Quais devem ser os parâmetros éticos e jurídicos para a utilização da inteligência artificial?

As respostas oferecidas pelas recentes Diretrizes da União Europeia para a inteligência artificial confiável

---

Ana Frazão

Advogada. Professora de Direito Civil e Comercial da UnB. Ex-Conselheira do CADE.

No último dia 8 de abril, a Comissão Europeia divulgou as Diretrizes Éticas para a Inteligência Artificial Confiável. Partindo da preocupação de que a inteligência artificial, ao mesmo tempo em que traz benefícios substanciais para os indivíduos e para a sociedade, também apresenta erros, riscos e impactos negativos que podem ser de difícil antecipação, identificação e mensuração, o Guia procura oferecer as orientações essenciais para endereçar tais problemas.

Importante premissa do Guia é a de que a inteligência artificial, para ser confiável, precisa ser lícita, ética e robusta, tanto da perspectiva técnica quanto da perspectiva social, considerando os riscos, ainda que não intencionais, que oferece para a democracia, as garantias legais (*rule of law)*, a justiça distributiva, os direitos fundamentais e mesmo a mente humana. Daí a premissa básica de que os sistemas de inteligência artificial precisam ser centrados no homem e alicerçados no compromisso de serem utilizados a serviço da humanidade, do bem comum e da liberdade.

Na medida em que é dirigido a todos os que, de alguma maneira, se interessam ou participam do *design*, desenvolvimento, implementação ou uso dos sistemas de inteligência artificial, assim como também àqueles que por eles serão afetados, as recomendações e exigências do Guia são de ampla abrangência, abarcando companhias, organizadores, pesquisadores, serviços públicos, governo, agências,

instituições, organizações da sociedade civil, indivíduos, trabalhadores e consumidores.

Entretanto, diante das preocupações com a igualdade, a não discriminação e a solidariedade, percebe-se que os tutelados prioritariamente pelas determinações do Guia são (i) os grupos considerados vulneráveis, tais como crianças, pessoas com deficiências, minorias étnicas e outros que foram historicamente colocados em desvantagem ou estão em risco de exclusão, e (ii) aqueles que se encontram em relações assimétricas, tais como empregados e consumidores. Como se verá adiante, um dos maiores objetivos do Guia é evitar resultados injustamente enviesados que possam prejudicar exatamente os já vulneráveis, o que impõe, dentre outras exigências, que os dados usados para treinar os sistemas de inteligência artificial sejam os mais inclusivos possíveis e representem diferentes grupos populacionais.

Princípios

O primeiro passo para a compreensão do Guia é entender os quatro princípios éticos que constituem os seus fundamentos: (i) o respeito pela autonomia humana, (ii) a prevenção de danos, (iii) a justiça e (iv) a explicabilidade.

O respeito pela autonomia humana requer a observância da dignidade e o tratamento das pessoas como sujeitos morais e não como objetos que podem ser esquadrinhados, classificados, avaliados em termos de notas ou medidas, arrebanhados, condicionados, coagidos, subordinados ou manipulados. Com isso, reforça-se o compromisso da inteligência artificial com a proteção do homem, o que envolve a sua integridade física e mental, o seu senso de identidade pessoal e cultural, a liberdade e a independência para tomar decisões. Envolve igualmente o igual acesso aos direitos, benefícios e oportunidades vinculados aos respectivos sistemas.

Consequentemente, o Guia propõe rígido controle sobre as ameaças da inteligência artificial sobre a saúde mental das pessoas, bem como sobre a vigilância injustificada e os riscos de enganos e manipulações injustas. Como os sistemas de inteligência artificial devem ser desenhados para aumentar, complementar e empoderar as habilidades cognitivas, sociais e culturais dos seres humanos, o Guia propõe que a alocação de funções entre homens e sistemas de inteligência artificial tenham o homem como preocupação central desde o seu *design* (*human-centric design principles*) e possibilitem oportunidade significativa para a escolha humana.

Por essa razão, acentua a necessidade de se assegurar a supervisão humana sobre todos os ciclos de vida de tais sistemas.

O princípio da prevenção de danos está associado à exigência de robustez técnica e segurança, o que será mais bem desenvolvido adiante. Já no que diz respeito à justiça, tal princípio tem uma dimensão substantiva e outra procedimental que, se devidamente observadas, podem até possibilitar que a inteligência artificial seja uma ferramenta para a implementação de justiça.

A primeira dimensão da justiça busca assegurar (i) igual e justa distribuição dos custos e benefícios entre as pessoas, de modo a garantir que indivíduos e grupos estarão livres de vieses injustos, discriminações e estigmatizações, bem como (ii) proporcionalidade entre meios e fins, bem como a necessidade de balanceamento entre os interesses e objetivos conflitantes. A segunda dimensão busca assegurar a possibilidade de contestar decisões tomadas por sistemas de inteligência artificial ou por humanos que os operam, assim como de obter respostas contra as impugnações. Daí a necessidade de que a entidade responsável pelas decisões seja identificável e que o processo em si seja explicável.

Tal aspecto já conecta o princípio da justiça com o da explicabilidade, considerado como crucial para a construção e a manutenção da confiança dos usuários nos sistemas de inteligência artificial. Isso significa que os processos devem ser transparentes e suscetíveis de comunicação aberta, assim como as decisões devem ser, na medida do possível, explicáveis para aqueles que são direta e indiretamente afetados por ela, até para que tenham condições de contestá-la, se for o caso.

Nesse ponto, o Guia reconhece que nem sempre é possível a explicação sobre as razões pelas quais o modelo gerou um particular resultado. Casos assim, também chamados de *black boxes*, requerem particular atenção e a adoção de outras medidas de explicabilidade, tais como a rastreabilidade, a auditabilidade e a comunicação transparente sobre as capacidades dos sistemas. O grau de explicabilidade também depende do contexto e da severidade das consequências de resultados equivocados ou sem a devida acurácia.

Exigências

Além dos quatro princípios éticos já mencionados, o Guia está também alicerçado em sete exigências, que devem ser avaliadas continuamente ao longo de todo o ciclo de vida do sistema de inteligência artificial: (i) *human agency* e supervisão

humana, (ii) robustez técnica e segurança, (iii) privacidade e governança de dados, (iv) transparência, (v) diversidade, não discriminação e justiça, (vi) bem estar e ambiental e social e (vii) *accountability*.

Dialogando diretamente com os princípios éticos, as exigências dirigem-se claramente aos desenvolvedores - aqueles que pesquisam, criam ou desenvolvem os sistemas de inteligência artificial -, e aos implantadores – aqueles que usam os sistemas em seus negócios, especialmente para oferecer produtos e serviços -. Observa-se, portanto, que o Guia sugere uma espécie de responsabilidade compartilhada entre todos os que exploram os sistemas de inteligência artificial.

A exigência de *human agency* e supervisão diz respeito ao fato de que os sistemas de inteligência artificial, em respeito à autonomia humana, devem possibilitar uma sociedade democrática e equitativa, a realização dos direitos fundamentais e, qualquer caso, a supervisão humana. Tal preocupação exige que, antes mesmo do desenvolvimento, haja a avaliação de riscos, processo que deve perdurar continuamente, inclusive por meio de mecanismos que possam receber *feedbacks* externos, especialmente quando os sistemas de inteligência artificial possam infrigir direitos fundamentais.

Tal exigência também dialoga com a necessidade de que os usuários sejam capazes de fazer decisões autônomas e informadas sobre sistemas de inteligência artificial, conhecer as ferramentas para compreender e interagir com tais sistemas em um grau satisfatório, assim como avaliá-los razoavelmente e contestá-los. Daí o especial cuidado que se deve ter com sistemas que influenciam e moldam o comportamento humano por meio de artifícios difíceis de serem identificados, especialmente quando se utilizam de processos subconscientes. Outro aspecto importante é a necessidade de se respeitar o direito de não ser sujeito a uma decisão totalmente automatizada quando ela produza efeitos jurídicos ou significativos sobre os indivíduos.

Ja a supervisão diz respeito à implementação de mecanismos de governança, alguns dos quais são expressamente mencionados pelo Guia, tais como (i) o *human-in-the-loop* (HITL), que se refere à capacidade de intervenção humana em cada ciclo de decisão do sistema, o que nem sempre é possível ou desejável, (ii) o *human-on-the-loop* (HOTL), que se refere à capacidade de intervenção humana durante o ciclo do *design* do sistema e o monitoramento da sua operação e o (iii) *human-in-command* (HIC), que se refere à capacidade de supervisionar a atividade global do sistema,

incluindo seus impactos econômicos, sociais, jurídicos e éticos, bem como a habilidade de decidir quando e como usar o sistema em determinada situação, o que pode incluir a decisão de não usá-lo em determinada situação, de estabelecer níveis de discricionariedade humana durante o seu uso ou mesmo de assegurar a habilidade de contornar a decisão do sistema.

Mais do que isso, é importante assegurar que agentes públicos tenham a habilidade de supervisionar tais sistemas por meio de mecanismos que podem ser exigidos em vários graus para apoiar segurança e medidas de controle, dependendo da área de aplicação ou dos riscos de cada sistema de IA. A ideia fundamental é que, via de regra, quanto menos supervisão humana possa haver sobre um sistema de inteligência artificial, mais serão necessários testes extensivos e uma rígida governança.

A exigência de robustez técnica e segurança, associada ao princípio da prevenção do dano, faz uma diferenciação entre danos, partindo da premissa de que deve haver a minimização de danos não intencionais ou não esperados e a prevenção de danos inaceitáveis. Também requer a acurácia dos sistemas de inteligência artificial, especialmente nas situações em que afetem diretamente as vidas humanas. Logo, espera-se que tais sistemas possam fazer julgamentos, predições, recomendações e decisões corretos, bem como indicar prováveis erros sempre que as predições incorretas não puderem ser evitadas.

Além disso, a robustez técnica e a segurança têm por finalidade assegurar a confiança e a reprodutibilidade, sendo esta última identificada como a possibilidade de verificar se um experimento de inteligência artificial exibe o mesmo comportamento quando reproduzido sob as mesmas condições, propriedade importante para que cientistas e reguladores possam descrever o que tais sistemas fazem.

No tocante à governança de dados e de privacidade, o Guia enfatiza a exigência da qualidade e da integridade dos dados, bem como do acesso a eles. Mesmo a coleta de dados deve ser cuidadosa, a fim de que vieses, erros e faltas de acurácia sejam resolvidos antes mesmo do início do treinamento do sistema, especialmente se for o caso de *self-learning*. Consequentemente, os bancos ou conjuntos de dados precisam ser testados e documentados em cada passo - planejamento, treinamento, testagem e implementação -, cuidado que também se aplica aos sistemas que são adquiridos de terceiros.

A exigência de transparência, segundo o Guia, está diretamente associada à rastreabilidade, à explicabilidade e à comunicação. No que diz respeito à primeira, ressalta o Guia que as bases de dados e os processos sejam documentados da melhor forma possível, assim como os processos decisórios, pois somente assim será possível identificar as razões pelas quais as decisões dos sistemas de inteligência artificial são erradas e prevenir erros futuros. Nesse sentido, a rastreabilidade promove auditabilidade e explicabilidade, vista esta última como a habilidade de explicar tantos os processos técnicos como as próprias decisões que podem trazer impactos sobre seres humanos.

A parte técnica da explicabilidade exige que as decisões dos sistemas de inteligência artificial possam ser compreendidas e rastreadas por seres humanos, assim como que as explicações sejam disponibilizadas em tempo adequado e de acordo com o grau de expertise daquele que será por ela afetado. Por mais que possam ocorrer *tradeoffs* entre a explicabilidade e a acurácia do sistema, isso não afasta a necessidade da explicabilidade.

Nesse ponto, o Guia ressalta a importância de se entender o grau com que um sistema de inteligência artificial influencia e molda o processo de decisão de uma organização e desenha suas escolhas. Daí sustentar que deve ser suscetível de explicação não apenas a racionalidade da implantação do sistema de inteligência artificial, como também o próprio modelo negocial.

Já a necessidade da comunicação diz respeito ao fato de que os sistemas de inteligência artificial não podem se apresentar como humanos para os usuários, pois estes têm o direito de serem informados quando interagem com tais sistemas, até para que possam não prosseguir, se assim desejarem.

A exigência de diversidade, não discriminação e justiça tem por objetivo evitar vieses injustos, preocupando-se com a inclusão e a diversidade ao longo de todos os ciclos de vida do sistema de inteligência artificial. Tem, portanto, clara conexão com o princípio da justiça e com as preocupações de que os sistemas de inteligência artificial possam sofrer a influência de vieses históricos por inadvertência, ausência de completude e maus modelos de governança. Especial preocupação também existe em relação à exploração intencional dos vieses dos consumidores ou ao engajamento em competições injustas, como a colusão ou a ausência de transparência.

Consequentemente, determina o Guia que vieses discriminatórios sejam removidos já na fase de coleta de dados, sempre que possível. Da mesma maneira, as

formas como os sistemas de inteligência artificial são desenvolvidos, como a programação de algoritmos, podem também sofrer de vieses injustos, o que precisa ser endereçado por meio de processos de supervisão para analisar e avaliar o sistema, de forma clara e transparente, no que diz respeito aos seus propósitos, restrições, exigências e decisões. O Guia ainda ressalta a importância da diversidade de opiniões e da participação dos interessados, assim como que os sistemas não tenham uma abordagem *one-size-fits-all*, mas sim sejam adaptados às características de cada usuário.

A exigência de bem estar social e ambiental reforça o comprometimento da inteligência artificial com o bem estar do homem em uma perspectiva macro, o que envolve igualmente os cuidados com a democracia.

Por fim, a exigência de *accountability*, que está intrinsecamente relacionada à justiça, tem também desdobramentos importantes sobre a possibilidade de auditagem (*auditability*) sobre os dados, os algoritmos e os processos de design dos sistemas de inteligência artificial. Nesse ponto, o Guia tem o cuidado de ressaltar que tal abordagem não necessariamente implica que a informação sobre os modelos de negócios e a propriedade intelectual envolvida sejam sempre disponibilizadas abertamente. Para o Guia, a avaliação por auditores internos e externos e a disponibilização de tais relatórios de avaliação podem contribuir para a confiabilidade na tecnologia. Entretanto, em aplicações que afetam direitos fundamentais, os sistemas precisam ser abertos para uma auditoria independente.

A *accountability* também está relacionada à minimização de danos e ao reporte dos impactos negativos. Consequentemente, impõe para desenvolvedores e implementadores os deveres de identificar, avaliar, documentar e minimizar os potenciais impactos negativos dos sistemas de inteligência artificial, bem como se utilizarem de avaliações de impacto.

Para alcançar a *accountability*, o Guia reconhece que existem alguns *tradeoffs*. Entretanto, entende que estes precisam ser endereçados de maneira racional e metodológica, de acordo com o "estado da arte", a fim de que possam ser reconhecidos e avaliados, do ponto de vista axiológico, de acordo com o risco aos princípios éticos e aos direitos fundamentais. O Guia é ainda categórico ao afirmar que, nas situações em que nenhum *tradeoff* puder ser considerado eticamente aceitável, a conclusão possível é a de que o desenvolvimento, a implantação e o uso do sistema de inteligência artificial não podem ser feitos dessa maneira.

O Guia é também explícito a respeito do papel de quem decide o *tradeoff*, determinando que o tomador da decisão seja responsável pela maneira como o mesmo um *tradeoff* considerado apropriado endereçado, impondo-lhe também a obrigação de rever continuamente a adequação da sua decisão, a fim de assegurar mudanças necessárias que possam ser feitas no sistema.

A *accountability* ainda tem a ver com a capacidade do sistema de se corrigir diante da constatação de um impacto adverso e injusto.

Como se pode observar a partir da brevíssima síntese ora exposta, as Diretrizes da Comissão Europeia endereçam de forma muito satisfatória as principais preocupações a respeito da utilização crescente da inteligência artificial e tornam-se um excelente ponto de referência para as discussões em torno do assunto, especialmente no tocante (i) aos direitos dos que são afetados por sistemas de inteligência artificial, ainda mais quando vulneráveis, assim como (ii) aos deveres e responsabilidades dos desenvolvedores e dos implementadores dos referidos sistemas.