MAPEAMENTO DE PRINCÍPIOS DE INTELIGÊNCIA ARTIFICIAL

AUTORES:
Caroline Burle e Diogo Cortiz

SUMÁRIO:

ceweb.br nic.br cgi.br

Introduction

Este é um mapeamento multissetorial e não-exaustivo de princípios de Inteligência Artificial. Mapeamos seis iniciativas internacionais, sendo duas do setor governamental (Comissão Europeia e Departamento de Defesa Norte-americano), duas do setor empresarial (Google e Microsoft), uma organização internacional (Organização para a Cooperação e Desenvolvimento Econômico - OCDE) e outra composta de academia e setor empresarial (Academia de Inteligência Artificial de Pequim).

Selecionamos essas iniciativas com o intuito de garantir pluralidade regional (Estados Unidos, Europa e Ásia) – não foi encontrado nenhum material produzido para o Brasil até o momento da escrita deste texto – e multissetorial (setor privado, academia, governos e organização internacional).

Analisamos os princípios de Inteligência Artificial encontrados em cada uma dessas iniciativas, com base em seis dimensões: Equidade (Fairness); Confiabilidade e Segurança (Reliability & Safety); Impacto Social (Social Impact); Responsabilidade (Accountability); Privacidade & Segurança (Privacy & Security); e Transparência (Transparency).

As seis dimensões foram elaboradas pelos autores deste texto, a partir das leituras prévias de cada documento. Após encontrar similaridades entre os documentos, compreendemos que essas dimensões seriam adequadas para mapear os princípios de Inteligência Artificial das iniciativas analisadas.

Princípios de Inteligência Artificial por dimensões

Equidade (*Fairness*)

Comissão Europeia

O desenvolvimento, a implantação e o uso de sistemas de Inteligência Artificial devem ser justos. Reconhecem que existem muitas interpretações diferentes de equidade (ou justiça), e dividem em duas dimensões: substantiva e procedural. A dimensão substantiva implica o compromisso de garantir uma distribuição equitativa e justa de benefícios e custos, assim como garantir que indivíduos e grupos estejam livres de preconceitos injustos, discriminação e estigmatização. Se desvios injustos puderem ser evitados, os sistemas de Inteligência Artificial podem até aumentar a justiça social. A igualdade de oportunidades em termos de acesso à educação, bens, serviços e tecnologia também deve ser promovida. O uso de sistemas de IA nunca deve levar as pessoas a serem enganadas ou injustificadamente prejudicadas em sua liberdade de escolha. Além disso, a equidade implica que os

profissionais de IA respeitem o princípio da proporcionalidade entre meios e fins e considerem cuidadosamente como equilibrar interesses e objetivos concorrentes. A dimensão procedural busca reparação efetiva contra as decisões tomadas pelos sistemas de Inteligência Artificial e pelos humanos que os operam. A entidade responsável pela decisão deve ser identificável e os processos de tomada de decisão devem ser explicáveis.

Departamento de Defesa Norte-americano

O Departamento de Defesa deve tomar medidas para evitar preconceitos não intencionais no desenvolvimento e implantação de sistemas de Inteligência Artificial de combate ou não-combate que inadvertidamente causariam danos a pessoas.

Academia de Inteligência Artificial de Pequim

Pesquisa e Desenvolvimento de Inteligência Artificial deve adotar abordagens éticas de design para tornar o sistema confiável. Isso pode incluir, mas não se limita a: tornar o sistema o mais equitativo possível, reduzir possíveis discriminações e preconceitos, melhorar sua transparência, prover explicação e previsibilidade e tornar o sistema mais rastreável, auditável e responsável.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

Os sistemas de Inteligência Artificial devem ser projetados de maneira a respeitar o estado de direito, os direitos humanos, os valores democráticos e a diversidade, e devem incluir salvaguardas apropriadas - por exemplo, possibilitando a intervenção humana sempre que necessário - para garantir uma sociedade equitativa e justa.

Google

Evitar criar ou reforçar preconceitos injustos sobre as pessoas, particularmente aquelas relacionadas a características sensíveis, como raça, etnia, gênero, nacionalidade, renda, orientação sexual, habilidade e crenças políticas ou religiosas.

Microsoft

Os sistemas de Inteligência Artificial devem tratar todas as pessoas de maneira equitativa.

Confiabilidade e Segurança (*Reliability & Safety*)

Comissão Europeia

Os sistemas de Inteligência Artificial não devem causar nem exacerbar danos ou afetar adversamente os seres humanos. Devem proteger a dignidade humana, bem como a integridade mental e física. Os sistemas de IA e os ambientes em que operam devem ser seguros, tecnicamente robustos e deve-se garantir que não estejam abertos ao uso malicioso. Pessoas vulneráveis devem receber maior atenção e ser incluídas no desenvolvimento, implantação e uso de sistemas de IA. Também deve ser dada atenção especial a situações em que os sistemas de IA podem causar ou exacerbar impactos adversos devido a assimetrias de poder ou informações, como entre empregadores e funcionários, empresas e consumidores ou governos e cidadãos. A prevenção de danos também implica a consideração do ambiente natural e de todos os seres vivos.

Departamento de Defesa Norte-americano

Os sistemas de inteligência artificial do Departamento de Defesa devem ter um domínio de uso explícito e bem definido, e a segurança, a proteção e a robustez de tais sistemas devem ser testadas e garantidas por todo o ciclo de vida desse domínio de uso. Os sistemas de Inteligência Artificial do Departamento de Defesa devem ser projetados para cumprir sua função pretendida e, ao mesmo tempo, possuir a capacidade de detectar e evitar danos ou interrupções não intencionais, e para o desengajamento ou desativação humana ou automatizada de sistemas implantados que demonstram comportamento de escala ou outro não intencional.

Academia de Inteligência Artificial de Pequim

Esforços contínuos devem ser feitos para melhorar a maturidade, robustez, confiabilidade e controlabilidade dos sistemas de Inteligência Artificial, de modo a garantir a segurança dos dados, a segurança do sistema de IA e a segurança do ambiente externo, no qual o sistema de IA é implantado.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

Os sistemas de Inteligência Artificial devem funcionar de maneira robusta, segura e protegida ao longo de seus ciclos de vida, e os riscos em potencial devem ser avaliados e gerenciados continuamente.

Google

Ser construído e testado para segurança para evitar resultados não intencionais que criam riscos de danos. Os sistemas de Inteligência Artificial serão projetados e desenvolvidos de acordo com as melhores práticas de pesquisa de segurança de

IA. As tecnologias de IA serão testadas em ambientes restritos e a sua operação monitorada após a implantação.

Microsoft

Os sistemas de Inteligência Artificial devem ter um desempenho confiável e seguro.

Impacto Social (*Social Impact*)

Comissão Europeia

Os direitos fundamentais em que se baseia a União Europeia visam garantir o respeito pela liberdade e autonomia dos seres humanos. Os seres humanos que interagem com os sistemas de IA devem ser capazes de manter a autodeterminação completa e eficaz sobre si mesmos e participar do processo democrático. Os sistemas de IA não devem injustificadamente subordinar, coagir, enganar, manipular, condicionar ou agrupar humanos. Em vez disso, eles devem ser projetados para aumentar, complementar e capacitar as habilidades cognitivas, sociais e culturais humanas. A alocação de funções entre humanos e sistemas de IA deve seguir os princípios de design centrado no ser humano (*human centered design*) e deixar oportunidades significativas para a escolha humana. Isso significa garantir a supervisão humana sobre os processos de trabalho nos sistemas de IA. Os sistemas de IA também podem mudar fundamentalmente a esfera de trabalho. Devem apoiar os seres humanos no ambiente de trabalho e ter como objetivo a criação de trabalho significativo.

Departamento de Defesa Norte-americano

Não consta.

Academia de Inteligência Artificial de Pequim

A Inteligência Artificial deve ser projetada e desenvolvida para promover o progresso da sociedade e da civilização humana, promover o desenvolvimento sustentável da natureza e da sociedade, beneficiar toda a humanidade e o meio ambiente e melhorar o bem-estar da sociedade e da ecologia. O desenvolvimento da Inteligência Artificial deve refletir a diversidade e a inclusão e ser projetado para beneficiar o maior número possível de pessoas, especialmente aquelas que seriam facilmente negligenciadas ou sub-representadas nos aplicativos de IA. Encoraja-se o estabelecimento de plataformas abertas de Inteligência Artificial para evitar monopólios de dados ou plataforma. Além de compartilhar os benefícios do desenvolvimento da IA na maior extensão possível e de promover oportunidades iguais de desenvolvimento para diferentes regiões e indústrias.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

A Inteligência Artificial deve beneficiar as pessoas e o planeta, impulsionando o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar.

Google

Ser socialmente benéfico e manter altos padrões de excelência científica. Levar em conta uma ampla gama de fatores sociais e econômicos e prosseguir onde acreditam que os benefícios prováveis gerais excedam substancialmente os riscos e desvantagens previsíveis. Disponibilizar informações precisas e de alta qualidade usando a IA, de forma a respeitar as normas culturais, sociais e legais nos países em que opera. Avaliar cuidadosamente quando disponibilizar as tecnologias da Google em bases não comerciais.

Microsoft

Os sistemas de IA devem capacitar todos e envolver as pessoas.

Responsabilidade (Accountability)

Comissão Europeia

Inclui auditoria, minimização e relatórios de impacto negativo e trade-offs. O requisito de responsabilidade está intimamente ligado ao princípio da equidade. É necessário que sejam criados mecanismos para garantir a responsabilidade e a prestação de contas dos sistemas de IA e seus resultados, antes e depois de seu desenvolvimento, implantação e uso.

Departamento de Defesa Norte-americano

Os seres humanos devem exercer níveis adequados de julgamento e permanecer responsáveis pelo desenvolvimento, implantação, uso e resultados dos sistemas de Inteligência Artificial do Departamento de Defesa.

Academia de Inteligência Artificial de Pequim

Pesquisadores e desenvolvedores de Inteligência Artificial devem ter considerações suficientes para os possíveis impactos e riscos éticos, legais e sociais trazidos por seus produtos e tomar ações concretas para reduzi-los e evitá-los.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

As organizações e indivíduos que desenvolvem, implantam ou operam sistemas de Inteligência Artificial devem ser responsabilizados pelo seu bom funcionamento, de acordo com os princípios acima.

Google

Ser responsável perante as pessoas. Projetar sistemas de Inteligência Artificial que ofereçam oportunidades apropriadas para feedback, explicações relevantes e apelo. As tecnologias de IA estarão sujeitas à direção e controle humanos adequados.

Microsoft

Os sistemas de Inteligência Artificial devem ter responsabilidade algorítmica.

Privacidade & Segurança (*Privacy & Security*)

Comissão Europeia

Inclui respeito pela qualidade da privacidade e integridade dos dados e acesso aos dados. Intimamente ligada ao princípio de prevenção de danos está a privacidade, um direito fundamental particularmente afetado pelos sistemas de IA. A prevenção de danos à privacidade também exige governança de dados adequada que cubra a qualidade e a integridade dos dados utilizados, sua relevância à luz do domínio em que os sistemas de IA serão implantados, seus protocolos de acesso e a capacidade de processar dados de maneira que proteja a privacidade.

Departamento de Defesa Norte-americano

Não consta.

Academia de Inteligência Artificial de Pequim

A pesquisa e desenvolvimento da Inteligência Artificial deve servir à humanidade e estar em conformidade com os valores humanos, bem como com os interesses gerais da humanidade. Privacidade, dignidade, liberdade, autonomia e direitos humanos devem ser suficientemente respeitados. A IA não deve ser usada contra, utilizar ou prejudicar seres humanos.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

Não consta.

Google

Incorporar princípios de privacidade desde a concepção do projeto (*privacy by design)*. Incorporar princípios de privacidade no desenvolvimento e uso de nossas tecnologias de IA. Dar oportunidade de aviso e consentimento, incentivar arquiteturas com salvaguardas de privacidade e fornecer transparência e controle adequados sobre o uso de dados.

Microsoft

Os sistemas de Inteligência Artificial devem ser seguros e respeitar a privacidade.

Transparência (*Transparency*)

Comissão Europeia

É crucial para criar e manter a confiança dos usuários nos sistemas de Inteligência Artificial. Isso significa que os processos precisam ser transparentes, as capacidades e o objetivo dos sistemas de IA comunicados abertamente e as decisões - na medida do possível - explicáveis para os afetados direta e indiretamente. Sem essas informações, uma decisão não pode ser devidamente contestada. Uma explicação sobre por que um modelo gerou uma saída ou decisão específica (e que combinação de fatores de entrada contribuiu para isso) nem sempre é possível. Esses casos são chamados de algoritmos de 'caixa preta' e requerem atenção especial. Nessas circunstâncias, outras medidas para prover explicação (por exemplo, rastreabilidade, auditabilidade e comunicação transparente sobre as capacidades do sistema) podem ser necessárias, desde que o sistema como um todo respeite os direitos fundamentais.

Departamento de Defesa Norte-americano

A disciplina de Engenharia de Inteligência Artificial do Departamento de Defesa deve ser suficientemente avançada para que os técnicos especialistas possuam um entendimento adequado da tecnologia, processos de desenvolvimento e métodos operacionais de seus sistemas de IA, incluindo metodologias transparentes e auditáveis, fontes de dados e procedimentos e documentação de projeto.

Academia de Inteligência Artificial de Pequim

A Pesquisa e Desenvolvimento da Inteligência Artificial deve adotar abordagens éticas de design para tornar o sistema confiável. Isso pode incluir, mas não se limita a: tornar o sistema o mais justo possível, reduzir possíveis discriminações e

preconceitos, melhorar sua transparência, prover explicação e previsibilidade e tornar o sistema mais rastreável, auditável e responsável.

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

Deve haver transparência e divulgação responsável em torno dos sistemas de Inteligência Artificial para garantir que as pessoas entendam os resultados baseados em IA e possam desafiá-los.

Google

Fornecer transparência apropriada e controle do uso de dados. Este princípio é citado dentro de Privacidade & Segurança.

Microsoft

Os sistemas de Inteligência Artificial devem ser compreensíveis.

Algumas considerações

A Comissão Europeia e a Academia de Inteligência Artificial de Pequim abordam os seis princípios listados nesse mapeamento e as duas iniciativas detalharam bem cada princípio de Inteligência Artificial analisado. A Microsoft também aborda todos os princípios listados, porém é bastante concisa, sem detalhá-los.

Nesse sentido, é interessante observar que a Comissão Europeia, órgão executivo da União Europeia, politicamente independente, é composta por uma equipe de 28 Comissários (um de cada país da UE). Já a Academia de Inteligência Artificial de Pequim é formada por universidades e empresas chinesas. Analisa-se, portanto, que as duas iniciativas são compostas por diversas entidades, o que pode ter trazido maior riqueza e detalhamento aos princípios.

O princípio de Equidade (*Fairness*) é abordado pelas seis iniciativas analisadas. A Comissão Europeia traz maior riqueza de detalhes sobre esse princípio e propõe uma divisão entre equidade substantiva e procedural. Afirma que a primeira vista o compromisso de garantir uma distribuição equitativa e justa de benefícios e custos, assim como garantir que indivíduos e grupos estejam livres de preconceitos injustos, discriminação e estigmatização, enquanto a segunda busca reparação efetiva contra as decisões tomadas pelos sistemas de Inteligência Artificial e pelos humanos que os operam.

Ainda sobre equidade, a Academia de Inteligência Artificial de Pequim sugere adotar abordagens éticas de design para tornar o sistema confiável. A OCDE acrescenta que os sistemas de IA devem incluir salvaguardas apropriadas - por exemplo, possibilitando a intervenção humana sempre que necessário - para garantir uma sociedade justa.

Sobre o princípio de Confiabilidade e Segurança (*Reliability & Safety*), as seis iniciativas analisadas explicitam esse princípio. Todas afirmam que os sistemas de Inteligência Artificial devem ter um desempenho confiável e seguro. A Comissão Europeia detalha que devem proteger a dignidade humana, bem como a integridade mental e física. Além de enfatizar que os sistemas de IA e os ambientes em que operam devem ser seguros, tecnicamente robustos e deve-se garantir que não estejam abertos ao uso malicioso. O Departamento de Defesa Norte-americano acrescenta que a segurança, a proteção e a robustez de tais sistemas devem ser testadas e garantidas por todo o ciclo de vida desse domínio de uso.

A dimensão do Impacto Social (*Social Impact*) é analisada por cinco organizações, apenas o Departamento de Defesa Norte-americano não especifica esse princípio. A Comissão Europeia afirma que os sistemas de IA não devem injustificadamente subordinar, coagir, enganar, manipular, condicionar ou agrupar humanos. Em vez disso, eles devem ser projetados para aumentar, complementar e capacitar as habilidades cognitivas, sociais e culturais humanas.

A Academia de Inteligência Artificial de Pequim acrescenta sobre o Impacto Social que a IA deve ser projetada e desenvolvida para promover o progresso da sociedade e da civilização humana, promover o desenvolvimento sustentável da natureza e da sociedade, beneficiar toda a humanidade e o meio ambiente e melhorar o bem-estar da sociedade e da ecologia. A Google diz que é necessário disponibilizar informações precisas e de alta qualidade usando a IA, de forma a respeitar as normas culturais, sociais e legais nos países em que opera.

Todas as iniciativas mapeadas citam o princípio Responsabilidade (*Accountability*). De modo geral afirmam que é necessário criar mecanismos para garantir a responsabilização e a prestação de contas dos sistemas de IA e seus resultados, antes e depois de seu desenvolvimento, implantação e uso. A Academia de Inteligência artificial de Pequim acrescenta que os pesquisadores e desenvolvedores de IA devem ter considerações suficientes para os possíveis impactos e riscos éticos, legais e sociais trazidos por seus produtos e tomar ações concretas para reduzi-los e evitá-los.

Das seis iniciativas mapeadas, apenas quatros explicitam o princípio de Privacidade & Segurança (*Privacy & Security*). A Comissão Europeia afirma que inclui respeito pela qualidade da privacidade e integridade dos dados e acesso aos dados. Além de estar intimamente ligada ao princípio de prevenção de danos. A Academia de Inteligência Artificial de Pequim acrescenta que privacidade,

dignidade, liberdade, autonomia e direitos humanos devem ser suficientemente respeitados. A Google diz que deve incorporar princípios de privacidade no desenvolvimento e uso de nossas tecnologias de IA. E a Microsoft que os sistemas de Inteligência Artificial devem ser seguros e respeitar a privacidade.

Sobre o princípio de Transparência (Transparency), apenas a Google não especifica, mas menciona dentro de Privacidade & Segurança. De maneira geral, as iniciativas afirmam que os processos precisam ser transparentes. A Comissão Europeia enfatiza que as capacidades e o objetivo dos sistemas de IA devem ser comunicados abertamente e as decisões - na medida do possível - explicáveis para os afetados direta e indiretamente. O Departamento de Defesa Norte-americano acrescenta que devem incluir metodologias transparentes e auditáveis, fontes de dados e procedimentos e documentação de projeto. E a Academia de Inteligência Artificial de Pequim que é preciso adotar abordagens éticas de design para tornar o sistema confiável.

Verificamos, portanto, que três princípios: Equidade, Confiabilidade & Segurança e Responsabilidade são abordados pelas seis iniciativas mapeadas. Os princípios de Impacto Social e Transparência são explicitados por cinco iniciativas. E apenas o princípios de Privacidade & Segurança é detalhado por quatro das seis iniciativas mapeadas.

Referências

Em ordem alfabética por iniciativa

Academia de Inteligência Artificial de Pequim

Beijing AI Principles
Universidades Chinesas e setor privado
https://www.baai.ac.cn/blog/beijing-ai-principles

Comissão Europeia

Building Guidelines for Trustworthy AI
European Commission (High Level Group on Artificial Intelligence)
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Departamento de Defesa Norte-americano

AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense
Defense Innovation Board

https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Google

Google AI Principles
Google
https://www.blog.google/technology/ai/ai-principles/

Microsoft

Microsoft AI Principles
Microsoft
https://www.microsoft.com/en-us/ai/our-approach-to-ai

Organização para a Cooperação e Desenvolvimento Econômico (OCDE)

OECD (Organisation for Economic Co-operation and Development) Principles on AI
OECD (Organisation for Economic Co-operation and Development)
https://www.oecd.org/going-digital/ai/principles/