ENCCLA 2021

ESTRATÉGIA NACIONAL DE COMBATE À CORRUPÇÃO E À LAVAGEM DE DINHEIRO

RESUMO EXECUTIVO

Guia de boas práticas em *Big Data* e Inteligência Artificial

AÇÃO 08/2021:

GUIA DE BOAS PRÁTICAS EM *BIG DATA* E INTELIGÊNCIA ARTIFICIAL

Coordenação: AGU, PF

Colaboradores: ADPF, AGU, AJUFE, ANAPE, ANPR, BCB, BNDES, CADE, CAIXA, CG/DF, CGE/MG, CGM/SP, CGU, CJF, CNJ, CNMP, COAF, CONACI, CONCPC, DRCI, GNCOC, INSS, MD, MP/DFT, MP/GO, MP/MG, MP/MS, MP/PB, MP/RJ, MP/RN, MP/SE, MP/SP, MPF, MPM, MTP, PC/DF, PC/MA, PC/RJ, PC/RS, PC/SC, PC/SP, PF, PG/DF, PGFN, PGE/RS, PGM/SP, RFB, SECONT/ES, SEGES/ME, SF, TCU.

SUMÁRIO

1. HISTÓRICO

A implantação de políticas de *big data* e inteligência artificial tem-se apresentado como necessidade ínsita à organização dos órgãos de persecução e controle no combate à corrupção e lavagem de capitais. O constante e contínuo volume de informações, em nítido crescimento exponencial, dificulta sobremaneira detecção, prevenção e repressão a práticas ilícitas. Recursos analógicos de colheita e análise de dados atualmente não apresentam a mesma eficácia de outrora.

No entanto, ainda que neste contexto, *big data* e inteligência artificial nem sempre desaguam nos resultados práticos esperados. Isso se deve, muitas vezes, a equivocados planejamentos de implantação, ausência de infraestrutura e conhecimentos técnicos ou, até mesmo, porque a solução ao problema apresentado não se relaciona, em verdade, a qualquer das aludidas políticas.

Ao contrário do pensamento muitas vezes difundido, *big data* e inteligência artificial podem, muitas vezes, não ser aplicáveis à solução de problemas relacionados ao combate à corrupção e lavagem de capitais. Antes de mais nada mostra-se preciso, essencial, conhecer realmente a fundo o problema que se apresenta para, somente após, decidir-se pela aplicação (ou não) de determinada política de persecução e controle, ainda que restrita a recursos analógicos.

Mesmo dentre os casos manifestamente afetos à colheita e análise massiva de volumosos dados, a arquitetura construída para este mister pode conter equívocos perceptíveis apenas depois de longos períodos de investimento, seja em recursos financeiros ou humanos, maculando a própria credibilidade das políticas de *big data* e inteligência artificial.

De rigor frisar, neste aspecto, que estas políticas demandam altos custos para implantação e desenvolvimento e, ainda assim, dado o dinamismo da qualidade e volume das informações obtidas, não se afiguram estáticas, podendo vir a se tornar obsoletas em curtos intervalos de tempo. Imprescindível, à evidência, que a estrutura a ser implementada seja dotada de plasticidade suficiente não só a absorver as mudanças que se fizerem necessárias, mas principalmente aprimorar seus respectivos resultados.

Não obstante deslizes e eventuais fracassos, inegável a existência de instituições que atingiram níveis de excelência, tanto na implantação, execução, como obtenção de resultados lastreados em *big data* e inteligência artificial. O compartilhamento destas experiências é essencial para que outras instituições não repitam os mesmos erros, imprimam maior velocidade na implantação de estruturas digitais e, por fim, aprimorem os mecanismos já existentes.

Neste contexto é que se verifica a edição do presente guia de boas práticas em *big data* e inteligência artificial, tendo por escopo auxiliar gestores e instituições a implementar, da forma mais célere e eficiente possível, mecanismos avançados de combate à corrupção e lavagem de capitais.

2. DIAGNÓSTICO

Antes de se definir qual o destino ou patamar a ser atingido, imperioso conhecer e apontar, de fato, o verdadeiro ponto de partida ou origem do desafio a ser vencido. Em outras palavras, antes de se discutir qual a melhor forma de se construir o guia de boas práticas em *big data* e inteligência artificial, necessário conhecer como o tema vem sendo tratado pelas principais instituições de combate à corrupção e lavagem de capitais.

Nesse sentido, no âmbito da Estratégia Nacional de Combate à Corrupção e Lavagem de Capitais (ENCCLA) difundiu-se, no ano de 2021, questionário a diversas instituições perquirindo dados estatísticos a respeito da forma como o assunto é abordado atualmente.

O resultado da pesquisa demonstrou que, em geral, há grande deficiência na estruturação e organização de dados coletados, passo inicial importante para a implantação das políticas em apreço. Em muitas instituições há grandes acervos físicos, em alguns casos lentamente em processo de estruturação, em muitos outros, porém, sem perspectiva alguma disso. De outra parte, também se constatou que consideráveis instituições já recebem os dados de forma estruturada, seja originalmente, seja porque oriundos de outra instituição que, em momento anterior, concluíra previamente referida estruturação.

Nas situações em que superada a dificuldade atinente à estruturação e catalogação de dados, a pesquisa revelou que algumas instituições não possuem logística apropriada para tratamento e interpretação dos dados coletados. Em alguns casos isso se verifica pela total inobservância de regras uniformes de ortossemântica, isto é, ausência de utilização de vocabulários controlados, tanto na estruturação quanto catalogação dos dados. Em outros, pela existência de lacunas intransponíveis nos dados e elementos obtidos, a lhes retirar todo o valor social que poderiam ter em futuras análises por mecanismos de inteligência artificial.

Apurou-se, outrossim, a ausência de mecanismos de controle de veracidade dos dados e informações coletadas, salvo quando atinentes à própria atividade-fim da entidade mantenedora da informação. Diante do grande volume de dados produzidos minuto a minuto na sociedade atual, muitas instituições acabam por absorvê-los de outras instituições, e assim sucessivamente, numa reação em cadeia. Estes dados, invariavelmente, diante dos múltiplos e sucessivos compartilhamentos, em algum momento acabam por retornar à instituição de origem que, sem conhecer esta circunstância (de que os dados dela partiram inicialmente), acaba por validá-los sob o argumento de que se compatibilizam com aqueles já constantes em seus bancos de dados. Todavia, em verdade, cuida-se dos mesmos dados, que circularam entre instituições sem que houvesse controle algum de veracidade e que, quando de fato realizado, acabou por se lastrear em falsa premissa.

Como se não bastasse, a ausência de mecanismos de controle de veracidade revelou, a partir das pesquisas efetuadas, que atualmente não há qualquer mecanismo de aferição do balanceamento das informações e dados obtidos pelas instituições de combate à corrupção e lavagem de capitais. Esse desbalanceamento potencializa o estabelecimento de vieses interpretativos equivocados e resultados, por conseguinte, pífios ou inexpressivos em se tratando de *big data* e inteligência artificial.

A pesquisa também revelou a ausência de critérios similares entre as instituições no que atine à catalogação dos dados eventualmente já estruturados. Aludidas diferenças justificam-se até certo ponto em razão dos diferentes fins institucionais, a fomentar critérios de etiquetagem diversos a partir da respectiva funcionalidade esperada. Contudo, a pesquisa apontou que instituições com atividades finalísticas absolutamente muito próximas ou similares adotam critérios, nesta seara, absolutamente díspares. Como consequência, embora diante da mesma situação ou elemento fático, acabam por atribuir-lhes dissonantes escalas de importância e abrangência, o que dificulta, encarece e até mesmo inviabiliza a coordenação de políticas interinstitucionais.

Esse detalhe chama a atenção na medida em que a maioria das instituições afirma ter implementado políticas de *big data* e inteligência artificial, a revelar considerável empenho de recursos humanos e financeiros. Entretanto, não se verificam interoperabilidades interinstitucionais, sendo comum observarem-se estruturas replicadas em vários órgãos, ainda que arquitetonicamente diversas, para análise de um mesmo dado ou elemento. Não se cuida apenas da replicação de trabalhos ou custos (elevados, frise-se), mas principalmente da inviabilidade de aproveitamento destas políticas, coordenadamente, entre diferentes instituições – panorama cuja modificação afigura-se de urgência inquestionável.

Por outro lado, o diagnóstico também retrata a preocupação das instituições em que os dados coletados, estruturados e catalogados possam ser lidos por meio da utilização de softwares com licenciamento gratuito. Há de se reconhecer que, apesar deste intuito, a avançada natureza de determinado dado coletado muitas vezes exige a aplicação de software privado sem similar no mercado, quanto mais de licenciamento gratuito. Contudo, ainda que possível a utilização de softwares gratuitos, a interoperabilidade das políticas de *big data* e inteligência artificial ainda se apresenta por demais incipiente.

Destaque-se, outrossim, que muitas instituições apontaram que suas políticas de *big data* e inteligência artificial não estariam alinhadas aos objetivos específicos das atribuições institucionais. Com o ímpeto de alinhar políticas institucionais ao desenvolvimento tecnológico – mormente quando a sociedade patenteia a equivocada impressão de que *big data* e inteligência artificial representariam a solução de boa parte dos problemas institucionais -, acabam por implementá-las de forma açodada, sem o devido planejamento ou, quando muito, tendo em consideração muito mais atividades-meio do que atividades-fim institucionais. Além de olvidarem que muitos problemas não demandariam *big data* ou inteligência artificial como caminhos de solução, o equívoco cometido acaba por demais burocratizar aquilo que já é simples, desacreditando culturas de política de dados e dificultando, ainda mais, a correção dos vícios identificados.

Também preocupante a informação trazida pelo diagnóstico de que a grande maioria das instituições registra escassez de habilidades em *big data* e inteligência artificial. Embora o diagnóstico não deixe isso claro, aludida escassez pode, de fato, ser devida à falta de cultura institucional quanto às políticas em apreço, à ausência de centro de divulgação de conhecimento e, por fim, ao eventual desalinhamento observado em relação às atividades-fim das instituições. Em outras palavras, a escassez seria mais operacional do que de implantação propriamente dita. Esta premissa pode ser extraída a partir da confirmação, por boa parte das instituições, quanto à ausência de centros de excelência em *big data* e inteligência artificial, muito embora asseverem, por outro lado, possuírem em seus quadros profissionais especializados em estatística, álgebra, ciência da informação, tecnologia da informação etc.

Coleta, estruturação e catalogação dos dados exigem, por sua vez, a existência de infraestruturas capazes de armazená-los e, acima de tudo, de manter a higidez de todos os sistemas, notadamente por meio de rotinas para preservação da segurança dos dados. Neste aspecto, a pesquisa levada a efeito revelou que grande parte das instituições adota múltiplas rotinas para preservar a segurança dos dados. Dentre elas destacam-se a implementação de acessos controlados, trilhas de auditoria, criptografia de dados, mecanismos de redundância e políticas claras de segurança da informação. Porém, de outra parte, a armazenagem dos dados ainda se concentra principalmente em meios físicos, suscetíveis a maior degradação e com custo de manutenção elevado. Ainda que de forma tímida, nota-se o esforço das instituições, a partir dos dados coletados, na busca do aumento significativo de armazenamento de dados em nuvem ou, de modo pulverizado, em diversos equipamentos.

A par desse esforço, constatou-se que mais da metade das instituições que participaram da pesquisa utilizam algoritmos para analisar os dados coletados. No que diz respeito ao uso da inteligência artificial, assim como já observado em outros aspectos da pesquisa, muitas instituições criaram estruturas próprias visando à análise dos mesmos dados e tendo os mesmos objetivos em pauta, implicando elevados custos diante da replicação de estruturas idênticas que seriam dispensáveis caso houvesse diretrizes básicas estabelecidas ou mesmo coordenação interinstitucional nesse sentido. Em compensação, pontualmente identificou-se que algumas instituições já evitam a duplicidade de estruturas por meio do compartilhamento com instituições afins, ainda que de natureza diversa, implicando índices de resultados bastante satisfatórios.

Quanto à natureza da política de inteligência artificial implementada nas instituições, a tônica pauta-se, atualmente, pela diversidade. Normalmente os algoritmos são desenvolvidos pelas próprias instituições que os utilizam, havendo raríssimos casos de compartilhamento destas plataformas. Quase em sua totalidade os algoritmos são supervisionados, embora não existam padrões que permitam identificar mensuração ou controle dos resultados obtidos em larga escala. Os resultados, em verdade, acabam sendo valorados em situações casuisticamente postas sob análise, mas sem levar em consideração aspectos globais ou interinstitucionais. Há, por sua vez, interesse na habilitação de algoritmos não supervisionados quanto a aspectos informacionais bastante específicos, porém não levados a efeito por conta não apenas de restrições legais, mas também em razão da necessária consolidação destas políticas no ambiente institucional, incompatível com algoritmos *blackbox*.

Por fim o diagnóstico revela o esforço das instituições na implantação de políticas de *big data* e inteligência artificial e a preocupação constante em respeitar as restrições legais advindas da política regulatória de acesso e divulgação de dados no Brasil, como a Lei Geral de Proteção a Dados e a tendência cada vez mais restritiva ora identificada. Disso exsurge, com maior tenacidade, a urgência no estabelecimento de políticas aptas a incrementar tanto o compartilhamento de dados quanto a interoperabilidade de sistemas, ainda que pelo uso de ferramentas de inteligência artificial, com o escopo de redução de custos e maximização de resultados, continuamente aferíveis por todas as instituições que participarem de sua respectiva construção.

Este o sentido do presente guia de boas práticas.

3. BOAS PRÁTICAS

Boas práticas devem ser entendidas não como algo impositivo ou inquestionável. Ao contrário, pretendem melhor orientar o gestor público no processo de tomada de decisões. Têm por objetivo a realização de projetos de forma mais célere e eficiente, com redução de custos, maximização dos resultados possíveis e plasticidade suficiente a permitir sua contínua evolução diante do constante dinamismo social.

Dinamismo ainda mais notório quando se trata de políticas de *big data* e inteligência artificial. Como se não bastassem as contínuas e profundas alterações da sociedade, a própria tecnologia registra evoluções e mudanças ainda mais rápidas e surpreendentes. A arquitetura escolhida para implantação destas políticas deve ser capaz de suportar todas estas mudanças, por considerável período, sob pena de restarem absolutamente inviabilizadas. Por demandarem elevando custo de investimento (financeiro e humano) e impactarem diretamente a consolidação institucional desta cultura (política de resultados), imprescindível que seus alicerces suportem diversas configurações ao longo do tempo.

As maiores críticas às políticas de *big data* e inteligência artificial, em verdade, não se devem à natureza delas, mas à frustração decorrente das altas expectativas em suas implantações. Para afastar ou minimizar este desagrado, necessário conhecer no que serão aplicadas, como serão e, acima de tudo, o que se pode esperar a partir disso. Neste contexto a troca de experiências interinstitucionais resta fundamental: não que se vá replicar ou repetir o já realizado por outro ente - até porque as realidades de cada instituição divergem e nem sempre aquilo que seja proveitoso a uma delas será em relação às demais -, mas principalmente entender quais foram os acertos e erros identificados, e como foram superados.

3.1. Conhecer o âmbito de aplicação

Muito se têm falado a respeito de b*ig data* e inteligência artificial, muitas vezes como solução para os mais diversos e variados problemas. No entanto, b*ig data* e inteligência artificial nem sempre serão a solução. Nem tudo pode ser por eles tratado; ao contrário, há inúmeros contextos em que, além de não se aplicarem, insistir nesta perspectiva pode provocar riscos sérios de danos irreparáveis, que acabariam por prejudicar sua credibilidade.

O uso da ferramenta incorreta não a torna imprestável, mas apenas retrata, em suma, a incompetência de seu usuário. Não há como extrair, por exemplo, parafusos sextavados utilizando-se de uma chave de fenda: a ineficiência não é da ferramenta, mas do seu operador, que não soube distinguir qual seria a aplicável naquela situação. E há ainda quem negue, sem fundamento, que o parafuso sextavado seja, de fato, um parafuso – tudo na tentativa de justificar a própria incompetência.

Portanto, antes de conhecer o que de fato vêm a ser *big data* e inteligência artificial, necessário que o gestor compreenda quais seus verdadeiros âmbitos de aplicação. Em breves linhas, pode-se afirmar serem aplicáveis quando identificada a necessidade de coleta, estruturação, armazenamento e catalogação contínua de elevado volume de dados, a exigir rápida e simultânea interpretação para fundamentar a adoção de medidas

ou atividades imediatas, sob pena de não atingirem as finalidades esperadas. O sincronismo constante entre o recebimento de elevado volume de dados e a resposta institucional emitida, para que surja efeitos, deve se dar no menor tempo possível, quase em concomitância.

Diante do contínuo volume de informações recebidas, eventual demora no agir institucional pode ensejar comportamentos ou atitudes que, quando implementados, já não se apliquem à realidade social naquele momento verificável. Isso, além de acarretar a inutilidade dos esforços envidados, agride a imagem da própria política de *big data* e inteligência artificial adotada. No entanto, o erro é do gestor público que, assim como o ferramenteiro que insistiu na utilização da chave de fenda para retirada de parafuso sextavado, não soube utilizar corretamente o ferramental posto à sua disposição.

O raciocínio descrito afigura-se ainda mais apropriado no âmbito dos órgãos e instituições de persecução e controle no combate à corrupção e lavagem de capitais. As organizações voltadas à prática de atos ilícitos modificam seu modus operandi de acordo, dentre outros aspectos, com a evolução das políticas utilizadas pelas instituições para combatê-las. A fotografia do ilícito, por assim dizer, obtida pelas instituições de combate à corrupção e lavagem de capitais, deve ser temporalmente o mais próxima possível do momento em que deflagrada a atuação estatal de combate. E como o ilícito é dinâmico, característica essencial para sua permanência, os órgãos de persecução e controle precisam reunir o maior volume de dados no menor tempo possível, tratá-los, catalogá-los, interpretá-los e, imediatamente, agir para buscar desestruturar as atividades ilícitas.

O dinamismo, o extenso volume de dados, o necessário tratamento e interpretação das informações, aliados à necessidade de rápida resposta por parte das instituições, sintetiza, ainda que de maneira singela, o âmbito de aplicação das políticas de *big data* e inteligência artificial no combate à corrupção e lavagem de capitais.

3.2. Conhecer o problema

Conhecido o âmbito de aplicação, incumbe ao gestor afastar qualquer projeto de aplicação de políticas de *big data* e inteligência artificial fora deste contexto. Caso esteja no âmbito de aplicação, ainda assim, imprescindível que o gestor, antes de buscar soluções ao problema proposto, procure conhecê-lo da maneira mais profunda e completa possível.

Qualquer solução, de que nível for, inicia-se a partir do entendimento do problema a que se propõe extinguir. A má compreensão do problema, por sua vez, pode implicar a adoção de soluções que não atendam às expectativas formuladas ou, até mesmo, expressamente a elas contrárias. E, neste caso, mais uma vez se tentará imputar a responsabilidade à própria natureza das políticas de *big data* e inteligência artificial quando, em verdade, cuida-se de equívoco na identificação do problema.

Aproveitando-se da comparação já utilizada anteriormente, o operador da ferramenta precisa saber diferenciar um parafuso sextavado (a exigir a aplicação de ferramental próprio para extração) de outro autoatarraxante, por exemplo. Só a partir desta compreensão é que poderá reconhecer, dentre as ferramentas disponíveis, qual delas se afigura aplicável. A má compreensão do gestor, muitas vezes traído pela experiên-

cia, consiste justamente em atribuir a mesma definição a problemas de naturezas distintas, utilizando-se de ferramentas incompatíveis e deixando de obter os resultados esperados.

Em se tratando de políticas de *big data* e inteligência artificial o problema torna-se ainda mais sério porque não se cuida de uma mera ferramenta. Em verdade implica a estruturação de arquitetura administrativa de alto custo, com elevados investimentos financeiros e humanos, muitas vezes inservível a outros fins que não os inicialmente previstos - normalmente quando não seguidas as boas práticas constantes do presente guia. Ademais, exige o desenvolvimento de cultura institucional para que seus integrantes a utilizem de fato, constatem os resultados obtidos. Caso as instituições estruturem estas políticas sem conhecer, a fundo, o problema a que serão aplicadas, o risco da obtenção de resultados insignificantes ou reprováveis afigura-se altíssimo (não conseguirá extrair o parafuso sextavado com a chave de fenda escolhida, e.g.), denegrindo indevidamente a credibilidade das políticas de *big data* e inteligência artificial no âmbito institucional e prejudicando, sem dúvida alguma, novos investimentos.

Portanto, antes de se buscar a melhor solução, de rigor conhecer em profundidade o próprio problema.

3.3. Primar pela interação humana na construção do modelo classificador

Concluindo-se pelo cabimento da aplicação de políticas de *big data* e inteligência artificial e conhecido o problema que se deseja solucionar, importante pensar a construção de modelo classificador que permita pô-las em prática, ou seja, coletar, estruturar, catalogar, armazenar e interpretar dados, atribuindo-lhes valor social suficiente a permitir a adoção de atitudes concretas e imediatas no combate à corrupção e lavagem de capitais.

A construção do modelo classificador representa uma das estruturas mais demoradas e custosas no âmbito de aplicação das políticas em apreço. Ao mesmo tempo, alterá-lo posteriormente pode se tornar inviável ou altamente custoso. Por fim, ignorar erros cometidos na sua escolha pode prejudicar, de forma irreversível, a credibilidade de qualquer instituição.

Ainda que a escolha do modelo classificador, dentre as hipóteses possíveis, tenha sido a mais acertada em determinado momento, o dinamismo das relações sociais pode torná-lo obsoleto em curto intervalo de tempo. E como se mostra altamente custosa a criação de qualquer modelo classificador, dificilmente instituições ou gestores públicos estarão dispostos a investir, novamente, na criação de outro modelo que possa se tornar inservível a qualquer momento.

Infelizmente, ao tratar de políticas de *big data* e inteligência artificial, muitas instituições acabam por dar preferência a critérios de automação quase que independentes, desde a coleta até a própria interpretação de dados, conferindo à interação humana apenas o caráter de supervisão. Em geral estes modelos classificadores possuem custos, ainda que elevados, menores se comparados a outros, e ainda por cima podem ser desenvolvidos de forma mais célere. No entanto, a preferência pelos critérios de automação em detrimento de considerável parcela de intervenção humana torna-os menos suscetíveis a se adaptarem a novas realidades. Em outras palavras, apresentam resultados rápidos, porém registram vida curta.

Para que o modelo classificador possa se adaptar rapidamente às modificações socialmente identificadas, imprescindível que a interação humana esteja presente em cada uma das fases de sua construção. Não basta, neste aspecto, que o ser humano exerça a supervisão das estruturas digitais criadas; a atividade humana precisa, necessariamente, compor cada um dos pilares destas estruturas, sejam elas relacionadas a *big data* ou inteligência artificial. Como num tecido complexo, na medida em que as alterações da sociedade impactarem o modelo classificador, compondo a ação humana cada um dos pilares de sua estrutura juntamente com a tecnologia, a percepção de mudanças ou variáveis poderá ser melhor e mais rapidamente identificada, permitindo a realização dos ajustes no tempo e modo necessários à sua manutenção como modelo válido de coleta, estruturação, catalogação, armazenamento, análise e interpretação de dados.

O grande e imperdoável erro das instituições em políticas de *big data* e inteligência artificial reside justamente em supervisionar aludidas políticas com base exclusivamente na aferição de resultados. Em outras palavras, delega-se à tecnologia praticamente toda a construção do modelo classificador, de sorte que, quando os resultados deixam de ser os esperados, não se sabe como, nem o que, poderia ser modificado para alinhá-lo à nova realidade. Isso, no mais das vezes, acarreta o abandono do modelo classificador e início de construção de um novo, com elevados custos e nem sempre atingindo aquilo que se espera.

Há ainda a errônea impressão, diga-se de passagem, de que o acerto do modelo classificador para uma situação singular permita seja aplicado a outras similares. Essa circunstância fica bastante evidente quando a instituição apura a eficiência de políticas de *big data* e inteligência artificial de forma casuística, em um ou outro caso concreto. Não se pode pensar a construção de modelo classificador visando a uma situação ou caso concreto específico, até porque contrária à própria natureza das políticas de *big data* e inteligência artificial, atrelada a massivos volumes de dados. No entanto, como a estruturação se deu com base na dependência de parâmetros majoritariamente tecnológicos, cabendo à interação humana exclusivamente a supervisão dos resultados, sendo estes positivos (ainda que casuisticamente), supõe-se seja um verdadeiro sucesso. Porém, basta uma pequena modificação da realidade social (às vezes ínfima), e o modelo classificador torna-se obsoleto e imprestável.

Somente com a premissa de que a interação humana deva estar presente em cada uma das fases de estruturação do modelo classificador pode-se lhe atribuir a plasticidade necessária para adaptar-se, constantemente, ao dinamismo social identificado e continuar, de forma sólida (o paradoxo aqui não é mera figura de estilo), a entregar resultados com o rigor e eficiência buscados.

A interação humana não pode se resumir, portanto, à supervisão das políticas de *big data* e inteligência artificial: precisa integrá-las, de fato.

3.4. Zelar pelo balanceamento de dados

Como decorrência lógica das boas práticas anteriores encontra-se o inafastável controle do balanceamento dos dados obtidos. Ignorar esta necessidade pode gerar a criação de perspectivas contrárias à realidade, implicando falsos resultados capazes de levar o gestor público ou mesmo instituições a tomarem decisões, no mínimo, equivocadas; às vezes, catastróficas.

Desde que a interação humana integre as políticas de *big data* e inteligência artificial, em cada uma das fases de estruturação do modelo classificador, mais praticável será a aferição do balanceamento dos dados. Muitas instituições acabam tendo destaque na formação do volume de dados por circunstâncias diversas, não necessariamente vinculadas à natureza dos dados obtidos. Citam-se os exemplos daquelas que já recebem dados de forma estruturada, a facilitar a respectiva disseminação; outras com atribuições institucionais necessariamente atreladas ao fluxo de grande volume de dados. Inegável, assim, que a formação dos bancos de dados em políticas de *big data* e suas interpretações relacionem-se, diretamente, à ótica destas instituições a respeito dos problemas sociais identificados, dentre eles, por óbvio, o combate a atos de corrupção e lavagem de capitais.

Contudo, nenhuma das instituições, por mais abrangentes que possam ser suas atribuições, consegue extrair completa descrição de determinado problema ou fato sociais: ao contrário, seus dados e declarações correspondem a uma parcial visão, ainda que não admitida expressamente, permeada pela tendência de suas atribuições institucionais. Cita-se o caso, por exemplo, da comum definição de *hot spots* ("áreas quentes"), pelas secretarias de segurança pública de todos os entes da federação, a partir da análise de boletins de ocorrência formalizados. Apesar da validade da iniciativa, mostra-se absolutamente equivocada se ignorar, por exemplo, que em áreas dominadas por organizações criminosas dificilmente haverá a comunicação, pela comunidade, de práticas ilícitas, seja pelo temor de represálias, seja pela existência de verdadeiro estado paralelo na comunidade, a estabelecer leis e mecanismos de cumprimento próprios. A leitura automática dos dados, desprovida de critérios de balanceamento respectivos, pode levar a efeito a afirmação de que determinada área se afigura segura quando, de fato, cuida-se absolutamente do contrário, uma vez que dominada por nefastas organizações criminosas.

Fácil perceber, portanto, a necessidade do contínuo controle de balanceamento dos dados obtidos, notadamente pela imprescindível atuação de profissionais vinculados à disciplina ciência de dados e informação. No exemplo narrado, a própria credibilidade das políticas de *big data* e inteligência artificial restaria fortemente fragilizada (o paradoxo é proposital) não por conta de sua estrutura ou operação, mas pela reprovável falha de seus instituidores ao descuidarem do balanceamento dos dados obtidos. Ainda quanto ao exemplo apresentado, caso houvesse a interação humana na obtenção destes dados, a observância de altas taxas de evasão escolar (denotando o recrutamento de jovens por organizações criminosas), de atendimentos médicos no sistema único de saúde relacionados a atos de violência (incompatíveis com o número de boletins de ocorrência registrados) ou mesmo a discrepância da "calmaria" identificada em relação a outras localidades próximas (a evidenciar o controle de regiões por organização criminosas), além da experiência acumulada dos órgãos de persecução e controle, poderiam, em conjunto, identificar o desbalanceamento eventual de dados e ajustar o modelo classificador para levar em consideração estas circunstâncias no momento em que se proceder à catalogação e interpretação das informações obtidas.

Muito embora o objetivo deste guia de boas práticas seja, justamente, evitar casuísmos (até porque seu escopo relaciona-se ao maior espectro de aplicação possível), impossível tratar do balanceamento de dados sem a imersão em exemplos concretos. Bastante claro, portanto, que o desbalanceamento de dados, além de comprometer a correta interpretação das realidades sociais, pode ocasionar o total fracasso das políticas

de b*ig data* e inteligência artificial, inviabilizando, inclusive, novos investimentos.

No entanto, como as instituições abrigam parciais visões da realidade social, o balanceamento de dados somente será possível, em princípio, a partir da maior interoperabilidade interinstitucional de sistemas e dados e desde que seguida, principalmente a boa prática a seguir elencada.

3.5. Controlar a veracidade dos dados

Não basta que os dados obtidos sejam balanceados: é preciso submetê-los a constantes e contínuos processos de aferição de veracidade. Balanceamento e veracidade são boas práticas de natureza associada; em outras palavras, o controle de balanceamento dos dados permite estabelecer mecanismos de aferição da veracidade do mesmo modo que mecanismos de aferição de veracidade permitem a construção de critérios de análise de balanceamento dos dados coletados. Por mais circular que possa parecer o raciocínio, o pensamento em voga ajusta-se exatamente a essa premissa: círculos contínuos não possuem início.

A melhor implantação das políticas de *big data* e inteligência artificial têm, essencialmente por natureza, a preocupação quanto ao compartilhamento de informações e interoperabilidade de sistemas. Como já aduzido em tópicos anteriores, há instituições que, por circunstâncias diversas, coletam por natureza grande volume de dados que acabam sendo repassados ou replicados a diversas outras instituições. Certamente, em algum momento posterior, as mesmas informações acabam retornando à instituição que inicialmente as inseriu no circuito de dados. Desconhecendo esta circunstância, diante da equivalência das informações, a própria instituição de origem acaba por validar a informação original – ausente, porém, qualquer procedimento de aferição real dos dados.

Há ainda possíveis situações em que, dada a credibilidade conquistada por determinada instituição no seio social, tende-se a aceitar naturalmente os dados ou informações por ela transmitidos como verdadeiros, seja porque se supõe (frise-se, apenas uma suposição) terem sido submetidos a algum controle de veracidade (que, em verdade, não houve), seja porque eventual falsidade do dado seria absolutamente incompatível com a imagem que aquela instituição possui perante seus pares.

Com isso quer-se afirmar que o controle de veracidade, além de essencial, deve ocorrer em dois níveis distintos: o primeiro, interno, isto é, a própria instituição deve prever, ao implantar políticas de *big data* e inteligência artificial, critérios claros e contínuos de aferição da veracidade dos dados coletados; o segundo, externo, de sorte que a instituição remetente dos dados deixe claro se existem, ou não, políticas de aferição de veracidade implantadas.

O controle de veracidade dos dados, em se tratando de boas práticas em *big data* e inteligência artificial, jamais pode se balizar em suposições. Urge que, tanto no recebimento quanto na transmissão das informações, a contínua política de controle de veracidade seja constante (o pleonasmo não é mera figura de estilo, posto que pode ser contínua quanto aos dados coletados, porém inconstante quanto à manutenção deste controle sobre os mesmos dados) e transparente.

Uma vez que mecanismos de controle de veracidade dos dados ainda é algo absolutamente casuístico na realidade nacional (em verdade, o controle de veracidade em muitas instituições dá-se pontualmente, e não de forma estatística ou com base em ciência de dados), caso a instituição consiga implementá-los como boa prática já representará enorme avanço no que atine a *big data* e inteligência artificial. O ideal, no entanto, é que a aferição da veracidade de dados aconteça de forma replicada, ainda que sazonalmente ou por amostragem, ainda que abarcando dados já anteriormente analisados, uma vez que, diante do dinamismo social, no qual se inserem as constantes mutações dos atos de corrupção e lavagem de capitais, o conhecimento de outros dados, até então ignorados por órgãos de persecução e controle, pode seguramente invalidar critérios que fundamentaram conclusões ou resultados anteriores.

Como boa prática, assim, o controle de veracidade jamais pode se basear na confiança entre instituições, mas em critérios matemáticos e estatísticos cientificamente demonstráveis.

3.6. Utilizar vocabulário controlado

Conhecidos o âmbito de aplicação das políticas de *big data* e inteligência artificial, o problema que se busca tratar, integrada a interação humana ao modelo classificador escolhido, instituídos mecanismos de controle de balanceamento e veracidade dos dados, necessário estruturá-los de forma a permitir sejam agrupados, catalogados e devidamente armazenados.

Contudo, a estruturação não terá utilidade alguma se a instituição ignorar a necessidade de, previamente, estabelecer regras claras de ortossemântica para esta finalidade. Seria o mesmo que continuamente admitir o recebimento de livros em uma biblioteca sem critérios claros de identificação, catalogação e agrupamento; pior, equivaleria a recebê-los continuamente sem que se pudesse sequer entender do que tratam ou identificar em qual idioma redigidos.

Importante reconhecer que a estruturação de dados, apesar do grande avanço e sucesso obtido por diversas instituições, é um dos grandes e, ao lado da definição do modelo classificador, mais custosos pilares de construção de políticas de *big data* e inteligência artificial. Por isso, mais uma vez, o profundo conhecimento do problema que se pretende abordar é importantíssimo para que as instituições, dentre os dados coletados, elejam quais informações prioritariamente devem ser extraídas. Definido este aspecto, quanto a este conjunto de informações, a instituição precisa estabelecer regras ortossemânticas claras que permitam sejam corretamente identificadas.

Cumpre esclarecer neste ponto que, embora sob o mesmo signo linguístico, termos e dados podem possuir semânticas absolutamente distintas. “Machado”, por exemplo, é termo que pode se referir tanto a instrumento de corte como a ilustre escritor brasileiro. Erros na catalogação deste termo, por divergências semânticas, podem causar danos invencíveis tanto no balanceamento dos dados obtidos quanto na aferição da respectiva veracidade do conjunto informacional.

Portanto, antes de iniciar a estruturação dos dados propriamente, as instituições necessitam definir qual informação consideram de extração prioritária e, a partir deste universo, criar regras de ortossemântica gráfica

e vocabular que propiciem a estruturação de dados propriamente dita. Esta providência permitirá não apenas imprimir maior velocidade à implantação das políticas de *big data* e inteligência artificial como também identificar, o quanto antes, a existência de dados incompletos que, caso não retificados, ensejarão fatores críticos de desbalanceamento de dados e prejuízos ao controle de veracidade necessário.

3.7. Estruturar dados

Devem as instituições públicas e seus respectivos gestores aterem-se à forma como os dados recebidos serão estruturados. Não se ignora que muitas instituições, seja pela própria natureza de suas atribuições finalísticas, seja pelo avanço no uso de tecnologias incomuns à realidade brasileira, já possuem alto grau de estruturação dos dados recebidos – em algumas situações em sua totalidade, inclusive.

No entanto, a grande maioria das instituições possuem grandes acervos físicos documentais ou, ainda que recebam dados em formato digital, não desenvolvem qualquer política de estruturação destes dados. Nos casos em que identificados acervos físicos não tratados, incumbe ao gestor público imprimir o máximo esforço para que a estruturação, em formatos digitais, seja contínua e constantemente buscada, sendo imperiosa a fixação de metas realmente possíveis e observância regular deste processo. Já naqueles casos em que instituições recebem dados em formatos digitais têm-se a falsa impressão de que, por este simples fato, estariam estruturados, o que representa gritante equívoco.

Como visto, a estruturação de dados envolve muito mais do que a digitalização de acervos físicos, apesar de ser um passo importante e considerável neste assunto. Além dos formatos digitais, é preciso que a estruturação permita catalogar as informações prioritárias rápida e uniformemente, dado o contínuo e volumoso fluxo de dados inerentes às políticas de *big data* e inteligência artificial. Quanto mais estruturado estiver o acervo de dados de uma instituição, não apenas mais célere e eficaz será o desenvolvimento de mecanismos de inteligência artificial, mas certamente menos custoso em termos financeiros e orçamentários.

Qualquer uma das boas práticas descritas no presente guia não deve ser entendida como fase estanque e isolada do processo de implantação de políticas de *big data* e inteligência artificial. Para melhor aproveitamento destas políticas é essencial que sejam monitoradas e desenvolvidas de forma associada. Este apontamento faz-se necessário uma vez que, definida a forma de estruturação de dados, instituições e gestores passam a concentrar seus esforços em outras áreas, entendendo por encerrado referido assunto. Todavia, além do dinamismo social, que pode tornar obsoletos padrões de estruturação de forma repentina e imprevista, deve-se lembrar que, em se tratando de políticas de *big data* e inteligência artificial, da fase de coleta dos dados à de concreta atuação das instituições com base na interpretação destes elementos, deve-se observar o mínimo espaço de tempo possível, objetivando-se intervenções quase que concomitantes à própria obtenção dos dados (*streaming act*). Por esta razão, essencial a contínua intervenção humana na integração e supervisão das políticas de estruturação, a avaliar se os critérios definidos em momento anterior permanecem adequados frente à realidade.

3.8. Utilizar softwares de licenciamento gratuito

Não basta estruturar corretamente dados: para que possam revelar o valor social agregado é preciso que consigam ser, efetivamente, lidos e interpretados do modo mais fluido possível. A ferramenta utilizada para isso, na maior parte dos casos, reside justamente na escolha do software correto para que a interpretação seja viável.

Veja-se o seguinte exemplo: um arquivo gráfico pode ser aberto em editores de texto, ainda que muito rudimentares; no entanto, a visualização de seu conteúdo ocorrerá em linguagem de programação (muitas vezes incompleta, aliás, pela própria limitação do software em abrir arquivos distintos da natureza para a qual pensado e criado) ou em códigos binários. A imagem gráfica, em si, jamais poderá ser vista diretamente por meio da utilização de um editor de textos. Para a correta compreensão do arquivo gráfico, necessário aplicar softwares apropriados de leitura.

Mesmo quando utilizados softwares apropriados para abertura e leitura de dados, a pluralidade de opções existente no mercado nacional - referentes a diversas versões do mesmo produto, inclusive -, pode implicar incompatibilidades e leituras parciais dos elementos contidos no material examinado, circunstância custosa e altamente nociva ao desenvolvimento de políticas de *big data* e inteligência artificial. E não se cuida apenas de situações em que se busca a leitura de documento produzido em software licenciado por meio de outro gratuito: o inverso também ocorre, muitas vezes, com perda considerável de dados.

De qualquer forma, impossível (e até inviável) que instituições ou gestores públicos utilizem softwares únicos para análise de dados de acordo com suas respectivas naturezas. Inegável que, diante do considerável volume de dados e da celeridade obrigatória de análise por mecanismos de inteligência artificial, em muitos casos não haverá software de licenciamento gratuito que possa se desincumbir desta tarefa. Há situações que demandam a criação de softwares, inclusive, específicos para a leitura de dados no contexto do combate à corrupção e lavagem de capitais. Isso sem mencionar os custos de aquisição de licenças, absolutamente variáveis e dependentes, sem dúvida, da capacidade orçamentária de cada instituição ou órgão.

Portanto a boa prática em epígrafe objetiva, primordialmente, suavizar ou mesmo neutralizar estes obstáculos. Sempre que houver software gratuito de análise de dados e desde que dotado de fatores que garantam a segurança da informação, instituições e gestores públicos a ele devem dar preferência. Caso isso não se mostre possível, havendo pluralidade de opções de softwares não gratuitos, deve-se optar pela aquisição de licenças daqueles que maior compatibilidade revelarem com softwares de licenciamento gratuito, ainda que de forma parcial. Somente quando superadas estas circunstâncias se afigura justificável a aquisição ou desenvolvimento de softwares de licenciamento remunerado por instituições ou gestores públicos no âmbito do combate a atos de corrupção e lavagem de capitais.

Qualquer que seja a opção escolhida pela instituição ou gestor público, dada a sensibilidade do tema combate à corrupção e lavagem de capitais, gratuito ou remunerado, o software escolhido necessariamente deverá apresentar características que mantenham, simultaneamente, a segurança e sigilo dos dados obtidos. Não se pode descartar a hipótese de inserção dolosa, por parte de organizações voltadas a práticas ilícitas, de softwares no mercado nacional (ou mesmo internacional), ainda que aparentemente bem intencionados, mas

que podem ter objetivo, de fato, facilitar o monitoramento dos dados coletados pelos órgãos de persecução e controle e identificar estratégias eventualmente adotadas.

3.9. Armazenar corretamente os dados coletados

Pouca ou nenhuma valia terá a coleta e leitura de dados sem que exista ambiente próprio para que sejam armazenados e, quando necessário, pronta e rapidamente localizados. Há diversas formas de armazenamento possíveis atualmente, porém cada uma delas registra características tanto favoráveis como desfavoráveis. A escolha deve se vincular aos fins institucionais perseguidos e à natureza dos dados coletados.

Sem desconsiderar as infinitas variáveis possíveis, os sistemas de armazenamento de dados atualmente utilizados em sua maioria pelas instituições de persecução e controle correspondem aos meios magnéticos (conectam-se aos equipamentos por meio de drivers, como discos rígidos), ópticos (CD, DVD e Blu-ray), magneto-óptico (mídias portáveis duráveis com grande capacidade de armazenamento), eletrônico (memórias em estado sólido – SSD), em rede (Network Attached Storage – NAS, servidores dedicados conectados a uma rede que compartilham dados com vários usuários) e em nuvem (cloud computing).

Em princípio pode-se estimar que os meios magnéticos, ópticos e magneto-ópticos, para fins de implantação, sejam os menos onerosos em termos financeiros. No entanto, com o passar do tempo, em vista do grande volume de dados armazenado e da constante necessidade de ampliação destes meios, não só os custos passam a ser proibitivos: às vezes o próprio espaço físico disponível para alocação destes recursos torna-se insuficiente. Estas circunstâncias os tornam, a longo prazo, mais onerosos em termos financeiros se comparados às demais formas de armazenamento.

Também apresentam grande vulnerabilidade para perdas ou violações os meios magnéticos, ópticos ou magneto-ópticos. Em razão da natureza física que ostentam, sujeitam-se mais facilmente às intempéries do tempo e a atos criminosos. Ainda que presentes mecanismos de replicação com o escopo de preservar a higidez dos dados coletados, não há de se descartar a possibilidade de atuações orquestradas em larga escala, por organizações voltadas a práticas ilícitas, com o objetivo de eliminação física destes dados, ou mesmo desbalanceamento daquilo que já fora coletado, mediante a implantação de falsos dados.

Há quem defenda, outrossim, que a existência de barreiras de segurança retiraria dos aludidos meios a vulnerabilidade em pauta. Apesar da adequação deste pensamento, geralmente os custos voltados à manutenção da integridade e segurança dos dados armazenados nos meios magnéticos, ópticos ou magneto-ópticos pode facilmente superar os custos estimados para a adoção de outros meios de armazenamento mais eficazes, como eletrônico, em rede ou em nuvem.

Os benefícios do armazenamento em rede são muitos, dentre eles destacam-se a facilidade de adoção, baixos custos de implantação, escalabilidade horizontal (não há necessidade de aquisição de mais discos rígidos, bastando solicitar mais recursos sem que seja preciso desligar a rede), e a possibilidade da adoção de diversos mecanismos de segurança, tanto de alimentação quanto de acesso aos dados. Ademais, o armazenamento em rede permite maior agilidade das instituições de persecução e controle na coleta e análise

de dados, posto que estarão disponíveis em diversificados e variáveis pontos de acesso, permitindo acesso independentemente do tipo de equipamento casuisticamente utilizado ou do software nele eventualmente instalado.

Embora o armazenamento eletrônico possa parecer tão vulnerável quanto os meios magnéticos e ópticos, a natureza da tecnologia utilizada no seu desenvolvimento dificulta a prática de atos de invasão ou vandalismo cibernético, notadamente por permitir o uso de variados graus de segurança de acesso e pela maior velocidade de processamento, propiciando elevada vigilância e monitoramento dos centros de dados, facilmente replicáveis, inclusive, em rede (armazenamento eletrônico e em rede possuem alto grau de compatibilidade, permitindo replicação das informações e duplicidade de critérios de autopreservação logística destes dados).

O armazenamento em nuvem, por outro lado, apesar da alta capacidade de armazenamento e do pouco espaço físico demandado das instituições que por ele optarem, traz alto risco quanto ao sigilo e segurança dos dados armazenados. Ainda que se afigurem sólidos e confiáveis os mecanismos de segurança postos à disposição da instituição que o utilizar, a responsabilidade pelo gerenciamento destes dados normalmente é atribuída a entes privados, que, em verdade, acabam ficando na posse destes dados. Como o combate à corrupção e lavagem de capitais envolve temas sensíveis, não se mostra aconselhável, muito menos prudente, que estas informações fiquem sob a responsabilidade de entes privados, ainda que se comprometam a manter rígidos controle de integridade e segurança cibernética. Recomendável, portanto, que o armazenamento em nuvem seja desenvolvido e implantado pela própria instituição de persecução e controle, o que demanda o desembolso de maiores recursos financeiros no início dos trabalhos, se comparado às outras formas de armazenamento.

Conclui-se, por fim, que muito embora o armazenamento em nuvem seja o mais custoso às instituições, a longo prazo revela-se o mais seguro e rentável, desde que por ela criado, implementado e desenvolvido. Entretanto, de rigor que esteja atrelado a mecanismos de duplicidade, com a garantia da higidez da segurança e integridade do acervo coletado. No entanto, caso ainda se afigure proibitivo às instituições por questões financeiras e orçamentárias, aconselhável a utilização de mecanismos de armazenamento em rede, ainda que combinados com os ambientes magnético, óptico e magneto-óptico.

3.10. Estabelecer níveis de acesso aos dados coletados

Definido o meio (ou meios) de armazenamento dos dados, a higidez e segurança esperada somente poderá ser alcançada se houver a clara definição de níveis de segurança de acesso a estes mesmos dados. A boa prática aconselha que as definições sejam claras para que a transparência no trânsito e acesso aos dados permita, inclusive, a implantação de critérios de aferição da veracidade úteis e eficientes. Da mesma forma, os níveis de acesso devem ser variados para permitir a interoperabilidade das informações, ainda que entre instituições diversas, sem prejudicar o sigilo ou segurança necessários à manutenção da higidez do dado ou de sua respectiva interpretação.

Não obstante, ainda que escalonados os níveis de acesso, importante a manutenção contínua de trilhas de auditoria capazes de aferir, em tempo real ou por meio de pesquisa histórica, quais usuários acessaram de-

terminados dados. Há quem defenda, inclusive, a necessidade de o usuário justificar antecipadamente a necessidade de seu acesso a determinado dado, ainda que de forma suscinta e automática, de sorte a permitir futura auditoria. Outros, por sua vez, afirmam que podem coexistir em perfeita harmonia níveis de acesso condicionados a justificativa prévia (ainda que de forma automática) ao lado de níveis de acesso incondicionados (notadamente aos órgãos de persecução e controle voltados à elaboração de estratégias construídas a partir de informações de inteligência).

A coexistência de níveis de acesso de natureza distinta afigura-se, em princípio, a melhor prática a ser seguida, a depender, evidentemente, da natureza do órgão de persecução e controle que os utilizar. Sem ignorar a existência de profunda discussão sobre o tema, descabida no âmbito singelo do presente guia, importante definir se, de fato, as informações obtidas serão utilizadas em estratégias de imputação (instrução de procedimentos apuratórios ou de responsabilização) ou de inteligência (desvinculada dos aludidos procedimentos, objetivando definir a melhor forma de abordagem de determinado fato social, lícito ou não). Quando se fala em estratégias de imputação, diante das garantias constitucionais e da pormenorizada regulamentação infraconstitucional no acesso e proteção a dados, até para evitar nulidade dos procedimentos, aconselhável que os níveis de acesso possam identificar, além do usuário, a justificativa de uso por ele apresentada. Já em se tratando de estratégias de inteligência, aludida identificação estará muito mais atrelada aos mecanismos de controle interno do próprio órgão de persecução e controle, podendo ser dispensável em determinados níveis de acesso, desde que o órgão consiga, de alguma forma, monitorar o comportamento e atuação de seus integrantes.

Muito embora se verifique já ser prática comum em diversos órgãos de persecução e controle, aconselhável também que os níveis de acesso, ainda que rasos, sejam dotados de mecanismos que permitam aferir se o usuário que pleiteia o acesso é, de fato, quem afirma ser. Há inúmeras alternativas, como acesso em duas etapas, biometria, controle de voz, reconhecimento facial, padrões de digitação etc. Entretanto, não se pode esquecer que todos estes dados podem também ser coletados automaticamente, sem alarde algum, por mecanismos de inteligência artificial, circunstância que se afigura altamente preocupante. Preocupante porque se algum mecanismo de inteligência artificial for capaz de ludibriar sistemas de segurança relacionados a níveis de acesso, quando o órgão ou instituição se der conta deste problema certamente o volume de dados vazados terá seguramente sido bastante expressivo.

3.11. Identificar valor social nas informações

É preciso, antes de tudo, clareza institucional quanto aos objetivos ou informações que devem ser obtidas com a coleta e armazenamento de dados. A partir destes parâmetros, necessário investigar qual o valor social esperado. Informações sem valor social agregado algum, além de ocuparem o espaço daquilo eventualmente útil, muitas vezes inviabilizam o procedimento da correta tomada de decisões pela míope visão gerada da realidade social.

Reconhece-se, sem dúvida alguma, que identificar o valor social que deverá ser extraído dos dados coletados é tarefa difícil, nem sempre perceptível mesmo após reiteradas análises dos conteúdos e da convergência

dos fins institucionais. Caso a situação assim se apresente, melhor que a instituição procure, o quanto antes, setorizar as informações obtidas, catalogando-as de acordo com sua natureza ou conteúdo semântico, por exemplo. Aludida setorização, por sua vez, permitirá a contínua depuração dos dados, autorizando a rápida identificação e correção daqueles que se apresentarem incompletos ou semanticamente irregulares.

Seja para identificar o valor social das informações, seja para setorizá-las, importante que a instituição estruture bases de conhecimento ou modelos semânticos. Caso reste impossibilitada, por razões orçamentárias ou humanas, importante se faz buscar auxílio por meio de acordos de cooperação com órgãos ou instituições congêneres (ou que tenham por objeto analisar os mesmos fatos sociais em questão) para que busquem estabelecer interoperabilidades recíprocas de sorte a compartilharem bases de conhecimento ou modelos semânticos.

Importante diferenciar, neste aspecto, que o compartilhamento de bases de conhecimento ou modelos semânticos não envolve, em princípio, compartilhamento de dados entre as instituições. Compartilhamento de dados, embora absolutamente relacionado a boas práticas em políticas de big data e inteligência artificial, é tema extremamente denso cujo deslinde não constitui objeto do presente trabalho. No tópico aqui tratado, o dimensionamento reside no compartilhamento da tecnologia que permita setorizar dados, mediante estrutura semântica implementada em órgão diverso, para permitir que a instituição beneficiária possa, de fato, extrair valor social de seus bancos de dados.

A necessidade desta extração implicará melhor aproveitamento da estrutura de *big data* e inteligência artificial implantada pela instituição. Da abundância de dados obtidos, em volume e velocidade crescentes, uma pequena parcela (alguns estudos afirmam que, em média, dez por cento) é, de fato, tratada e analisada pelas instituições. O restante acaba por ocupar grande parcela do armazenamento de dados sem implicar, direta ou indiretamente, análise por mecanismos de inteligência artificial (muito menos humana, destaque-se). Com a melhor estruturação de bases de conhecimento e modelos semânticos, com a setorização adequada, estes dados "excedentes" poderão ter seu valor social mais facilmente identificado, fomentando o compartilhamento de informações entre instituições distintas com nítida redução de tempo e energia dispensadas.

E, por conseguinte, aqueles dados que já possuem valor social identificado pelas instituições, sem dúvida, terão suas análises e interpretações otimizadas, permitindo-se maior controle de veracidade, exatidão e balanceamento, implicando resultados confiáveis e dinâmicos. A integração e interoperabilidade de dados e aplicações, ao otimizar a extração de valor social dos dados, fará com que o processo de tomada de decisões naturalmente migre da análise manual de relatórios estatísticos para outro patamar, qual seja, o de recomendações, calibradas como pontos de apoio ou, em alguns casos específicos, como procedimento de automação destas decisões, sempre com a integração e supervisão humanas.

3.12. Alinhar políticas de *big data* e inteligência artificial aos fins institucionais

Após definir quais valores sociais busca-se extrair dos dados coletados, com a integração deles e interoperabilidade de aplicações, de forma setorizada ou não, para que a instituição de persecução e controle atinja

os melhores resultados, importante que alinhe as políticas de *big data* e inteligência artificial aos seus fins institucionais.

Embora de difícil percepção, quando da integração entre instituições, seja pelo elevado desenvolvimento de algumas delas ou pela credibilidade social que ostentam, há a tendência de que os modelos adotados sejam replicados, sem grandes alterações, nas demais. No entanto, esta replicação, apesar de trazer resultados, nem sempre alcança, em verdade, os fins institucionais propriamente ditos. Ademais, a experiência revela que as políticas de *big data* e inteligência artificial normalmente intercambiáveis interinstitucionalmente restringem-se às atividades-meio, dificilmente às atividades-fim.

Isso se afigura prejudicial à credibilidade das políticas de *big data* e inteligência artificial. À sociedade pouco importa como determinada instituição se organiza para atingir determinado objetivo. A sociedade espera, de fato, que a instituição atinja os fins pelos quais sua criação e estruturação foram motivadas. Processos judiciais, por exemplo, com tramitação anomalamente rápida e célere porque implementadas avançadas políticas de *big data* e inteligência artificial nos serviços da serventia judicial não entregam o valor social esperado pela sociedade se, em sua maioria, forem considerados injustos ou ilegais. Da mesma forma, a rápida e fluida análise e interpretação de grande volume de dados continuamente coletados não terá utilidade alguma se não estiverem atreladas aos fins institucionais.

O avanço na implantação das políticas de *big data* e inteligência artificial deve, de fato, se preocupar com a eliminação da nociva burocracia ainda presente nas instituições de persecução e controle. Todavia, a isso não deve se limitar: deve ser, no mínimo, instrumento capaz de fomentar a aplicação de estruturas semelhantes nas atividades-fim destas mesmas instituições.

Quando se verifica o alinhamento das políticas de *big data* e inteligência artificial aos fins institucionais, aludido pilar dota-as de inegável plasticidade diante do constante dinamismo social, permitindo a continuidade no desenvolvimento e validação das técnicas implantadas mesmo que o panorama de piso tenha sofrido severas alterações. Ao contrário, quando não existe o alinhamento em pauta, as políticas são implementadas geralmente de forma isolada, casuisticamente, para a solução de problemas ou desafios específicos. Têm-se a impressão, inclusive, de que a todo momento uma nova política está sendo implementada, sem fatores que permitam afirmar se tratar de evolução ou continuidade de políticas anteriores. Isso se afigura prejudicial não apenas em termos financeiros e humanos, mas principalmente à imagem destas políticas vislumbradas no seio de cada uma das instituições.

3.13. Fomentar cultura institucional

Estando claro o alinhamento das políticas de *big data* e inteligência artificial com os fins institucionais, facilitada a criação e manutenção de cultura institucional sobre o assunto. Muitas vezes o maior obstáculo a ser vencido na implantação destas políticas não reside no fator orçamentário ou na capacitação humana. Verifica-se a resistência de boa parte dos integrantes da instituição porque consideram perigoso partilhar o processo de tomada de decisões com uma máquina ou instrumento de inteligência artificial. Outros veem

a ampliação destas políticas como uma ameaça correcional e até mesmo à própria manutenção dos seus postos de trabalho.

O desenvolvimento da cultura institucional em políticas de *big data* e inteligência artificial será favorecido caso as instituições pautem em evidências o processo de tomada de decisões. As evidências, por seu turno, devem estar pautadas na constante divulgação e monitoramento dos níveis de eficiência alcançados pela adoção destas políticas, sempre primando pela contínua avaliação por parte dos respectivos membros e colaboradores institucionais, em constante interação com mecanismos de inteligência artificial.

A aferição contínua dos resultados obtidos, além de necessariamente dever compor o plano de desenvolvimento da cultura institucional ora retratada, permite imprimir maior plasticidade às políticas de *big data* e inteligência artificial. Uma vez compreendida a cultura destas políticas pelos integrantes da instituição, ainda que ordenada por evidências e controle contínuo dos resultados, maior será a capacidade de adaptá-las às constantes modificações sociais sem que isso implique retrocessos ou perdas em relação a todo o trabalho anteriormente desenvolvido.

Além do mais, o fomento da cultura institucional vinculada às políticas de *big data* e inteligência artificial, dada a multiplicidade de ideais e realidades inerentes a esta característica, permite melhor validação das técnicas implantadas e maior rapidez na transformação de fatores eventualmente necessária.

3.14. Criar centros de ampliação e divulgação do conhecimento

A criação destes centros, além de facilitar a obtenção de informações por qualquer membro da instituição, muito contribuiu para a celebração de acordos de cooperação interinstitucional com o objetivo de compartilhar interoperabilidade de dados, soluções implementadas, algoritmos de inteligência artificial e, não menos importante, evitar a replicação de erros cometidos, seja na própria ou em outra instituição.

A experiência demonstra que a criação de centros de ampliação e divulgação do conhecimento nem sempre depende do empenho de consideráveis recursos financeiros ou da constituição de equipes numerosas. Circunstância recomendável, no entanto, e que deverá ser observada na medida do que for possível às instituições, é que aludidos centros de ampliação e divulgação do conhecimento primem, invariavelmente, pela multidisciplinariedade de seus integrantes.

Referida multidisciplinariedade impacta, de forma positiva, na contínua aferição do acerto (ou equívoco) da política de *big data* e inteligência artificial escolhida, bem como de seus resultados. Muitas vezes determinada escolha institucional apresenta ótimos resultados apenas em um flanco (ou frente, como costumam se referir os profissionais de gestão) das atividades desempenhadas pela instituição. Nem sempre a escolha "ótima" para determinado setor institucional revela-se sequer "razoável" para outras atividades-fim institucionais. Os centros de ampliação e divulgação do conhecimento têm a capacidade, portanto, dado seu caráter multidisciplinar, de identificar aludidas inconsistências e sugerir a promoção de alterações para tornar a atuação institucional a mais uniforme e homogênea possível frente às finalidades que lhe são atribuídas.

Existindo centros de ampliação e divulgação do conhecimento, maior será a conscientização dos membros e colaboradores institucionais quanto à importância de aderirem às políticas de *big data* e inteligência artificial. Com a canalização das informações para aludido centro, mais eficiente será a confecção de mapas de calor a partir das informações coletadas e maior será a divulgação destes elementos internamente na própria instituição, permitindo a aferição contínua dos resultados obtidos, incrementando análises gráficas e textuais importantes no procedimento de tomada de decisões, com a melhora na alocação de recursos e formas de abordagem de problemas. Permitirá, também, o constante treinamento das habilidades humanas na interação com os mais variados mecanismos de inteligência artificial, aprimorando fluxos de feedback essenciais à plasticidade e manutenção da política implementada.

3.15. Definir modelos apropriados de aprendizagem de máquina

Tão importante quanto coletar, catalogar e armazenar dados é definir qual modelo de aprendizagem de máquina afigura-se mais indicado para atender aos fins institucionais. Há vários ramos de aprendizagem de máquina (*machine learning*) com arquiteturas de processamento distintas, sendo necessário definir quais modelos melhor atendem às necessidades institucionais.

Não se ignora a profundidade do tema relativo à natureza dos modelos de aprendizagem de máquina, bem como sua intrínseca relação com as disciplinas ciência de dados e estatística. Para os restritos fins deste guia de boas práticas apontam-se essencialmente dois modelos de aprendizagem de máquina utilizados, em regra, pelas instituições de persecução e controle quando da implantação de políticas de *big data* e inteligência artificial.

O primeiro destes modelos permite diferenciar com relativa clareza a fase de extração dos dados daquela referente à respectiva interpretação e classificação da informação pelos algoritmos. A verificação humana, neste modelo de aprendizagem, afigura-se presente tanto na aferição dos resultados obtidos como no acompanhamento do próprio procedimento de aprendizagem de máquina em si mesmo.

Já no segundo modelo de aprendizagem em apreço, a verificação humana, ainda que possível, restringe-se ao acompanhamento dos resultados obtidos a partir do reconhecimento de padrões e aprendizado computacional em inteligência artificial. Em outras palavras, neste modelo de aprendizagem não se afigura possível diferenciar a fase de extração da de classificação de dados, ainda que presente a verificação humana sobre os resultados recebidos.

Devem as instituições primar pela utilização de modelos de aprendizagem de máquina da primeira espécie, por permitirem maior controle do acerto das políticas de *big data* e inteligência artificial implementadas. De outra parte, isolando-se as fases de extração e catalogação dos dados pelos algoritmos, imprime-se maior transparência aos resultados alcançados, permitindo a utilização em estratégias de imputação ou instrução para detectar e punir responsáveis pela prática de atos de corrupção e lavagem de capitais. Igualmente, em decorrência da necessária fundamentação dos atos institucionais, constitucionalmente prevista, essa forma de utilizar conjuntos de algoritmos adequa-se muito mais à natureza dos sistemas punitivo e processual vigentes, assegurando-se o respeito à proteção dos dados conforme regime regulatório também em vigor.

De outra parte, como a aprendizagem por algoritmos ocorre pelas experiências e resultados obtidos e não pelo volume, em si, dos dados que os alimentam, a utilização deste modelo de aprendizagem permite o desenvolvimento, simultaneamente e em larga escala, tanto de algoritmos prescritivos (a analisar realidades ou fatos já verificados) quanto preditivos (estimando realidades ou fatos futuros), auxiliando a implantação, também necessária, de política de governança de modelos de inteligência artificial.

No que tange ao segundo modelo de aprendizagem analisado, que não permite, ainda que contando com a verificação humana, diferenciar a fase de extração da respectiva catalogação dos dados, importante esclarecer que, em geral, ocorre por redes neurais convolucionais, estabelecidas por camadas de percpetrons multi-layers indecomponíveis, que não permitem enxergar o trabalho desenvolvido pelos algoritmos, muitos menos quais dados especificamente teriam sido utilizados na operação. São os chamados modelos de aprendizagem por conjunto de algoritmos caixa-preta ou *blackbox*.

O modelo de aprendizagem destes algoritmos pode abordar, inclusive, redes de relacionamento associadas ou não à prática de ilícitos. Desta forma, os algoritmos podem se utilizar de informações que não se afiguram apropriadas para fins de imputação de responsabilidades, seja no âmbito administrativo ou judicial, mas que, de algum modo, revelaram durante o procedimento de aprendizado automático alguma relevância para estabelecimento do resultado ou conclusão gerados. Em virtude da tendência nacional de restrição legislativa crescente, uma vez que a natureza e conteúdo dos dados analisados pelos algoritmos não se mostra visível (muito menos identificável) ao usuário desta tecnologia, muitas instituições os veem como uma alternativa para o desenvolvimento de modelos preditivos, sem a infringência de qualquer política de privacidade, para permitir a melhor alocação de recursos e pessoas no combate a atos de corrupção e lavagem de capitais. Cuidar-se-ia, apenas, do âmbito de informações de inteligência, não utilizadas para incriminar ou responsabilizar quem quer que seja, mas para melhor definição das estratégias a serem seguidas pelas instituições. Dado o risco inerente a esta política, bem como a dependência que provoca no processo de tomada de decisões em relação a fatores de automação, de rigor seja utilizada pelas instituições apenas em casos pontuais e bastante específicos.

3.16. Treinar e verificar algoritmos

Definida qual natureza algorítmica melhor atende aos interesses e objetivos institucionais, importante manter fluxos de trabalho constantemente voltados ao treinamento e verificação destes algoritmos. Destaque-se que a continuidade deste treinamento não se restringe aos parâmetros de automação, devendo alcançar a própria interação humana.

Para tanto, a instituição deve ter como boa prática a implementação de mecanismos capazes de monitorar, continuamente, as habilidades em *big data* e inteligência artificial de todos que interajam, de algum modo, com referidas políticas. Imprescindível que a instituição deixe bastante claro aos seus integrantes quais políticas adota no desenvolvimento de algoritmos, isto é, se prevalecem mecanismos de autossuficiência de criação (desenvolvimento de algoritmos próprios), se existe a possibilidade de celebração de acordos de cooperação técnica neste ponto (compartilhamento de experiência interinstitucionais, ainda que os algoritmos

sejam desenvolvidos no ambiente interno de cada uma das instituições), se há preferência pela setorização dos módulos de desenvolvimento dos algoritmos (ainda que exista cooperação interinstitucional, cada ente fica responsável por uma parcela do desenvolvimento, de sorte que o algoritmo final possa ter aplicações distintas) ou ainda compartilhamento de algoritmos de inteligência artificial (cada ente desenvolve o próprio algoritmo, porém por acordo de cooperação técnica autoriza-se o compartilhamento recíproco das tecnologias conquistadas).

De todo modo, o treinamento e verificação de algoritmos deve pautar-se, sempre, por resultados obtidos a partir de evidências, ainda que se utilizem modelos de aprendizagem por *deep learning* (mesmo que indecomponíveis as fases de extração e catalogação automática dos dados, a verificação humana sobre os resultados obtidos será sempre necessária). Não existe qualquer contradição neste aspecto, desde que realizada a escolha correta da natureza dos algoritmos utilizados frente aos resultados buscados. Ainda que diante, eventualmente, de algoritmos *blackbox* de cunho exclusivamente preditivo, a análise dos mapas ou manchas de calor estimados comparativamente à realidade que, de fato, futuramente vier a ser constatada, permitirá à interação humana verificar os resultados e definir se aquele modelo algorítmico atende, ou não, às necessidades institucionais.

Já em se cuidando de modelos prescritivos, em que a interação humana no curso da aprendizagem algorítmica se faz presente de forma mais notória, o treinamento e verificação devem ter por escopo fomentar a automática e assistida detecção de eventos relacionados à corrupção e lavagem de capitais. A partir do monitoramento constante dos resultados obtidos, o processo de tomada de decisões poderá atingir maiores graus de automação, afigurando-se mais célere e eficiente, porém sem descartar a necessária interação humana no respectivo treinamento e verificação.

Em estágios mais avançados de treinamento e desenvolvimento de algoritmos, os fluxos de trabalho poderão permitir a interação entre algoritmos e conjuntos de algoritmos de diversas naturezas. Esta conexão autorizará o acompanhamento e comparação de manchas e mapas de calor, permitindo metrificar as evoluções temporais respectivamente por eles registradas, resultando na automatização de procedimentos de auditoria capazes de otimizar a alocação de recursos e pessoas pelas instituições envolvidas no combate a atos de corrupção e lavagem de capitais.

Com a consolidação clara e precisa de políticas de treinamento e verificação em inteligência artificial, mais propício será o ambiente para executar a política de governança de dados e permitir, ainda que de forma limitada e controlada, a colaboração da sociedade civil na estruturação não apenas de procedimentos, como também na obtenção e aferição dos resultados obtidos.

4. CONCLUSÃO

O presente guia não pretende, de modo algum, esgotar o tema alusivo às boas práticas a serem seguidas pelas instituições de persecução e controle. Reflete, pura e simplesmente, as principais preocupações e alinhamentos registrados pelos integrantes da Estratégia Nacional de Combate à Corrupção e Lavagem de Capitais (ENCCLA) no curso de inúmeras reuniões realizadas, com a contribuição inestimável da experiência de diversas instituições e apontamentos da academia. Também não se afigura como seu objetivo adentrar análise de termos técnicos ou peculiaridades institucionais. As boas práticas procuram não privilegiar instituições, muito menos apontar quais estariam em estágios avançados de implantação ou o contrário disso. Também não se buscou descrever modelos que devessem ser seguidos à risca pelas instituições, sob pena de fracasso.

As ideias retratadas aqui procuram, apenas, planificar tudo o que de bom já foi realizado e implementado por diversos órgãos e instituições, para assim orientar a elaboração de políticas de *big data* e inteligência artificial de modo mais harmônico possível. Busca-se imprimir maior celeridade às implantações, reduzir drasticamente os custos envolvidos (de toda ordem, frise-se), e fomentar a interoperabilidade interinstitucional de tecnologias, seja na obtenção, classificação e armazenamento de dados; seja na interpretação, obtenção de resultados e convergência de processos de tomada de decisão.

Todos os assuntos tratados no presente guia de boas práticas são, ainda que minimamente, interrelacionados. Poder-se-ia destrinchar, ainda mais, cada um dos tópicos acima discriminados. Isso, porém, retiraria a objetividade esperada do presente trabalho. Nada impede, por outro lado, que constitua um marco que permita ao usuário aprofundar-se em qualquer um dos temas aqui explanados de acordo com respectiva necessidade identificada.

O gestor público, portanto, deve conseguir valorar, dentre as boas práticas aqui delineadas, quais se afiguram aplicáveis à realidade de que dispõe e, dentre estas, quais se mostram de implementação mais urgente e necessária. Espera-se de toda forma que, ao se levar em consideração as boas práticas descritas, as instituições de combate a atos de corrupção e lavagem de capitais consigam harmonizar as políticas de *big data* e inteligência artificial, dotando-as de continuidade e plasticidade suficientes a alcançar os valorosos fins almejados em benefício de uma sociedade que, embora dinâmica, ainda se afigura muito ressentida com antigos problemas, e que demanda de órgãos e instituições a premente e inadiável busca de novas soluções.