft teams up with OpenAI to exclusively license GPT-3 language model. *Official Microsoft Blog*. Sept. 09, 2020. <https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/>. Acesso em: 20 mai.2022
- SCSP Generative AI. The future of innovation power. *Special Competitive Studies Project*. 2023. <https://www.scsp.ai/reports/gen-ai/>

- SEARLE, J.R. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3), 1980, pp.417-424. doi:10.1017/S0140525X00005756
- SEJNOWSKI, T.J. (2018) *The Deep Learning Revolution*. MIT Press
- SÉRVIO, G. Project Astra: Google revela assistente de IA capaz de ‘ver e lembrar’. *Olhar Digital*. Maio 14, 2024. <https://olhardigital.com.br/2024/05/14/internet-e-redes-sociais/project-astra-google-revela-assistente-de-ia-capaz-de-ver-e-lembrar/>
- SENFLEBEN, M. Generative AI and Author Remuneration. *International Review of Intellectual Property and Competition Law* 54, pp. 1535-1560. 2023. Disponível em: <https://ssrn.com/abstract=4478370>.
- SEVILLA, J. et al. Can AI Scaling Continue Through 2030?". *Published online at epochai.org*. 2024. <https://epochai.org/blog/can-ai-scaling-continue-through-2030>
- SEVILLA, J. et al. Compute trends across three eras of machine learning. *ArXiv* 2022 <https://arxiv.org/abs/2202.05924>
- SHAFTER, P. Why big tech companies are open-sourcing their AI systems. *The Conversation*. Feb 16, 2016. <https://theconversation.com/why-big-tech-companies-are-open-sourcing-their-ai-systems-54437>
- SHAPIRO, K. Antitrust limits to patent settlements. *RAND Journal of Economics* 34 (2), 2003, pp.391–411
- SHANFELD, E. Ghostwriter’s ‘Heart on My Sleeve,’ the AI-Generated Song Mimicking Drake and the Weeknd, Submitted for Grammys. *Variety*. Sep 6, 2023. <https://variety.com/2023/music/news/ai-generated-drake-the-weeknd-song-submitted-for-grammys-1235714805/>
- SHEN, X. China leads world in number of generative AI start-ups to receive funding in first half of 2023, report finds. *SCMP*. July 10, 2023. <https://www.scmp.com/tech/tech-trends/article/3227197/china-leads-world-number-generative-ai-start-ups-receive-funding-first-half-2023-report-finds>. Acesso e: 10 dez.2023
- SHILOV, A. Nvidia's AI GPUs Are Selling for up to \$70,000 in China. *Tom's Hardware*. July 29, 2023. <https://www.tomshardware.com/news/price-of-Nvidia-compute-gpu-can-hit-70000-in-china>. Acesso em: 12 ago. 2024 (2023a)
- SHILOV, A. China imposes new export restrictions on gallium and germanium. *Anandtech*, Aug 2, 2023. <https://www.anandtech.com/show/19989/china-imposes-new-export-restrictions-on-gallium-and-germanium>. (2023b)
- SHROBE, H. The Innovative Applications of Artificial Intelligence Conference: Past and Future. *AI Magazine* 17(4), 1996. <https://doi.org/10.1609/aimag.v17i4.1236>
- SHUMAILOV, I. et al. The curse of recursion: training on generated data makes models forget. *ArXiv*. 2023. <https://arxiv.org/abs/2305.17493>
- SIA The State of the US Semiconductor Industry. Semiconductor Industry Association. 2021 <https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf>
- SIDAK, J.G. The Meaning of Frand, Part I: Royalties. *The Journal of Competition Law & Economics*, 9(4), 2013, pp. 931–1055. doi:10.1093/joclec/nht040
- SIMONITE, T. Despite Pledging openness Companies rush to patent AI Tech. *Wired*. Jul 31, 2018. <https://www.wired.com/story/despite-pledging-openness-companies-rush-to-patent-ai-tech/>.
- SINGH, P. What are Multimodal Models? *Analytics Vidhya*. Dec.25, 2023. <https://www.analyticsvidhya.com/blog/2023/12/what-are-multimodal-models>. Acesso em: 14 jan.2024
- SINGHAL, K. A Scalable Approach for Partially Local Federated Learning. Google Research. Dec 16, 2021. <https://research.google/blog/a-scalable-approach-for-partially-local-federated-learning/>. Acesso em: 12 jan. 2022
- SMITH, T. Why top AI talent is leaving Google’s DeepMind. *Sifted*. Nov 3, 2023. <https://sifted.eu/articles/deepmind-talent>. Acesso em: 12 ago. 2024
- SOCHER, R. et al. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011 https://icml.cc/2011/papers/125_icmlpaper.pdf. Acesso em: 10 set.2023
- SOLAIMAN, I. The Gradient of Generative AI Release: Methods and Considerations. *ArXiv*. 2023. <https://arxiv.org/abs/2302.04844>
- SRNICEK, N. Paths Forward for the Study of the Digital Economy. In J. Muldoon, & W. Stronge (Eds.), *Platforming Equality: Policy Challenges for the Digital Economy* (pp. 85-90). 2020. https://kclpure.kcl.ac.uk/portal/files/135510358/Srnicsek_Nick_2020_Paths_Foward_for_the_Study_of_the_Digital_Economy_.pdf
- SRNICEK, N. (2017) *Platform Capitalism*. Cambridge: Polity Press.
- STATE OF AI REPORT COMPUTE INDEX V3 (2024). Feb 06. <https://stateofai.substack.com/p/state-of-ai-report-compute-index-v3>. Acesso em: 07 fev.2024
- STATE OF AI REPORT COMPUTE INDEX. State of AI Report Compute Index v3 Update. Air Street Press. 2024. Disponível em: <https://press.airstreet.com/p/state-of-ai-report-compute-index-v3>. Acesso em: 01 out 2024
- STIERLE, M. Artificial Intelligence Designated as Inventor – An Analysis of the Recent EPO Case Law. *GRUR International* 69(9), 2020, pp. 918–924. <https://academic.oup.com/grurint/article/69/9/918/5880762>

- STIGLITZ, J.E. Knowledge as a Global Public Good, in Kaul, I., Grunberg, I. e Stern, M. (eds), *Global Public Goods: International Cooperation in the 21st Century*, New York, 1999; online edn, Oxford Academic. <https://doi.org/10.1093/0195130529.003.0015>.
- STONE, B. e VANCE, A. Apple's obsession with secrecy grows stronger. *The New York Times* June 22, 2009. <https://www.nytimes.com/2009/06/23/technology/23apple.html>. Acesso em: 08 jun.2023
- STOCKMAN, F. How the Biden Administration Took the Pen Away From Meta, Google and Amazon. *The New York Times* Nov 27, 2023. <https://www.nytimes.com/2023/11/27/opinion/google-facebook-ai-trade.html>
- SUN, Z. LEI, Z. e YIN, Z. Innovation policy in China: nationally promulgated but locally implemented. *Innovation policy in China: nationally promulgated but locally implemented. Applied Economics Letters*, 25(21), 2018, pp. 1481–1486. <https://doi.org/10.1080/13504851.2018.1430315>
- SURESH, S. Generative AI: Creating machines more human-like. Kaggle AI Report. Aug 1, 2023. <https://www.kaggle.com/code/sanjushasuresh/generative-ai-creating-machines-more-human-like/notebook>. Acesso em: 22 set. 2023
- SWANSON, D. G. e BAUMOL, W. J. Reasonable and non-discriminatory (RAND) royalties, standards selection, and control of market power. *Antitrust Law Journal*, 73(1), 2005, pp. 1-58. <https://www.jstor.org/stable/40037598>
- TADDY, M. The technological elements of artificial intelligence. *National Bureau of Economic Research*. 2018. DOI 10.3386/w24301
- TEECE, D. J. Technological Leadership and 5G Patent Portfolios: Guiding Strategic Policy formulation and Licensing Decisions. *California Management Review*, 63 (3), 2021, pp. 5-34. <https://doi.org/10.1177/00081256211007584>
- TEECE, D. J. Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world, *Research Policy*, 47(8), 2018, pp. 1367-1387. <https://doi.org/10.1016/j.respol.2017.01.015>. Acesso em: 11 dez. 2020 (2018a)
- TEECE, D. J. Enabling Technologies and the Underinvestment Problem. July 10, 2018. <https://ipwatchdog.com/2018/07/10/enabling-technologies-underinvestment-problem/id=98824/> (2018b)
- TEECE, D. J. Profiting from Innovation in the Digital Economy: Standards, Complementary Assets, and Business Models In the Wireless World. Tusher Center for the Management on Intellectual Capital. Working Paper Series N° 16, August. 2016. <https://www.haas.berkeley.edu/wp-content/uploads/Tusher-Center-Working-Paper-No.-16.pdf>. Acesso em: 03 mai. 2022
- TEECE, D.J. Next generation competition: new concepts for understanding how innovation shapes competition and policy in the digital economy. *Journal of Law, Economics and Policy* 9 (1), 2012, pp. 97–118
- TEECE, D.J. Reflections on profiting from innovation. *Research Policy* 35 (8), 2006, pp. 1131–1146. <https://doi.org/10.1016/j.respol.2006.09.009>. Acesso em: 12 mai. 2021
- TEECE, D. J. Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research policy* 15(6), 1986, pp. 285-305. Disponível em [https://doi.org/10.1016/0048-7333\(86\)90027-2](https://doi.org/10.1016/0048-7333(86)90027-2). Acesso em: 11 abr. 2020.
- TEECE, D. J. e SHERRY, E. On Patent 'Monopolies': An Economic Re-Appraisal. *Antitrust Chronicle*, Spring. 2027. <https://ssrn.com/abstract=2962208>
- TEECE, D.J.; PISANO, G. e SHUEN, A. Dynamic Capabilities and Strategic Management, *Strategic Management Journal*, V.18 (7), 1997, pp. 509-533. [https://doi.org/10.1002/\(SICI\)1097-0266\(199708\)18:7<509::AID-SMJ882>3.0.CO;2-Z](https://doi.org/10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z)
- TENCENT. Tencent Unveils Hunyuan, its Proprietary Large Foundation Model on Tencent Cloud. *Tencent*. Sept. 7, 2023. <https://www.tencent.com/en-us/articles/2201685.html>. Acesso em: 10 ago.2024
- THE BATCH. Training Data Free-For-All: Japan's AI data laws, explained. *The Batch*. June 14, 2023. <https://www.deeplearning.ai/the-batch/japan-ai-data-laws-explained/>. Acesso em: 10 ago.2024
- THE ECONOMIST. OpenAI's legal battles are not putting off customers—yet. Mar 6, 2024. <https://www.economist.com/business/2024/03/06/openais-legal-battles-are-not-putting-off-customers-yet> . The secrets of big tech. Acesso em: 07 mar.2024
- THE ECONOMIST. AI is setting off a great scramble for data. Aug 13, 2023. <https://www.economist.com/business/2023/08/13/ai-is-setting-off-a-great-scramble-for-data>. Acesso em: 15 ago.2023 (2023a)
- THE ECONOMIST. Big tech and geopolitics are reshaping the internet's plumbing. Dec 20, 2023. <https://www.economist.com/business/2023/12/20/big-tech-and-geopolitics-are-reshaping-the-internets-plumbing>. Acesso em: 25 dez.2023 (2023b)
- The Economist. One of the “godfathers of AI” airs his concerns. Jul 21, 2023. <https://www.economist.com/by-invitation/2023/07/21/one-of-the-godfathers-of-ai-airs-his-concerns>? Acesso em: 06 ago.2023 (2023c)
- THE ECONOMIST. Nvidia is not the only firm cashing in on the AI gold rush. May 29, 2023. <https://www.economist.com/business/2023/05/29/Nvidia-is-not-the-only-firm-cashing-in-on-the-ai-gold-rush>. Acesso em: 02 jun.2023 (2023d)

- The Economist. The AI boom has turbocharged Nvidia's fortunes. Can it hold its position? Jun 1, 2023. <https://www.economist.com/leaders/2023/06/01/the-ai-boom-has-turbocharged-Nvidias-fortunes-can-it-hold-its-position>. Acesso em: Acesso em: 02 jun.2023 (2023e)
- The Economist. A battle royal is brewing over copyright and AI. May 15, 2023. <https://www.economist.com/business/2023/03/15/a-battle-royal-is-brewing-over-copyright-and-ai>. Acesso em: 02 jun.2023 (2023f)
- THE ECONOMIST. The secrets of big tech. Apr 30, 2022. <https://www.economist.com/business/the-finance-secrets-of-big-tech/21808956>. Acesso em: 02 jun.2023
- THE ECONOMIST. How the pandemic has changed the weather in the technology industry. Oct 30, 2021. <https://www.economist.com/business/2021/10/30/how-the-pandemic-has-changed-the-weather-in-the-technology-industry?> Acesso em: 12 dez.2021 (2021a)
- THE ECONOMIST. The IT establishment is dressing in new clothes. Nov 4, 2021. <https://www.economist.com/business/the-it-establishment-is-dressing-in-new-clothes/21806076>. Acesso em: 12 dez.2021 (2021b)
- THOPPILAN, R. et al. LaMDA: Language Models for Dialog Applications. *ArXiv*. 2022. <https://arxiv.org/abs/2201.08239>
- TYAGI, K. Synthetic Data, Data Protection and Copyright in an era of Generative AI. 2024. 10.13140/RG.2.2.20871.41120.
- TIGRE, P.B. e MARQUES, F.S. Apropriação tecnológica na economia do conhecimento: inovação e propriedade intelectual de software na América Latina. *Economia e Sociedade* 18 (3), 2009, pp. 547-566. <https://www.scielo.br/j/ecos/a/D6CjncGN4PyYvmf7yktMzbL/?format=pdf&lang=pt>
- TOMIC, S. D. K. How Artificial Intelligence (AI) Innovators Profit from Innovation in Digital Platform Ecosystems: An Explorative Business Model Study. Master Thesis submitted for the degree of "Master of Business Administration", Continuing Education Center, Vienna. 2018
- TOOLE, A.A. et al. Inventing AI Tracing the diffusion of artificial intelligence with U.S. patents. USPTO, IP DATA HIGHLIGHTS N.5, October. 2020
- TORRISI, S.; GAMBADELLA, A.; GIURI, P.; HARNOFF, D.; HOISL, K. e MARIANI, M. Used, Blocking and Sleeping Patents: Empirical Evidence from a Large-Scale Inventor Survey, *Research Policy* 45 (7), 2016, pp. 1374-1385
- TRAJTENBERG, M.A. AI as the next GPT: a Political-Economy Perspective. *NBER Working Paper N. 24245*. Cambridge, MA. January, 2018. DOI 10.3386/w24245
- TRINH, T.H. et al. Solving olympiad geometry without human demonstrations. *Nature* 625, 2024, pp. 476-482. <https://doi.org/10.1038/s41586-023-06747-5>
- TSANG, Y.P. e LEE, C.K.M. Artificial intelligence in industrial design: A semi-automated literature survey. *Engineering Applications of Artificial Intelligence* 112, 2022. <https://doi.org/10.1016/j.engappai.2022.104884>.
- UNCTAD. Digital Economy Report 2024: Shaping an environmentally sustainable and inclusive digital future. United Nations Publications. N. York USA. 2024. https://unctad.org/system/files/official-document/der2024_en.pdf
- UNCTAD. Digital Economy Report, Value Creation and Capture: Implications for Developing Countries. United Nations Conference on Trade and Development. N. York USA. 2019 <https://unctad.org/publication/digital-economy-report-2019>
- US GEOLOGICAL SURVEY. Mineral commodity summaries 2024: U.S. Geological Survey, 212 p. 2024. <https://doi.org/10.3133/mcs2024>.
- USTR. Key Barriers to Digital Trade. United States Trade Representatives. March, 2017. <https://ustr.gov/about-us/policy-offices/press-office/fact-sheets/2017/march/key-barriers-digital-trade>
- THE WHITE HOUSE. FACT SHEET: Securing a Made in America Supply Chain for Critical Minerals. Feb. 22, 2022. <https://www.whitehouse.gov/briefing-room/statements-releases/2022/02/22/fact-sheet-securing-a-made-in-america-supply-chain-for-critical-minerals/>
- TROTT, P. e HARTMANN, D. Why "Open Innovation" is Old Wine in New Bottles. *International Journal of Innovation Management* 13 (4), 2009, pp. 715-736
- UMPLEBY, S. Heinz von Foerster and the Mansfield Amendment. *Cybernetics And Human Knowing* 10 (3-4), 2003, pp.187-190. https://www.researchgate.net/publication/233661621_Heinz_von_Foerster_and_the_Mansfield_Amendment
- VAN DEN OORD, A. et al. WaveNet: A Generative Model for Raw Audio. *ArXiv*. 2016. <https://arxiv.org/abs/1609.03499>
- VAN DIJCK, J. Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, 23(9), 2021, pp. 2801-2819. <https://doi.org/10.1177/1461444820940293>
- VAN LENT, M.; FISHER, W. e MANCUSO, M. An Explainable Artificial Intelligence System for Small-unit Tactical. In: Proceedings of the national conference on artificial intelligence. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900-907.

- VAN VEEN, F. e LEIJNEN, S. The neural network zoo. The Azimov Institute. Sept 2019. <https://www.asimovinstitute.org/neural-network-zoo/>
- VANNUCCINI, S.; PRYTKOVA, E. Artificial Intelligence's New Clothes? From General Purpose Technology to Large Technical System. SSRN. October 4, 2020. <https://ssrn.com/abstract=3704011>
- VARIAN, H. Artificial Intelligence, Economics, and Industrial Organization. NBER Working Paper No. 24839. Cambridge, MA. July 2018
- VASWANI, A. et al. Attention Is All You Need. *ArXiv*. 2017. <https://arxiv.org/abs/1706.03762>
- VERGARA, I. Comment Satya Nadella a redonné à Microsoft toute sa puissance technologique. *Le Figaro*. Jan 31, 2024. <https://www.lefigaro.fr/secteur/high-tech/comment-satya-nadella-a-redonne-a-microsoft-toute-sa-puissance-technologique-20240131>. Acesso em: 31 jan.2024
- VERTINSKY, L. Thinking Machines and Patent Law in Barfield, W., Pagallo, U. (eds) Research Handbook of Artificial Intelligence, Edward Elgar Publishing (Cap 18). 2018. <http://dx.doi.org/10.2139/ssrn.3036030>. Acesso em: 06 ju. 2023.
- VEZZOSO, S. The first new brandeisian decision in Europe?: Insights from the Utah Statement. *Concurrences* N° 2-22. Art. N° 106132, 2022, pp. 70-77. <https://www.concurrences.com/en/review/issues/no-2-2022/articles/the-first-new-brandeisian-decision-in-europe-insights-from-the-utah-statement>. Acesso em: 15 jul.2023
- VILLALOBOS, P. et al. Will we run out of data? Limits of LLM scaling based on human-generated data. *ArXiv*. 2024. <https://arxiv.org/abs/2211.04325>. Acesso em: 05 jun. 2024
- VILLALOBOS, P. et al. Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning. *ArXiv*. 2022. <http://arxiv.org/abs/2211.04325>.
- VINCENT, J. Meta's powerful AI language model has leaked online — what happens now? *The Verge* Mar 8, 2023. <https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse>. Acesso em: 14 mar. 2023
- VIPRA, J. e WEST, S.M. Computational power and AI. AI Now Institute. Sept 23, 2023. <https://ainowinstitute.org/publication/policy/compute-and-ai>. Acesso em: 02 jul . 2024
- VOLPICELLI, G. ChatGPT broke the EU plan to regulate AI. *Politico*. Mar 23, 2023 <https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/>
- VON KÜGELGEN, J. et al. On the Fairness of Causal Algorithmic Recourse. *ArXiv*. 2022. <https://arxiv.org/abs/2010.06529>
- WADE, R.H. The American paradox: ideology of free markets and the hidden practice of directional thrust, *Cambridge Journal of Economics* 41 (3), 2017, pp. 859–880. <https://doi.org/10.1093/cje/bew064>
- WADE, R.H. Return of Industrial Policy. . *International Review of Applied Economics* 26 (2), 2012, pp. 223-239. DOI: 10.1080/02692171.2011.640312
- WALSH, J. Pentagon Plans Boost for Basic Research: Secretary Brown's memo should help clear the confusion on criterion of relevance left by Mansfield Amendment. *Science* 205(4406), 1979, pp. 566-568. DOI: 10.1126/science.205.4406.566
- WANG, A. et al. Poster: A SDN-based Network Layer for Edge Computing. SEC '19, November 7–9, 2019. Arlington, VA, USA. <https://par.nsf.gov/servlets/purl/10139183>. Acesso em: 12 jan. 2022
- WANG, Y. et al. Case Study: How Alibaba Uses AI Chatbots to Serve a Billion Customers. *AI Business* Feb 14, 2024. <https://aibusiness.com/ml/the-alibaba-challenge-how-to-effectively-engage-with-a-billion-customers->. Acesso em: 04 mai.2024
- WANG, R. How amended regulations to China's Patent Law will impact the IP landscape. *IAM* Jan 10, 2024. [https://www.iam-media.com/article/how-amended-regulations-chinas-patent-law-will-impact-the-ip-landscape#:~:text=The%20Fourth%20Amendment%20to%20the%20Patent%20Law%20introduces%20a%20patent,\(patent%2Dterm%20extension\).](https://www.iam-media.com/article/how-amended-regulations-chinas-patent-law-will-impact-the-ip-landscape#:~:text=The%20Fourth%20Amendment%20to%20the%20Patent%20Law%20introduces%20a%20patent,(patent%2Dterm%20extension).)
- WARREN, E. Big Tech's Big Con: Rigging Digital Trade Rules to Block Antitrust Regulation. *USTR Report*. May, 2023. [https://www.warren.senate.gov/imo/media/doc/USTR REPORT.pdf](https://www.warren.senate.gov/imo/media/doc/USTR%20REPORT.pdf)
- Webster et al. Forum: Analyzing an Expert Proposal for China's Artificial Intelligence Law. *Digichina* August 23, 2023. <https://digichina.stanford.edu/work/forum-analyzing-an-expert-proposal-for-chinas-artificial-intelligence-law/>.
- WEF. Artificial Intelligence Collides with Patent Law. World Economic Forum White Paper. April, 2018. http://www3.weforum.org/docs/WEF_48540_WP_End_of_Innovation_Protecting_Patent_Law.pdf. Acesso 13/02/2021
- WEST, S.M. e VIPRA, J. Computational Power and AI: Comment Submission. AI Now Institute. Jun 22, 2023. <https://ainowinstitute.org/publication/policy/computational-power-and-ai>. Acesso em: 19 ago. 2024
- WHITE, M. J. How much does an AI supercomputer cost? Try 100 billion. *Digital Trends*. March 29, 2024. <https://www.digitaltrends.com/computing/microsoft-openai-working-on-a-data-center-project/>. Acesso em: 28 abr. 2024
- WIDDER, D.G.; WEST, S. e WHITTAKER, M. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. Aug, 2023. <http://dx.doi.org/10.2139/ssrn.4543807>. Acesso em: 03 jan. 2024

- WIGGERS, K. Amazon hires founders away from AI startup Adept. *TechCrunch* Jun 28, 2024. <https://techcrunch.com/2024/06/28/amazon-hires-founders-away-from-ai-startup-adept/?guccounter=1>. Acesso em: 28 jun. 2024.
- WILLIAMSON, O. E. Transaction-Cost Economics: The Governance of Contractual Relations. *Law Journal of Law and Economics* 22(2), 1979, pp. 233-261.
- WIPO. Generative Artificial Intelligence. Patent Landscape Report. Geneva: WIPO. 2024. <https://doi.org/10.34667/tind.49740> (2024a)
- WIPO. AI Inventions. IP and Frontier Technologies. Geneva: WIPO. 2024. Disponível em: <https://doi.org/10.34667/tind.49113>. Acesso em: 12 out. 2024 (2024b)
- WIPO. Generative AI Navigating intellectual property. IP and Frontier Technologies Factsheet. 2024. Disponível em: <https://www.wipo.int/edocs/pubdocs/en/wipo-pub-rn2024-8-en-generative-ai-navigating-intellectual-property.pdf>. Acesso em: 20 out. 2024 (2024c)
- WIPO. Interventions to the WIPO Conversation on IP and AI. 2020. https://www.wipo.int/about-ip/en/artificial_intelligence/policy.html. Acesso em: 08 mar.2020. (2020a)
- WIPO. WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session. Revised Issues paper on intellectual property policy and artificial intelligence. WIPO Secretariat. May 21, 2020. https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1_rev.pdf. Acesso em: 10 jul.2020. (2020b)
- WIPO. Technology Trends 2019: Artificial Intelligence. Geneva: World Intellectual Property Organization. 2019. <https://www.wipo.int/publications/en/details.jsp?id=4386>. Acesso: 06 dez.2019
- WONG-ERVIN, K. e WRIGHT, J. D. Intellectual Property and Standard Setting. *The Federalist Society Review* 17 (3). George Mason Law & Economics Research Paper No. 2016, pp. 16-48. <http://dx.doi.org/10.2139/ssrn.2878955>
- WAGNER, J. China's Cybersecurity Law: What You Need to Know. *The Diplomat*. June 1, 2017. <https://thediplomat.com/2017/06/chinas-cybersecurity-law-what-you-need-to-know/>
- WU, T. The Curse of Bigness: Antitrust in the New Gilded Age. Columbia Global Reports. 2018
- XIANG, C. OpenAI's GPT-4 Is Closed Source and Shrouded in Secrecy. *Vice*. March 16, 2023. <https://www.vice.com/en/article/openais-gpt-4-is-closed-source-and-shrouded-in-secrecy/>. Acesso em: 14 ago.2023
- XIN, Y.; WANG, M.. e ZHOU, Y. The development trend of artificial intelligence in medical: A patentometric analysis. *Artificial Intelligence in the Life Sciences*, Vol 1, 2021. <https://doi.org/10.1016/j.ailsci.2021.100006>.
- XU, P. ZHUN, X. e CLIFTON, D.A. Multimodal Learning with Transformers: A Survey. *ArXiv*. 2023. <https://arxiv.org/abs/2206.06488>
- XU, J. What does China's newly launched National Data Bureau mean to China and global data governance? *Internet Policy Review*, Apr. 25, 2023. <https://policyreview.info/articles/news/chinas-national-data-bureau-and-global-data-governance>. Acesso em: 22 ago. 2024
- YANG, S. Ending patent subsidies in China. *Mondaq*. Jan 18, 2022. https://www.mondaq.com/china/patent/1150552/ending-patent-subsidies-in-china#_edn10. Acesso em: 25 set.2023
- YANISKY-RAVID, S. e LIU, X. When Artificial Intelligence Systems Produce Inventions: The 3A Era and an Alternative Model for Patent Law. *Cardozo Law Review* (39), 2018, pp. 2215-2263. <http://dx.doi.org/10.2139/ssrn.2931828>. Acesso em: 20 ago. 2021.
- YIN, R. Qualitative research from start to finish. New York; London: The Guilford Press, Second ed. 2015
- YOFFIE, D.B.; GAWER, A. e CUSUMANO, M.A. A Study of More Than 250 Platforms Reveals Why Most Fail. *HBR* May 29, 2019. <https://hbr.org/2019/05/a-study-of-more-than-250-platforms-reveals-why-most-fail>
- YOU, H. et al. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *ArXiv*. 2023 <https://arxiv.org/abs/2310.07704>
- YUI, T. Understanding Neural Networks. Medium. July 2, 2019. <https://towardsdatascience.com/understanding-neural-networks-19020b758230>. Acesso em: 14 jun. 2021
- ZAO et al. A Survey of Large Language Models. *ArXiv*. 2023. <https://arxiv.org/abs/2303.18223>
- ZHANG, D. et al. The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March, 2022. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf (2022a)
- ZHANG, S. et al. OPT: Open Pre-trained Transformer Language Models. *ArXiv*. 2022 <https://arxiv.org/abs/2205.01068> (2022b)
- ZELLERS, R. et al. Defending Against Neural Fake News. *ArXiv*. 2020. <https://arxiv.org/abs/1905.12616>
- ZELLERS, R. Why We Released Grover. *The Gradient*. May 29, 2019. <https://thegradient.pub/why-we-released-grover/>. Acesso em: 02 jul. 2023
- ZHENG, L. et al. Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning. *ArXiv*. 2022. <https://arxiv.org/abs/2201.12023>

- ZHOU, K.; LIU, T. e SHOW, L. Industry 4.0: Towards future industrial opportunities and challenges. *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, China, 2015, pp. 2147-2152. doi: 10.1109/FSKD.2015.7382284
- ZIEDONIS, R.H. Don't Fence Me In: Fragmented Markets for Technology and the patent acquisition strategies of Firms. *Management Science* 50 (6), 2004, pp. 804-820. <https://doi.org/10.1287/mnsc.1040.0208>
- ZINGG, R. Foundational Patents in Artificial Intelligence', in Jyh-An Lee, Reto Hilty, and Kung-Chung Liu (eds), *Artificial Intelligence and Intellectual Property*,(Oxford; online edn, Oxford Academic, 22 Apr. 2021, <https://doi.org/10.1093/oso/9780198870944.003.0005>.
- ZITRON, E. How does OpenAI survive. July 29, 2024. <https://www.wheresyoured.at/to-serve-altman/?ref=ed-zitrons-where-yes-your-ed-at-newsletter>. Acesso: 03 ago.2024. (2024a)
- ZITRON, E. OpenAI is a bad business. Oct 2, 2024 <https://www.wheresyoured.at/to-serve-altman/?ref=ed-zitrons-where-yes-your-ed-at-newsletter>. Acesso: 04 out.2024. (2024b)
- ZUBOFF, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books. 2019