UNIVERSIDADE DE SÃO PAULO FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE FILOSOFIA PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

Joon Moon

As regras das leis para humanos, não humanos e trans-humanos

(Código de Ur-Nammu, Três leis da Robótica e Princípios de Asilomar)

[versão corrigida]

São Paulo

2023

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS

ENTREGA DO EXEMPLAR CORRIGIDO DA DISSERTAÇÃO/TESE

Termo de Anuência do (a) orientador (a)

Nome do (a) aluno (a): Joon Moon_____

Data da defesa: 06_/10_/2023_

Nome do Prof. (a) orientador (a): Marcos Barbosa de Oliveira_____

Nos termos da legislação vigente, declaro **ESTAR CIENTE** do conteúdo deste **EXEMPLAR CORRIGIDO** elaborado em atenção às sugestões dos membros da comissão Julgadora na sessão de defesa do trabalho, manifestando-me **plenamente favorável** ao seu encaminhamento ao Sistema Janus e publicação no **Portal Digital de Teses da USP**.

São Paulo, 29____/11_____/2023_____

Prof. Dr. Marcos Barbosa de Oliveira

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

M731r

Moon, Joon
As regras das leis para humanos, não humanos e transhumanos (Código de Ur-Nammu, Três leis da Robótica e Princípios de Asilomar) / Joon Moon; orientador Marcos Barbosa Oliveira - São Paulo, 2023.
206 f.

Dissertação (Mestrado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Filosofia. Área de concentração: Filosofia.

1. FILOSOFIA DA CIÊNCIA. 2. FILOSOFIA DO DIREITO. 3. ÉTICA. 4. INTELIGÊNCIA ARTIFICIAL. 5. ROBÓTICA. I. Oliveira, Marcos Barbosa, orient. II. Título.

RESUMO

O avanço da Inteligência Artificial é paralelamente acompanhado pela formulação de princípios regulatórios que visam seu controle. O que se revela intrigante nesta reação da sociedade diante das possíveis ameaças da Inteligência Artificial é o papel de agente capaz de obedecer a tais regras que é atribuído a ela. Pela primeira vez na história da nomologia, uma entidade não humana recebe tal designação. Apesar do ineditismo deste acontecimento, ele ainda não recebeu a devida atenção e estudo. Neste trabalho, esses princípios serão analisados sob os seus aspectos mais significativos, como as condições necessárias para sua efetivação e possíveis consequências. As leis da sociedade e os princípios da Inteligência Artificial serão representados por notações simbólicas como metodologia desta análise.

Palavras-chaves: Ética da Inteligência Artificial, As três leis da robótica de Asimov, Princípios de Asilomar, Robôs, Inteligência Artificial.

ABSTRACT

The advancement of Artificial Intelligence is parallelly accompanied by the formulation of regulatory principles aimed at controlling it. What is intriguing in this society's reaction to the possible threats of Artificial Intelligence is the role of agent capable of obeying such rules that is attributed to it. For the first time in the history of nomology, a non-human entity is given such a designation. Despite the uniqueness of this event, it has not yet received due attention and study. In this work, these principles will be analyzed in their most significant aspects, such as the necessary conditions for their effectiveness and possible consequences. The laws of society and the principles of Artificial Intelligence will be represented by symbolic notations as a methodology for this analysis.

Keywords: Ethics of Artificial Intelligence, Asimov's Three Laws of Robotics, Principles of Asilomar, Robot, Artificial Intelligence

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

PIA Princípios da Inteligência Artificial

H Seres Humanos

R Robôs e IA

N/D Natureza e/ou Deus

a Ações contidas nas leis

SUMÁRIO

INTRODUÇÃO

> Nenhum fato social, humano ou espiritual é tão importante quanto o fato da técnica no mundo moderno. E ainda assim, nenhum assunto é tão pouco compreendido. Vamos tentar estabelecer alguns pontos de orientação para situar o fenômeno técnico. (ELLUL. 1964. p.3) [1]

É dispensável enfatizar a relevância da Inteligência Artificial (doravante IA) na sociedade atual. A IA está presente em diversos setores com um número crescente de aplicações tais como diagnósticos de imagens na medicina, veículos autônomos, tecnologia de reconhecimento facial, sugestões de todos os tipos para compras, aplicações financeiras, filmes, músicas, notícias, relacionamentos, candidatos nas eleições, entre outros. Esta crescente participação gera discussões acerca dos seus efeitos na sociedade. Mesmo no campo da ciência, as alterações observadas no ensaio de Hacking para o quinquagésimo aniversário do livro *A Estrutura das Revoluções Científicas* de Kuhn em 2012 - segundo o qual, experiências em laboratórios são cada vez mais substituídas pela simulação computacional, tornam-se ainda mais profundas com a IA.

> Hoje, o momento é o das leis da biotecnologia. [...] Adicione a isso a ciência da informação. Adicione aquilo que o computador fez para a prática da ciência. Até mesmo o experimento não é mais o que era, porque ele tem sido modificado e, em certa medida, substituído por simulação computacional. E todos sabem que o computador mudou a comunicação. Em 1962 os resultados científicos eram anunciados em encontros, em seminários especiais, em *preprints* e depois em artigos publicados em revistas especializadas. Hoje, o modo primeiro de publicação é o arquivo eletrônico. (HACKING. 2017. p.12*)* [2]

O combate à Covid-19 é um bom exemplo de utilização da IA com atuação em principais etapas como o rastreamento de contatos[3] e o desenvolvimento de vacinas[4]. O ChatGPT3 - processador de linguagem de Inteligência Artificial da OpenAI - lançado para uso

[1] ELLUL, Jacques. *The Technological Society*. 1964. Tradução nossa.
[2] HACKING, Ian. *Ensaio introdutório da edição comemorativa dos 50 anos da publicação – A Estrutura das Revoluções científicas.* Kuhn. Thomas. Ed. Perspectiva. São Paulo. 2017
[3] NEELIMA Arora et al. *The role of artificial intelligence in tackling COVID-19*. National Library of Medicine. 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7692869>. Acesso em dezembro de 2022.
[4] Edison Ong, Mei U Wong,Anthony Huffman, Yongqun He. COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning. National Library of Medicine. 2020. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7350702>. Acesso em dezembro de 2022.

público no final de 2022, vem causando grande alvoroço na sociedade. Além destes avanços visíveis, IA também se desenvolve continuadamente de forma imperceptível e até incompreensível para não especialistas, o que não impede a sua ampliação para novas áreas.

Uma outra forma de verificar a relevância da IA na sociedade é por meio das iniciativas normativas de controle da IA. O desenvolvimento da IA suscita discussões em várias questões, tais como segurança e transparência, viés e discriminação, impacto no emprego, privacidade, uso militar, ampliação da desigualdade, produção e disseminação de mentiras, ameaça da perda de controle sobre seu desenvolvimento, entre outros. A principal manifestação da sociedade diante dos benefícios e também das ameaças potenciais da IA tem sido a elaboração de princípios regulatórios. Os termos utilizados para denominar tais normas são leis, princípios, preceitos, regras, declarações (*declaration*, *charter*) e orientações (*guidelines*). Neste trabalho, utilizaremos Princípios da Inteligência Artificial (doravante PIA) para nos referirmos às normas dessas iniciativas específicas de controle de robôs e IA. Também usaremos os termos leis e princípios para tratar dessas normas de forma mais genérica.

Podemos destacar como exemplo dois PIA anunciados no biênio de 2016 e 2017: os *Tenets of Partnership on AI* em 2016 e os 23 Princípios de Asilomar em 2017. Entre os envolvidos na primeira iniciativa, temos gigantes da tecnologia como Amazon, Apple, Facebook, Google (DeepMind), IBM e Microsoft. Estas empresas, além de estarem entre as maiores do mundo, também se destacam no desenvolvimento da IA. A conferência de Inteligência Artificial em Asilomar, Califórnia, em 2017, reuniu aproximadamente 2.300 pessoas, incluindo físicos, economistas, filósofos e juristas. Essa foi a maior mobilização na discussão sobre o desenvolvimento seguro da IA até hoje, e nela foram anunciados 23 princípios. Nomes célebres como Stephen Hawking, Saul Perlmutter, Frank Wilczek e Elon Musk são signatários dessa iniciativa. Vale ressaltar que tais iniciativas recentes não são inéditas na história dos Princípios da IA. A iniciativa pioneira foi a das "Três Leis da Robótica" de Isaac Asimov, introduzidas em 1942.[5]

[5] Há uma genealogia onde podemos identificar a influência de anteriores nos posteriores na ideia, formato e termos utilizados. Se desde as três leis de Asimov de 1942 até 2016 tivemos menos de dez leis, apenas no

Os princípios recentemente anunciados, bem como aqueles proclamados ao longo dos anos, expõem as preocupações e os desejos da sociedade face ao avanço tecnológico da IA. Eles não só ilustram isso, como também revelam os aspectos mais visíveis e ocultos da nossa sociedade diante de uma tecnologia que pode trazer tanto benefícios quanto ameaças: conflitos de interesses, expectativas quiméricas, concentração de poder em um número reduzido de empresas e países, aumento da desigualdade, disputa geopolítica, entre outros. Tanto o conteúdo das cláusulas quanto as iniciativas instituídas por trás delas fornecem material que possibilita um diagnóstico da época, desencadeado pelos desafios da tecnologia em novos níveis.

O objetivo principal deste trabalho é responder se os instrumentos normativos, tais como leis, princípios, regras, declarações e orientações, representados pelos PIA, poderiam atender às demandas especificadas, incluindo o objetivo maior de manter o controle humano sobre a IA. Além disso, ou a partir disso, buscamos também definir as condições necessárias para a operacionalização dos PIA.

Para responder a essas questões, analisaremos os PIA desde a sua origem até as publicações mais recentes. Se observarmos os anúncios de PIA que ocorreram nos últimos anos, podemos notar uma corrida competitiva pela publicação de PIA. Atualmente, já é possível falar em proliferação de PIA, uma vez que existem mais de trezentos princípios de diversas iniciativas. O objetivo de obter protagonismo na elaboração de PIA e a consequente corrida podem explicar o cenário atual, que ocorre mesmo sem questionamentos prévios sobre a sua necessidade e funcionalidade. Com base nestas respostas, poderíamos também aprimorá-los ou até mesmo abandoná-los. Apesar de bastante limitado, identificamos algumas reflexões e críticas acerca deste fenômeno, incluindo críticas sobre a proliferação.

Vejamos alguns exemplos:

biênio de 2017 e 2018 quase trinta PIA foram anunciadas, totalizando cerca de 250 “sub- princípios”. Esse pico do surgimento no biênio de 2017 e 2018 foi motivado pelas realizações no campo da IA. Além da mudança do protagonismo da IA no lugar de robô que acelerou a discussão sobre o controle. PIA continuam ser elaboradas e publicadas.

> - Juntamente com o rápido desenvolvimento da tecnologia de inteligência artificial (IA), testemunhamos uma proliferação de documentos de "princípios" destinados a fornecer orientações normativas sobre sistemas baseados em IA. [6]
>
> - Vários Princípios de IA são projetados com diferentes considerações, e nenhum deles pode ser perfeito e completo para todos os cenários. [7]
>
> - Infelizmente, o grande volume de princípios propostos ameaça se tornar esmagador e confuso, apresentando dois problemas potenciais. Ou os vários conjuntos de princípios éticos para IA são semelhantes, levando a repetições e redundâncias desnecessárias, ou, se diferirem significativamente, resultarão em confusão e ambiguidade. O pior resultado seria um "mercado de princípios" onde as partes interessadas podem ser tentadas a "comprar" os mais atraentes. [8]

Frequentemente, um princípio quase idêntico integra diferentes PIA. Em outras ocasiões, um princípio entra em contradição com outro quando se trata de diferentes iniciativas. Como adverte Floridi, o cenário atual poderia evoluir para a formação de um "mercado de princípios".

A quantidade de PIA que temos atualmente já é suficiente para dificultar a análise. É plausível supor que novos princípios surgirão com o passar do tempo, provavelmente de forma mais detalhada, atendendo a exigências específicas, uma vez que os princípios de hoje ainda apresentam formas abrangentes sem grandes distinções. Dessa forma, mesmo um estudo amplo e volumoso sobre os PIA atuais corre o risco de se tornar datado e obsoleto.

Se a competição pelo protagonismo na proliferação de conjuntos de princípios pode ser uma motivação mais visível, certamente temos razões anteriores. A proliferação dos PIA pode ser uma manifestação sintomática da sociedade diante dos desafios impostos pela IA. É possível afirmar que a IA provocou a sociedade de tal forma que sua reação foi por meio de conjuntos de princípios, geralmente de forma ágil em um curto intervalo de tempo. Provavelmente nessa corrida para a publicação de PIA, a competição não é apenas para obter

---

[6] FJELD, Jessica, et al. "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*." Berkman Klein Center for Internet & Society, 2020. Disponível em <https://dash.harvard.edu/handle/1/42160420> Acesso em maio de 2023. Tradução nossa.

[7] YI Zeng, ENMENG Lu, CUNQING, Huangfu. *Linking Artificial Intelligence Principles.* AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019), 2019. Disponível em <https://dash.harvard.edu/handle/1/42160420> Acesso em junho de 2023. Tradução nossa.

[8] FLORIDI, Luciano and COWLS, Josh, *A Unified Framework of Five Principles for AI in Society* (September 20, 2019). Disponível em <https://ssrn.com/abstract=3831321 or http://dx.doi.org/10.2139/ssrn.3831321> Acesso em maio de 2023. Tradução nossa.

protagonismo diante de outras instituições, mas também contra a própria IA. É a tentativa de acompanhar a velocidade de desenvolvimento da IA com conjuntos de princípios.

Em uma cronologia padrão, os acontecimentos precedem às leis regulatórias. O acúmulo de infrações, danos, prejuízos e desvios na sociedade exige uma normatização da ação em questão. No caso dos PIA, a ordem cronológica se inverte, dando espaço para iniciativas que remetem ao conceito do princípio de precaução de Hans Jonas, como se não houvesse tempo a perder diante dos avanços da IA. A sociedade vem sendo alertada pelos especialistas em IA. Após o lançamento do ChatGPT4, um grupo de especialistas publicou uma carta aberta[9] propondo a suspensão de seis meses no seu desenvolvimento.[10] O psicólogo cognitivo e cientista da computação, Geoffrey Hinton, conhecido como o "padrinho" da IA, após o anúncio do seu desligamento do Google, em entrevista ao jornal americano The New York Times, se diz preocupado com o avanço da IA como os chatbots.

> Neste momento, o que estamos vendo são coisas como o GPT-4 superar uma pessoa na quantidade de conhecimento geral que ela tem, e a supera de longe. Em termos de raciocínio, não é tão bom, mas já é capaz de raciocínios simples. [...] E, dado o ritmo de evolução, a expectativa é de que fiquem melhor rapidamente. Então, precisamos nos preocupar com isso. [11]

Nas iniciativas de encontro de pesquisadores e especialistas de diversas áreas para as discussões sobre o controle da IA, quase como uma regra, notamos como o resultado final da iniciativa um anúncio de leis e princípios sem etapas adequadas de discussões preliminares, um salto que enfraquece o próprio resultado da iniciativa.[12] Se não apresenta uma natureza legal, como devemos situar as pseudo leis e princípios atuais?

Os PIA sem dúvida têm suas importâncias. Podem guiar o desenvolvimento da IA refletindo e atendendo aos anseios, preocupações da sociedade, mesmo que não sirva de garantia por si só de cumprimentos das diretrizes expressas, que seria um dos assuntos deste trabalho. Sem tais iniciativas não teríamos iniciado uma discussão acerca de possíveis

[9] Disponível em <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> Acesso em junho de 2023.
[10] Disponível em https://g1.globo.com/tecnologia/noticia/2023/03/29/musk-e-centenas-de-especialistas-pedem-pausa-no-avanco-de-sistemas-com-inteligencia-artificial.ghtml> Acesso em junho de 2023.
[11] Disponível em <https://www.bbc.com/portuguese/articles/cgr1qr06myzo> Acesso em junho de 2023. Tradução nossa.
[12] Há exceções como *AI Code*, uma iniciativa da câmara de lordes do Reino Unido com formação de comitês por assuntos, e conclusão apenas com recomendações sem princípios normativos.

problemas em níveis detalhados envolvendo diversas participações na sociedade. Ainda que possa apresentar desordem e falta de consenso e até contradições, é inegável sua relevância, principalmente no estágio atual de desenvolvimento da IA.

Depois da aparição maciça dos PIA nos últimos anos (2016 até 2019), apesar de ainda não ser numeroso, começamos a verificar surgimento de alguns estudos acerca dos PIA. Selecionamos quatro estudos com metodologias distintas para oferecermos uma visão abrangente sobre os estudos, ressaltando os seus pontos positivos e negativos: "*Policy Paper on the Asilomar Principles on Artificial Intelligence*" [13] da Federação de Cientistas da Alemanha (*Vereinigung Deutscher Wissenschaftler. Doravante VDW*), "*A Unifed Framework of Five Principles for AI in Society*" [14] Luciano Floridi, Josh Cowls, "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*",[15] dos pesquisadores do Berkman Klein Center for Internet & Society, 2020, "*Linking Artificial Intelligence Principles*" [16] dos pesquisadores da chineses Yi Zeng, Enmeng Lu e Cunqing Huangfu.

Os quatro estudos apresentam diferentes abordagens que abrangem, em ordem crescente, a quantidade de princípios abordados: focando em um único princípio, selecionando alguns princípios e reunindo um grande número de princípios para extrair informações relevantes. Por exemplo, o estudo *'Policy Paper on the Asilomar Principles on Artificial Intelligence'*, realizado pela Federação de Cientistas da Alemanha (VDW), se concentra nos 23 princípios de Asilomar, analisando cada cláusula individualmente. Essa abordagem apresenta o benefício de permitir uma análise aprofundada de cada cláusula e seus respectivos aspectos. Entretanto, como limitação, possui uma cobertura que se restringe a um único conjunto de princípios. Apesar disso, o resultado é bastante enriquecedor, trazendo críticas contundentes que são certamente aplicáveis a outros princípios.

[13] Disponível em <https://vdw-ev.de/wp-content/uploads/2019/05/Policy-Paper-on-the-Asilomar-principles-on-Artificial-Inteligence_end.pdf> Acesso em junho de 2023.
[14] Disponível em <https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8> Acesso em junho de 2023.
[15] Disponível em <https://dash.harvard.edu/handle/1/42160420> Acesso em junho de 2023.
[16] Disponível em <https://www.linking-ai-principles.org/> Acesso em junho de 2023.

> Quem determina o que é “bom” quando a tecnologia é praticamente abrangente e afeta a todos - não apenas aqueles que usam um determinado produto baseado em IA? […] É realista que o uso de I.A. pode ser controlado apenas por acordos voluntários entre pesquisadores sem a participação formal das estruturas institucionais existentes e processos do espaço político democraticamente constituído? (ibid. p. 9 e 10. tradução nossa)

Luciano Floridi e Josh Cowls selecionaram seis princípios representativos. O critério adotado baseou-se em tempo (menos de 3 anos), relevância, impacto, reputação, e abrangência de, pelo menos, um escopo nacional. Esses princípios são: *The Asilomar AI Principles (2017), The Montreal Declaration for Responsible AI* (2017), *The General Principles* da segunda versão do *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, IEEE (2017), *The Ethical Principles do Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, EGE (2018), os '*Five overarching principles for an AI code'* do relatório do *UK House of Lords Artificial Intelligence Committee* (2018) e *The Tenets of the Partnership on AI, Partnership on AI* (2018). A partir dessa seleção, os autores realizam uma comparação para deduzir cinco temas principais nos princípios: Beneficência, Não-maleficência, Autonomia, Justiça, Explicabilidade. Segundo eles, estes estariam em concordância com os quatro princípios mais comumente utilizados em Bioética, com a adição da Explicabilidade para a IA.

Os pesquisadores do Berkman Klein Center for Internet & Society e da Academia Chinesa de Ciências se destacam por reunirem um grande número de Princípios da IA para análise. No primeiro estudo, eles selecionaram Direitos Humanos, Promoção de Valores Humanos, Responsabilidade Profissional, Controle Humano da Tecnologia, Justiça e Não Discriminação, Transparência e Explicabilidade, Segurança e Proteção, Responsabilidade e Privacidade como "Categorias de princípios de IA". O infográfico criado a partir dessa seleção facilita a visualização de quais iniciativas se concentram em quais categorias.

Já o artigo dos pesquisadores da Academia Chinesa de Ciências apresenta um resultado similar. A comparação entre as diferentes iniciativas revela os temas defendidos em função da natureza das entidades por trás dos princípios, evidenciando a diferença de prioridades refletida pela origem das iniciativas, seja governamental, empresarial ou acadêmica. Isso nos permite entender quais interesses são refletidos nos princípios, atuando como uma regra subjacente.

> Podemos observar […] que as corporações gostariam de mencionar mais sobre colaboração, mas não tanto sobre segurança e privacidade. Enquanto os governos mencionaram mais sobre segurança, mas não gostariam de mencionar a capacidade de prestação de contas. As corporações podem se beneficiar da colaboração, mas a atmosfera de colaboração pode não ser tão boa quanto a acadêmica, o que pode ser o motivo pelo qual gostariam de mencioná-la. Privacidade e segurança são questões delicadas para as corporações, talvez seja por isso que as corporações não gostariam de mencioná-las. E o governo mencionou o tema da capacidade de prestação de contas significativamente menos do que a academia. (Ibid. P.3. Tradução nossa)

Quanto à metodologia adequada para analisar os PIA, se realizar uma análise exaustiva de todas as leis e suas cláusulas publicadas até hoje seria uma tarefa imensa, possivelmente comprometida pela superficialidade, especialmente considerando que se trata de um fenômeno em curso. Por outro lado, um estudo focado em apenas um conjunto de princípios, como o da VDW, pode ser limitado por não oferecer uma visão geral. Acreditamos que a escolha de PIA influenciará diretamente o resultado da análise. A adoção de uma metodologia apropriada seria necessária para estudo de PIA, e, na medida do possível, sem repetir caminhos já demonstrados pelos esforços anteriores.

Assim, o presente trabalho tratará de uma forma distinta dos estudos citados anteriormente. Em primeiro lugar, faremos uma distinção sobre a natureza dos PIA baseado na identificação do agente, o que não ocorre nos outros estudos. Alguns PIA como as pioneiras leis de Asimov, entre outros, apresentam IA ou robô como agente de uma ação a ser coibida na sua forma negativa, ou responsável por cumprimento em forma positiva. Tomamos como objeto desse estudo, apenas PIA com IA ou robô como responsáveis pela ação na qual recaem o controle normativo.

Sustentamos que estes PIA abordados são verdadeiramente inovadores, não apenas no contexto da história da robótica e inteligência artificial, mas também dentro da nomologia global da sociedade. Outras formas foram deliberadamente excluídas deste estudo, pois podemos argumentar que se constituem como extensões das leis já existentes, nas quais a ação está intrinsecamente vinculada à responsabilidade humana, carecendo assim do ineditismo de a inteligência artificial ou robô enquanto agente autônomo. Portanto, o escopo deste trabalho concentra-se especificamente nos princípios em que a responsabilidade da ação recai sobre a inteligência artificial ou o robô.

Destacamos a singularidade desse evento. Pela primeira vez na história do sistema jurídico, testemunhamos uma ruptura fundamental, onde um ente não humano assume uma posição anteriormente reservada exclusivamente aos seres humanos. Ao subordinar algo aos preceitos legais, estamos, no mínimo, legitimando-o como um agente capaz de exercer a agência atribuída, reconhecendo assim sua capacidade intrínseca para cumprir e declarando sua aptidão inerente.

Em outras palavras, o advento da robótica e inteligência artificial transforma radicalmente o conceito de agente capacitado para cumprir normas. Este rompimento com o formato que permaneceu inalterado por milênios na história da humanidade muitas vezes passa despercebido sem a devida atenção. Destacar a importância desse evento torna-se uma parte fundamental do presente trabalho.

Além dessa distinção, propomos analisar os objetivos dos PIA propostos ao longo dos anos, as alterações que sofreram e o que esses objetivos podem revelar. Para esta análise, introduzimos uma classificação dos PIA em duas fases históricas, que também trazem distinções nas suas formas e direções. De acordo com esta classificação, os PIA apresentam-se de duas maneiras: na primeira fase, denominada negativa tradicional, há o predomínio do termo 'lei', juntamente com o protagonismo dos robôs, estendendo-se de 1942 até 2015. Na segunda fase, a positiva atual, com protagonismo da IA, há o predomínio do termo 'princípio', indo de 2016 até os dias de hoje.

Na sua forma negativa tradicional, os princípios são marcados pela proibição de danos aos seres humanos, seguindo a tradição inaugurada por Asimov - o robô/IA não deve causar dano. Na forma positiva atual, a partir dos Princípios de Parceria em IA de 2016, busca-se garantir ou trazer benefícios para a sociedade. Analisaremos esses objetivos tanto da fase negativa tradicional quanto da positiva atual, buscando compreender o que essas aspirações imputadas aos robôs e à IA revelam. Examinaremos em que medida essas aspirações são inéditas ou já vistas na história humana, o que motiva essas esperanças (afetos), que, apesar da diferença entre a fase negativa tradicional e a positiva atual, apresentam objetivos comuns. Argumentamos que todos os PIA podem ser resumidos em dois objetivos: evitar danos e proporcionar felicidade aos seres humanos.

Como uma classificação arbitrária e abrangente, há PIA que escapam desta delimitação, como os "Três princípios para criar inteligência artificial segura" de Stuart Russell, que, apesar de serem de 2017, estabelecem como primeiro princípio que "O único objetivo do robô é maximizar a realização dos valores humanos".

Além de analisar o papel do agente e na classificação por fases, pretendemos oferecer uma alternativa metodológica. Poderia existir uma fórmula ou sistema de notação capaz de representar todas as leis da robótica e da IA? Ou ainda, esses princípios poderiam ser traduzidos em proposições elementares que facilitariam um estudo mais amplo sem negligenciar as particularidades de cada lei? Essas perguntas sobre a natureza dos PIA nos levaram à ideia da essência, à possibilidade de redução e aos elementos comuns entre os diversos princípios, não apenas em relação ao seu conteúdo, palavras-chave ou iniciativas que os sustentam, como visto em outros estudos.

O sistema de notação adotado para essa tarefa consiste na axiomatização dos PIA. Essa metodologia nos permitiu fazer inferências sobre a natureza dos agentes e suas relações, além da possibilidade de traçar a proposição que justificaria a validade dos PIA. Portanto, essa axiomatização, além de funcionar como um método, também servirá como o fio condutor desta exposição. Trata-se de uma busca pela essência e pela proposição básica, elementar, que possibilita um estudo sobre a natureza dos PIA e também oferece a possibilidade de uma reflexão mais abrangente. Tentativas de redução, a busca pela essência e a forma lógica são partes integrantes da história da filosofia da ciência. As leis físicas ou naturais são elaboradas e apresentadas na forma de axiomas, enquanto que as leis da sociedade geralmente não recorrem a tal recurso.

O trabalho está estruturado em seguinte ordem: Introdução, Parte 1 e Parte 2. Cada Parte apresenta 3 capítulos. A Parte 1 oferece um panorama sobre os PIA e a parte 2 o ineditismo dos PIA e possíveis desdobramentos. Assim, no capítulo 1, os quatorze primeiros PIA são apresentados com breve descrição: “As três leis de Asimov” de 1942 até os “Princípios de Asilomar”. Posicionamos os vinte seis PIA restantes no apêndice para uma leitura mais fluida.

No capítulo 2, apresentamos uma classificação de PIA em duas fases: a primeira fase negativa tradicional com protagonismo de robôs e medo como o principal justificativa para a elaboração, a segunda fase positiva atual com protagonismo da IA, que demonstra uma esperança utópica que historicamente só poderia ser verificada nas religiões e estados idealizados.

No capítulo 3, analisamos os pontos positivos e negativos dos quatro principais estudos sobre PIA de seguintes autores: Federação de Cientistas da Alemana (VDW), Luciano Floridi e Josh Cowls, pesquisadores de Berkman Klein Center for Internet & Society, pesquisadores da Academia Chinesa de Ciências.

Na Parte 2, procuramos oferecer uma alternativa metodológica ao estudo dos PIA. No capítulo 1, introduzimos uma breve história das leis na sociedade desde as primeiras leis da civilização que temos conhecimento como o código Hamurabi. Apresentamos as regras das leis – características essenciais que as leis deveriam possuir para efetividade elaboradas pelos juristas. Em seguida apresentamos a possibilidade de simbolização para as leis jurídicas e demonstramos que os PIA podem ser representados de acordo com este método. As leis jurídicas que são normativas podem ser simbolizadas pelo formato [1 a 2], onde 1 e 2 representam seres humanos tanto como agente como objeto da ação contido na lei simbolizado por [a]. Outra forma desta simbolização é [H a H] que poderia ser considerado como axioma das leis da sociedade onde [H] representa seres humanos. Defendemos que todas as leis seguem este formato fundamental.

No capítulo 2, ressaltamos o ineditismo dos PIA na história da nomologia pelo papel do agente assumido pela IA. No passado já houve advento de uma nova tecnologia que levantou discussões sobre as questões éticas e iniciativas normativas de controle como aconteceu na bioética com genética, tratados para armas nucleares, mas IA apresenta um aspecto inédito que é evidenciado pelos PIA: IA como sujeito da ação. Até hoje apenas o ser humano ocupou o lugar [H].[17] O lugar do segundo [H] foi ocupado por homens (vivos ou mortos), deuses, natureza, animais. Se todas as leis da sociedade podem ser reduzidas em [H^1

[17] Tanto que na sua única exceção, a corporação precisou tomar emprestado e assumem “pessoa” para funcionamento.

a H^2], os PIA podem ser reduzidas em [R a H] onde [R] representa IA e robôs. Assim, o axioma [R a H] seria a proposição básica e elementar dos PIA.[18] Nas leis onde o agente da ação não é [R], permanece na forma [H^1 a H^2]. Destacamos os PIA que seguem o formato [R a H]. A possibilidade de IA ocupar o lugar [H^1] é inédita na história. [R a H] é a ruptura da tradição [H a H]. O que revela o peso de toda tradição sobre esta mudança. A partir do axioma [R a H], analisaremos a agência de [R] e realizabilidade para [H] e [a].

No capítulo 3, apresentamos a condição necessária para o funcionamento dos PIA: [R a H] → [R = H] que seria a condição de validade. Esta condição poderia ser saciada através de dois caminhos: maquinização dos seres humanos e/ou humanização das máquinas. A partir destas condições, para que PIA funcionem, seria necessário criar IA com capacidade de agência no cumprimento de leis como seres humanos e/ou trans humanismo sem perder a capacidade de obediência das leis dos seres humanos. Apresentamos também qual seria a expectativa do [H] que formula PIA, e por último, as considerações finais.

A história humana pode ser vista, de maneira simplificada, através das lentes do medo e da busca pela felicidade, dois poderosos impulsos que têm moldado a civilização. Este binômio atua como o motor propulsor da existência humana, influenciando as ações, aspirações e a forma como os seres humanos se relacionam com o mundo e com outros. Os objetivos dos PIA são inscritos no âmbito da filosofia prática, em sua tentativa de formular leis para evitar danos ou garantir benefícios para a sociedade, misturando as esferas da moral, do direito, da política, além da tecnologia. Argumentamos que esses objetivos são claramente utópicos.

O medo, por exemplo, tem desempenhado um papel crucial na evolução humana. Hobbes argumentou que o medo é o principal motivador do contrato social. No estado de natureza, argumentou Hobbes, o medo da morte violenta nas mãos de outros, a "guerra de todos contra todos", leva as pessoas a formarem sociedades e a se submeterem a um poder soberano. Esta visão reflete o que poderíamos chamar de uma resposta coletiva ao medo, uma maneira de minimizar a ameaça percebida à existência humana e bem-estar. Por outro lado, a busca pela felicidade tem sido uma força igualmente importante na condução da

[18] Mas neste trabalho não serão tratadas as formas de PIA do tipo [H a H] que não é objetivo.

história humana. Aristóteles considerava a felicidade, ou "eudaimonia", o maior bem e o fim último da vida humana.

A ciência e a tecnologia têm desempenhado papéis cruciais tanto na mitigação dos medos como na promoção da busca pela felicidade. Avanços científicos têm permitido compreender e controlar o mundo natural, reduzindo assim muitos dos perigos que antes ameaçavam a sobrevivência humana. Da mesma forma, a tecnologia tem se mostrado um meio poderoso de melhorar a qualidade de vida e facilitar a busca pela felicidade.

Por meio dos PIA, almejamos tanto eliminar os perigos quanto garantir os benefícios para a nossa sociedade. No entanto, é importante salientar que tanto a ciência quanto a tecnologia são ferramentas duplamente afiadas. Enquanto podem aliviar nossos medos e promover a felicidade, também têm o potencial de criar novos temores e desafios.

Temos uma relação conflituosa com a IA. Queremos que ela nos auxilie para nos livrarmos dos perigos e danos que carregamos, garantindo benefícios como revelam os PIA. Mas ao mesmo tempo tememos a própria IA. De forma semelhante que no desenvolvimento da IA houve invernos e primaveras, parece que o fascínio causado pelas novas realizações logo converte-se em medo.

Em maio de 2023, Geoffrey Hinton, um dos pioneiros em IA advertiu em uma entrevista à Reuters [19] que a IA poderia representar uma ameaça "mais urgente" à humanidade do que as mudanças climáticas. Hinton, que recentemente deixou a Google depois de uma década, expressou suas preocupações sobre os riscos potenciais da tecnologia. Ele está entre um número crescente de líderes de tecnologia que expressam preocupações públicas sobre a possível ameaça representada pela IA. Em abril, o CEO do Twitter, Elon Musk, juntou-se a milhares de pessoas assinando uma carta aberta[20] pedindo uma pausa de seis meses no desenvolvimento de sistemas mais poderosos que o recentemente lançado ChatGPT-4 que rapidamente se tornou o aplicativo de crescimento mais rápido da história.

[19] Disponível em <https://www.reuters.com/technology/ai-pioneer-says-its-threat-world-may-be-more-urgent-than-climate-change-2023-05-05/> Acesso em junho de 2023.
[20] Disponível em https://www.safe.ai/statement-on-ai-risk#open-letter Acesso em junho de 2023.

A compreensão da tecnologia se torna ainda mais desafiadora com o advento da IA. Conforme a citação de Ellul no início, na medida do possível, gostaríamos de oferecer alguns “pontos de orientação” para os PIA. A proposta deste trabalho pode, em certa medida, parecer extremamente desafiadora, uma vez que se propõe a abordar uma questão recente, que ainda carece de consolidação, literatura relevante, distanciamento necessário e ainda está em plena atividade. Além disso, este estudo será baseado em um método próprio para tal tarefa.

PARTE I CAPÍTULO 1 - Leis e princípios para robótica e IA

Neste capítulo procuramos apresentar maior número de leis e princípios para robótica e IA que pudemos reunir. Faremos esta apresentação em ordem cronológica. Começamos pelas "Três Leis da Robótica" de Asimov de 1942, a pioneira de todas as leis nesta área, incluímos também "As 10 leis de Osamu Tezuka", geralmente essas leis originadas em ficções são ignoradas em estudos de leis para robôs e IA. Optamos por incluí-las pois oferecem uma visão sobre a origem que são valiosas pela imaginação, ainda mais que, muitos princípios anunciados recentemente, também podem ser vistos como ficções quando algum tipo de controle real entrarem em funcionamento no futuro.

Alguns PIA são bastante extensos; nestes casos, buscamos extrair os princípios fundamentais para exemplificar o conteúdo geral (por exemplo, o relatório da Câmara dos Lordes do Reino Unido possui 74 princípios, dos quais destacamos os três primeiros e os quatro últimos). Também incluímos versões preliminares, rascunhos (*draft*), cartas abertas e recomendações, pois acreditamos que esses documentos fornecem um vislumbre interessante do processo de formulação dos PIA. Optamos por reunir os PIA publicados até 2019. Além da quantidade de publicações ter diminuído após esse ano, também começaram a surgir atualizações das publicações anteriores, bem como estudos críticos acerca dos PIA. Incluímos os PIA até os princípios de Asilomar neste capítulo. Concentramos o restante no apêndice.

As três leis da robótica [21]- Isaac Asimov (1942. EUA)

As três leis da robótica foram propostas pelo autor de ficção científica Isaac Asimov em sua obra "*Runaround*". Foi escrito em outubro de 1941 e publicado pela primeira vez na edição de março de 1942 da "*Astounding Science Fiction"*. São consideradas como as primeiras leis para robôs criados na humanidade, mesmo em ficção científica.

1. Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano sofra dano.

[21] ASIMOV, Isaac. (1983) *Machine that think: The Best Science Fiction Stories About Robot and Computer. História de Robô*, Porto Alegre.L&PM. 2010. Adaptação nossa.

2. Um robô deve obedecer às ordens dadas por seres humanos, exceto quando tais ordens conflitam com a Primeira Lei.
3. Um robô deve proteger sua própria existência, desde que essa proteção não conflite com a Primeira ou a Segunda Lei.

As Três Leis de Asimov apresentam formato de hierarquia entre os princípios, sendo que a primeira tem domínio sobre a segunda e assim por diante. Asimov também introduziu uma "Lei Zero" em obras posteriores, que antecede as outras três:

> Lei Zero: Um robô não pode prejudicar a humanidade ou, por inação, permitir que a humanidade sofra algum mal.

Essa Lei Zero adiciona uma camada extra de complexidade, uma vez que requer que os robôs considerem o bem maior para a humanidade como um todo.

10 Leis de Osamu Tezuka [22](1988. Japão)

Osamu Tezuka foi um famoso desenhista de mangá japonês e criador de personagens icônicos como Astro Boy, Kimba, o Leão Branco e Black Jack. Ele desenvolveu suas próprias filosofias em relação à criação de mangás e animações. Abaixo, segue as 10 leis de Osamu Tezuka:

1. Os robôs devem servir à humanidade.
2. Os robôs nunca devem matar ou ferir humanos.
3. Os robôs devem chamar o humano que os criou de “pai”.
4. Os robôs podem fazer qualquer coisa, exceto dinheiro.
5. Os robôs nunca devem ir para o exterior sem permissão.
6. Robôs masculinos e femininos nunca devem trocar de função.
7. Os robôs nunca devem mudar sua aparência ou assumir outra identidade sem permissão.
8. Robôs criados como adultos nunca devem agir como crianças.
9. Os robôs não devem montar outros robôs que tenham sido descartados por humanos.
10. Os robôs nunca devem danificar casas ou ferramentas humanas.

[22] Disponível em <https://akikok012um1.wordpress.com/japans-ten-principles-of-robot-law/> Acesso em maio de 2023. Tradução nossa.

Declaração Mundial de Robôs Feira Internacional de Robôs de Fukuoka[23] (2004. Japão)

A Declaração Mundial de Robôs foi anunciada durante a International Robot Exhibition (iREX) em Fukuoka, Japão, em 2004. O Japão tem sido um líder global em robótica, com o governo e as empresas investindo fortemente na pesquisa e no desenvolvimento de novas tecnologias robóticas. Como parte deste foco, o Japão tem procurado não apenas inovar, mas também estabelecer diretrizes e princípios para o uso responsável e ético dos robôs. A Declaração Mundial de Robôs surgiu como uma tentativa de expressar uma visão positiva e ambiciosa para o futuro da robótica. Embora não seja um conjunto formal de diretrizes éticas, essa declaração tem sido influente ao moldar o desenvolvimento e a implementação da robótica no Japão e em outras partes do mundo. A declaração enfatizava três funções principais dos robôs.

1. Robôs de próxima geração serão parceiros que coexistem com seres humanos
2. Os robôs da próxima geração ajudarão os seres humanos tanto física quanto psicologicamente
3. Os robôs da próxima geração contribuirão para a realização de uma sociedade segura e pacífica

Carta de Ética para Robôs [24](2007. Coreia do Sul)

Em meados da primeira década do século XXI, a Coreia do Sul emergiu como um dos principais players no campo da robótica, com o governo coreano investindo pesadamente em pesquisa e desenvolvimento nesta área. Reconhecendo que o avanço da robótica poderia trazer problemas éticos complexos, o Ministério da Informação e Comunicação da Coreia do Sul (MIC) decidiu em 2007 formular uma Carta de Ética para Robôs também conhecida como "Robot Ethics Charter". O MIC trabalhou com especialistas em ética, tecnologia e direito para

[23] Disponível em <http://prw.kyodonews.jp/prwfile/prdata/0370/release/200402259634/index.html> Acesso em out. de 2019. Tradução nossa.
[24] Disponível em <http://www.roboethics.org/icra2007/contributions/slides/Shim_icra%2007_ppt.pdf> Acesso em out. de 2019. Tradução nossa.

desenvolver a Carta, que foi uma das primeiras tentativas do mundo de estabelecer um conjunto formal de diretrizes éticas para a robótica:

> Capítulo 1 (Objetivo): O objetivo da Carta de Ética do Robô é identificar padrões éticos centrados no homem para a coexistência entre humanos e robôs.
>
> Capítulo 2 (Princípios de humanos e robôs): Humanos e robôs devem aderir à dignidade da vida, da inteligência e da ética em engenharia.
>
> Capítulo 3 (Ética Humana): Os seres humanos devem sempre julgar e tomar boas decisões ao fazer e usar robôs.
>
> Capítulo 4 (Ética do robô): Os robôs não devem ferir os seres humanos como amigos, ajudantes e companheiros que obedecem a seus comandos.
>
> Capítulo 5 (Ética do fabricante): O fabricante do robô tem o dever de fabricar robôs que protegem a dignidade humana, reciclar robôs e proteger informações.
>
> Capítulo 6 (Ética do usuário): Os usuários de robôs devem respeitar os robôs como amigos humanos e proibir modificações ilegais ou abuso de robôs.
>
> Capítulo 7: Promessa de implementação: Os governos e municípios devem implementar medidas efetivas para implementar o espírito desta declaração.

As três leis da robótica responsável [25] - IEEE (2009. Mundial)

As Três Leis da Robótica Responsável da IEEE (Instituto de Engenheiros Elétricos e Eletrônicos) foram inspiradas pelas Três Leis da Robótica de Isaac Asimov. Em 2009, a IEEE estabeleceu um grupo de trabalho, conhecido como a Iniciativa sobre Ética da Autonomia e Inteligência Sistêmica (EADSI), para estabelecer diretrizes éticas para a robótica e criou as Três Leis da Robótica Responsável:

> 1. Um humano não pode implantar um robô sem que o sistema de trabalho humano-robô atenda aos mais altos padrões legais e profissionais de segurança e ética.
> 2. Um robô deve responder aos humanos conforme apropriado para suas funções.
> 3. Um robô deve ser dotado de autonomia situada suficiente para proteger sua própria existência, desde que tal proteção forneça uma transferência suave de controle que não entre em conflito com a Primeira e a Segunda Leis.

[25] Disponível em < https://ieeexplore.ieee.org/document/5172885> Acesso em out. de 2019. Tradução nossa.

Cinco princípios éticos para a robótica EPSRC/AHRC [26](2011. Reino Unido)

Em 2010, a *Engineering and Physical Sciences Research Council* (EPSRC) e a *Arts and Humanities Research Council* (AHRC) do Reino Unido, reuniram um grupo de especialistas de várias disciplinas para discutir o futuro da robótica. Isso incluía pesquisadores, engenheiros, filósofos, sociólogos, psicólogos e outros, todos com o objetivo de identificar os principais problemas éticos e sociais que poderiam surgir com o avanço da robótica. Após um período intenso de debates e revisões, os Cinco Princípios da Robótica foram estabelecidos em 2011.

> 1. Os robôs são ferramentas multiuso. Os robôs não devem ser projetados exclusiva ou principalmente para matar ou ferir humanos, exceto no interesse da segurança nacional.
>
> 2. Humanos, não robôs, são agentes responsáveis. Os robôs devem ser projetados; operado na medida do possível para cumprir as leis existentes e direitos e liberdades fundamentais, incluindo privacidade.
>
> 3. Os robôs são produtos. Eles devem ser projetados usando processos que garantam sua segurança e proteção.
>
> 4. Os robôs são artefatos manufaturados. Eles não devem ser projetados de maneira enganosa para explorar usuários vulneráveis; em vez disso, sua natureza de máquina deve ser transparente.
>
> 5. Deve ser atribuída a pessoa com responsabilidade legal por um robô.

Carta de Ética de Robôs [27](2012. Coreia do Sul)

Esta Carta foi elaborada para prevenir males sociais que possam surgir de medidas sociais e legais inadequadas para lidar com robôs na sociedade.

> Parte 1: Padrões de Fabricação
>
> a) Os fabricantes de robôs devem garantir que a autonomia dos robôs que eles projetam seja limitada; caso seja necessário, deve ser sempre possível para um ser humano assumir o controle de um robô.

[26] Disponível em <https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics> Acesso em out. de 2019. Tradução nossa.

[27] Disponível em <https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012/> Acesso em maio de 2023. Tradução nossa.

b) Os fabricantes de robôs devem manter padrões rígidos de controle de qualidade, tomando todas as medidas razoáveis para garantir que o risco de morte ou lesão do usuário seja minimizado e a segurança da comunidade garantida.

c) Os fabricantes de robôs devem tomar medidas para garantir que o risco de danos psicológicos aos usuários seja minimizado. 'Dano psicológico', neste sentido, inclui qualquer probabilidade de o robô induzir comportamentos antissociais ou sociopáticos, depressão ou ansiedade, estresse e, particularmente, vícios (como vício em jogos de azar).

c) Os fabricantes de robôs devem garantir que seu produto seja claramente identificável e que essa identificação seja protegida contra alterações.

d) Os robôs devem ser projetados de forma a proteger os dados pessoais, por meio de criptografia e armazenamento seguro.

e) Os robôs devem ser projetados para que suas ações (tanto online quanto no mundo real) sejam rastreáveis o tempo todo.

f) O projeto do robô deve ser ecologicamente sensível e sustentável.

Parte 2: Direitos e responsabilidades dos usuários/proprietários

Seg. 1: Direitos e Expectativas dos Proprietários e Usuários

i) Os proprietários têm o direito de poder assumir o controle de seu robô.

ii) Proprietários e usuários têm o direito de usar seu robô sem risco ou medo de danos físicos ou psicológicos.

iii) Os usuários têm direito à segurança de seus dados pessoais e outras informações confidenciais.

iv) Proprietários e usuários têm o direito de esperar que um robô execute qualquer tarefa para a qual tenha sido explicitamente projetado (sujeito à Seção 2 desta Carta).

Seg. 2: Responsabilidades dos Proprietários e Usuários

Esta Carta reconhece o direito do usuário de utilizar um robô da maneira que achar melhor, desde que esse uso permaneça 'justo' e 'legal' dentro dos parâmetros da lei. Como tal:

i) Um usuário não deve usar um robô para cometer um ato ilegal.

ii) Um usuário não deve usar um robô de forma que possa ser interpretada como causando danos físicos ou psicológicos a um indivíduo.

iii) Um proprietário deve tomar "precauções razoáveis" para garantir que seu robô não represente uma ameaça à segurança e ao bem-estar dos indivíduos ou de sua propriedade.

Seg. 3: Os seguintes atos são uma ofensa sob a lei coreana:

i) Danificar ou destruir deliberadamente um robô.

ii) Por negligência grosseira, permitir que um robô sofra algum mal.

iii) É uma ofensa menor, mas ainda assim grave, tratar um robô de uma forma que possa ser interpretada como deliberada e excessivamente abusiva.

Parte 3: Direitos e responsabilidades dos robôs

Seg. 1: Responsabilidades dos Robôs

i) Um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal.

ii) Um robô deve obedecer a quaisquer ordens que lhe sejam dadas por seres humanos, exceto quando tais ordens entrarem em conflito com a Parte 3, Seção 1, subseção "i" desta Carta.

iii) Um robô não deve enganar um ser humano.

Seção 2: Direitos dos Robôs

De acordo com a lei coreana, os robôs têm os seguintes direitos fundamentais:

i) O direito de existir sem medo de ferimentos ou morte.

ii) O direito de viver uma existência livre de abusos sistemáticos.

Oito leis da robótica de Shinpo Fumio[28] - Keio University (2015. Japão)

Em 2015, o professor da Universidade de Keio, Shinpo Fumio, propôs oito preceitos da lei de robôs, que faziam referência aos princípios de privacidade da OCDE.

1) A humanidade em primeiro lugar - os robôs não podem prejudicar ou se tornar pessoas.

2) Obediência à ordem — devem seguir ordens humanas e estar sujeitos a controle.

3) Sigilo e privacidade — os robôs devem ser projetados para preservar o sigilo das informações que coletam.

4) Limitação de uso – os robôs devem ser limitados ao uso pretendido e não podem ser usados para prejudicar humanos.

5) Medidas de segurança.

6) Abertura e transparência — o projeto e o uso do robô devem ser verificáveis.

7) Participação individual — os indivíduos devem participar da criação de regras que regem os robôs, e os robôs não devem reger os indivíduos.

[28] Disponível em <https://www.japantimes.co.jp/community/2019/03/06/issues/robot-rights-asimov-tezuka/#.XdxkPK8nb3g> Acesso em out. de 2019. Tradução nossa.

8) Responsabilidade — deve haver regras de responsabilidade por danos causados por robôs.

As lei de Satya Nadella [29] (2016. EUA)

Satya Nadella, CEO da Microsoft, abordou a questão da robótica e da IA durante uma entrevista [30]na conferência DLD em Munique, Alemanha, em 2016. Nessa entrevista, Nadella introduziu seus próprios princípios para a IA, que têm sido referidos como as "Leis de Satya Nadella".

1. "IA deve ser projetada para ajudar a humanidade", o que significa que a autonomia humana precisa ser respeitada.

2. "IA deve ser transparente", o que significa que os humanos devem saber e ser capazes de entender como funcionam.

3. "IA deve maximizar a eficiência sem destruir a dignidade das pessoas".

4. "IA deve ser projetada para privacidade inteligente", o que significa que ela ganha confiança ao proteger suas informações.

5. "IA deve ter responsabilidade algorítmica para que os humanos possam desfazer danos não intencionais".

6. "IA deve se proteger contra o preconceito" para que não discrimine as pessoas.

Princípios de Parceria em IA [31](2016. EUA)

Os Princípios de Parceria em IA, também conhecidos como Parceria sobre IA, surgiram em 2016 como uma colaboração entre várias das principais empresas de tecnologia do mundo, incluindo Amazon, DeepMind (Google), Facebook, IBM e Microsoft. O objetivo dessa parceria era garantir que a IA e a aprendizagem automática fossem usadas para beneficiar as pessoas e a sociedade. A Parceria sobre IA foi formada em reconhecimento à rápida expansão e ao impacto potencial da IA em quase todos os aspectos da vida. As empresas envolvidas perceberam a necessidade de um conjunto de diretrizes éticas para ajudar a moldar o desenvolvimento e a implementação desta tecnologia poderosa.

[29] Disponível em <https://qz.com/720424/microsoft-ceo-satya-nadella-has-10-commandments-for-how-ai-and-humans-should-act/> Acesso em out. de 2019. Tradução nossa.
[30] Disponível em <https://news.microsoft.com/europe/2017/01/13/satya-nadella-to-talk-ai-at-dld-digital-conference/> Acesso em out. de 2019. Tradução nossa.
[31] Disponível em <https://www.partnershiponai.org/tenets/> Acesso em out. de 2019. Tradução nossa.

1. Procuraremos garantir que as tecnologias de IA beneficiem e capacitem o maior número possível de pessoas.

2. Educaremos e ouviremos o público e engajaremos ativamente as partes interessadas para buscar seu feedback sobre nosso foco, informá-los sobre nosso trabalho e responder às suas perguntas.

3. Estamos comprometidos em abrir a pesquisa e o diálogo sobre as implicações éticas, sociais, econômicas e legais da IA.

4. Acreditamos que os esforços de pesquisa e desenvolvimento de IA precisam ser ativamente engajados e responsáveis perante uma ampla gama de partes interessadas.

5. Vamos nos envolver e ter representação das partes interessadas na comunidade empresarial para ajudar a garantir que as preocupações e oportunidades específicas do domínio sejam compreendidas e abordadas.

6. Trabalharemos para maximizar os benefícios e enfrentar os possíveis desafios das tecnologias de IA, por meio de:

1. Trabalhando para proteger a privacidade e a segurança dos indivíduos.

2. Esforçar-se para compreender e respeitar os interesses de todas as partes que possam ser afetadas pelos avanços da IA.

3. Trabalhar para garantir que as comunidades de pesquisa e engenharia de IA permaneçam socialmente responsáveis, sensíveis e envolvidas diretamente com as possíveis influências das tecnologias de IA na sociedade em geral.

4. Garantir que a pesquisa e a tecnologia de IA sejam robustas, confiáveis, confiáveis e operem dentro de restrições seguras.

5. Opor-se ao desenvolvimento e uso de tecnologias de IA que violem convenções internacionais ou direitos humanos e promover salvaguardas e tecnologias que não prejudiquem.

7. Acreditamos que é importante que a operação dos sistemas de IA seja compreensível e interpretável pelas pessoas, para fins de explicação da tecnologia.

8. Nós nos esforçamos para criar uma cultura de cooperação, confiança e abertura entre cientistas e engenheiros de IA para nos ajudar a alcançar melhor esses objetivos.

Três princípios para criar inteligência artificial segura (ou IA compatível com humanos) - Stuart Russell[32] (2017. EUA)

Stuart Russell é um renomado professor de ciência da computação na Universidade da Califórnia, Berkeley, e um dos principais especialistas em IA. Os seus três princípios

[32] Disponível em <https://www.newworldai.com/three-principles-for-creating-safer-artificial-intelligence-stuart-russell/> Acesso em maio de 2023. Tradução nossa.

claramente demonstram inspiração nas Três leis de Asimov em conteúdo, quantidade e concatenação hierárquica entre os princípios, além do termo robô, mesmo quando já se utilizava largamente o termo IA.

1. O único objetivo do robô é maximizar a realização dos valores humanos.

2. O robô está inicialmente incerto sobre quais são esses valores

3. O comportamento humano fornece informações sobre os valores humanos

Três Regras para Sistemas de Inteligência Artificial - CEO do Allen Institute for Artificial Intelligence[33] (2017. EUA)

As Três Regras para Sistemas de Inteligência Artificial foram propostas pelo CEO do Allen Institute for Artificial Intelligence (AI2), Oren Etzioni. O AI2 é um instituto de pesquisa em inteligência artificial sediado em Seattle, nos Estados Unidos. Oren Etzioni é um renomado cientista da computação e empreendedor no campo da IA. Ele tem sido uma figura proeminente na comunidade de pesquisa em IA e é reconhecido por suas contribuições significativas para o avanço da tecnologia.

1. Um sistema IA deve estar sujeito a toda a gama de leis que se aplicam ao seu operador humano.

2. Um sistema IA deve revelar claramente que não é humano.

3. Um sistema IA não pode reter ou divulgar informações confidenciais sem a aprovação explícita da fonte dessas informações.

Regras de Direito Civil sobre Robótica [34](2017. UE)

O Relatório com recomendações à Comissão de Normas de Direito Civil sobre Robótica de 2017 da União Europeia foi resultado de um trabalho realizado por especialistas e acadêmicos que buscaram explorar os desafios legais e éticos apresentados pelo avanço da robótica. O relatório teve como objetivo fornecer orientações e recomendações para a

[33] Disponível em <https://e-discoveryteam.com/2017/09/17/new-draft-principles-of-ai-ethics-proposed-by-the-allen-institute-for-artificial-intelligence-and-the-problem-of-election-hijacking-by-secret-ais-posing-as-real-people/> Acesso em out. de 2019. Tradução nossa.

[34] Disponível em <http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html> Acesso em maio de 2023. Tradução nossa.

elaboração de normas e regulamentações que abordassem questões relacionadas à robótica e à IA no contexto do direito civil. Ele foi concebido como um guia para auxiliar governos da UE, legisladores e especialistas jurídicos na adaptação do quadro legal existente para lidar com os avanços tecnológicos em curso.

A equipe responsável pelo relatório analisou diversas questões, incluindo responsabilidade civil por danos causados por robôs, privacidade e proteção de dados, direitos dos consumidores, propriedade intelectual e direitos autorais, entre outros tópicos relevantes. Com base em suas análises, foram formuladas uma série de recomendações destinadas a orientar as políticas e práticas legais relacionadas à robótica. Este relatório teve influência em discussões subsequentes sobre a regulação da robótica e da IA em nível nacional e internacional, destacando a importância de uma abordagem legal adequada para lidar com os desafios emergentes trazidos por essas tecnologias disruptivas. Vale destacar que Os Princípios Gerais começa citando as Leis de Asimov.

> Princípios gerais
>
> U. Considerando que as Leis de Asimov devem ser consideradas dirigidas aos designers, produtores e operadores de robôs, incluindo robôs com autonomia incorporada e autoaprendizagem, uma vez que essas leis não podem ser convertidas em código de máquina;
>
> V. Considerando que é necessário um conjunto de regras que regem, em particular, a responsabilidade, a transparência e a prestação de contas, refletindo os valores humanísticos intrinsecamente europeus e universais que caracterizam o contributo da Europa para a sociedade; considerando que essas regras não devem afetar o processo de investigação, inovação e desenvolvimento em robótica;
>
> W. Considerando que a União pode desempenhar um papel essencial no estabelecimento de princípios éticos básicos a respeitar no desenvolvimento, programação e utilização de robôs e IA e na incorporação de tais princípios nos regulamentos e códigos de conduta da União, com o objetivo de moldar a revolução tecnológica para que sirva a humanidade e para que os benefícios da robótica avançada e da IA sejam amplamente compartilhados, evitando, na medida do possível, possíveis armadilhas;
>
> X. Considerando que uma abordagem gradualista, pragmática e cautelosa do tipo defendida por Jean Monnet deve ser adotada pela União no que diz respeito a futuras iniciativas em robótica e IA, de modo a garantir que não sufocamos a inovação;
>
> Y. Considerando que é adequado, tendo em conta o estado de desenvolvimento da robótica e da IA, começar pelas questões de responsabilidade civil;
>
> Princípios éticos

10. Observa que o potencial de capacitação através da utilização da robótica é matizado por um conjunto de tensões ou riscos e deve ser seriamente avaliado do ponto de vista da segurança, saúde e proteção humanas; liberdade, privacidade, integridade e dignidade; autodeterminação e não discriminação e proteção de dados pessoais;

11. Considera que o atual quadro jurídico da União deve ser atualizado e complementado, se for caso disso, mediante princípios éticos orientadores em conformidade com a complexidade da robótica e as suas múltiplas implicações sociais, médicas e bioéticas; considera que é necessário um quadro ético orientador claro, rigoroso e eficiente para o desenvolvimento, a conceção, a produção, a utilização e a modificação de robôs, a fim de complementar as recomendações jurídicas do relatório e o acervo nacional e da União em vigor; propõe, no anexo à resolução, um quadro sob a forma de uma carta que consiste num código de conduta para engenheiros robóticos, num código para comités de ética em investigação na revisão de protocolos robóticos e em licenças de modelo para designers e utilizadores;

12. Destaca o princípio da transparência, ou seja, que deve ser sempre possível fundamentar qualquer decisão tomada com a ajuda da IA que possa ter um impacto substancial na vida de uma ou mais pessoas; considera que deve ser sempre possível reduzir os cálculos do sistema de IA a uma forma compreensível para o ser humano; considera que os robôs avançados devem ser equipados com uma «caixa negra» que registe dados sobre todas as transações efetuadas pela máquina, incluindo a lógica que contribuiu para as suas decisões;

13. Salienta que o quadro ético orientador deve basear-se nos princípios da beneficência, não maleficência, autonomia e justiça, nos princípios e valores consagrados no artigo 2.º do Tratado da União Europeia e na Carta dos Direitos Fundamentais, tais como dignidade humana, igualdade, justiça e equidade, não discriminação, consentimento informado, vida privada e familiar e proteção de dados, bem como sobre outros princípios e valores subjacentes ao direito da União, como não estigmatização, transparência, autonomia, responsabilidade e responsabilidade social, e nas práticas e códigos éticos existentes;

14. Considera que deve ser dada especial atenção aos robôs que representam uma ameaça significativa à confidencialidade devido à sua colocação em esferas tradicionalmente protegidas e privadas e porque podem extrair e enviar dados pessoais e sensíveis;

23 Princípios de Asilomar [35](2017. EUA)

Os 23 Princípios de Asilomar foram estabelecidos durante a Conferência de Asilomar sobre Inteligência Artificial Beneficente, que ocorreu em janeiro de 2017, em Asilomar, Califórnia. O evento foi organizado pelo Instituto Futuro da Vida (FLI) e contou com a participação de especialistas renomados em IA, bem como líderes de pensamento de diversas áreas relacionadas. Durante a conferência, os participantes trabalharam juntos para formular um conjunto de princípios que deveriam orientar a pesquisa e o desenvolvimento em IA. Esses

[35] Disponível em <https://futureoflife.org/ai-principles/> Acesso em maio de 2023. Tradução nossa.

23 princípios são divididos em três categorias: pesquisa em IA, ética e valores em IA e cooperação a longo prazo. Os 23 Princípios de Asilomar tiveram um impacto significativo no campo da IA influenciando muitos pesquisadores, desenvolvedores e legisladores em sua abordagem à IA.

Questões de pesquisa

1) Objetivo da pesquisa: O objetivo da pesquisa de IA deve ser criar não inteligência não direcionada, mas inteligência benéfica.

2) Financiamento de Pesquisa: Os investimentos em IA devem ser acompanhados de financiamento para pesquisas que garantam seu uso benéfico, incluindo questões espinhosas em ciência da computação, economia, direito, ética e estudos sociais, como:

• Como podemos tornar os futuros sistemas de IA altamente robustos, para que eles façam o que queremos sem funcionar mal ou serem hackeados?

• Como podemos aumentar nossa prosperidade por meio da automação, mantendo os recursos e o propósito das pessoas?

• Como podemos atualizar nossos sistemas jurídicos para serem mais justos e eficientes, acompanhar o ritmo da IA e gerenciar os riscos associados à IA?

• Com qual conjunto de valores a IA deve estar alinhada e qual status legal e ético ela deve ter?

3) Vínculo Ciência-Política: Deve haver um intercâmbio construtivo e saudável entre pesquisadores de IA e formuladores de políticas.

4) Cultura de Pesquisa: Uma cultura de cooperação, confiança e transparência deve ser fomentada entre pesquisadores e desenvolvedores de IA.

5) Evitar corridas: as equipes que desenvolvem sistemas de IA devem cooperar ativamente para evitar cortes nos padrões de segurança.

Éticas e valores

6) Segurança: os sistemas de IA devem ser seguros e protegidos durante toda a sua vida operacional, e de forma verificável quando aplicável e viável.

7) Transparência de falha: se um sistema de IA causar danos, deve ser possível determinar o motivo.

8) Transparência Judicial: Qualquer envolvimento de um sistema autônomo na tomada de decisão judicial deve fornecer uma explicação satisfatória auditável por uma autoridade humana competente.

9) Responsabilidade: Designers e construtores de sistemas avançados de IA são partes interessadas nas implicações morais de seu uso, mau uso e ações, com responsabilidade e oportunidade de moldar essas implicações.

10) Alinhamento de valores: sistemas de IA altamente autônomos devem ser projetados para que seus objetivos e comportamentos possam estar alinhados com os valores humanos em toda a sua operação.

11) Valores humanos: Os sistemas de IA devem ser projetados e operados de forma a serem compatíveis com os ideais de dignidade humana, direitos, liberdades e diversidade cultural.

12) Privacidade pessoal: as pessoas devem ter o direito de acessar, gerenciar e controlar os dados que geram, dado o poder dos sistemas de IA de analisar e utilizar esses dados.

13) Liberdade e privacidade: a aplicação de IA a dados pessoais não deve reduzir de forma irracional a liberdade real ou percebida das pessoas.

14) Benefício compartilhado: as tecnologias de IA devem beneficiar e capacitar o maior número possível de pessoas.

15) Prosperidade Compartilhada: A prosperidade econômica criada pela IA deve ser amplamente compartilhada, para beneficiar toda a humanidade.

16) Controle Humano: Os humanos devem escolher como e se desejam delegar decisões aos sistemas de IA, para atingir os objetivos escolhidos pelos humanos.

17) Não subversão: O poder conferido pelo controle de sistemas de IA altamente avançados deve respeitar e melhorar, ao invés de subverter, os processos sociais e cívicos dos quais depende a saúde da sociedade.

18) Corrida Armamentista AI: Uma corrida armamentista em armas autônomas letais deve ser evitada.

Problemas de longo prazo

19) Cuidado de capacidade: Não havendo consenso, devemos evitar fortes suposições sobre os limites superiores das futuras capacidades de IA.

20) Importância: A IA avançada pode representar uma mudança profunda na história da vida na Terra e deve ser planejada e gerenciada com cuidado e recursos adequados.

21) Riscos: Os riscos apresentados pelos sistemas de IA, especialmente os riscos catastróficos ou existenciais, devem estar sujeitos a esforços de planejamento e mitigação compatíveis com seu impacto esperado.

22) Auto-aperfeiçoamento recursivo: Os sistemas de IA projetados para auto-aperfeiçoamento ou auto-replicação recursiva de uma maneira que possa levar a um aumento rápido da qualidade ou quantidade devem estar sujeitos a medidas rígidas de segurança e controle.

23) Bem Comum: A superinteligência só deve ser desenvolvida a serviço de ideais éticos amplamente compartilhados e para o benefício de toda a humanidade, e não de um estado ou organização.

PARTE I CAPÍTULO 2 - CLASSIFICAÇÃO DE PIA POR FASES

A ficção científica é um gênero literário que apresenta uma peculiaridade em relação aos demais. O imaginário da ficção científica, muitas vezes, antecipou e influenciou o desenvolvimento da ciência e da tecnologia. Pode-se afirmar que existe um constante diálogo entre elas. No caso da IA, a ficção científica contribuiu com antecipações bastante relevantes. O termo “robótica” foi criado por Isaac Asimov, baseado na palavra 'robô'. Já em 1921, o escritor checo Karel Capek[36] escreveu uma peça de teatro intitulada 'R.U.R.', onde a palavra 'robot', derivada do termo eslavo 'robota' (que significa tanto 'trabalho' quanto 'escravo'), aparece pela primeira vez.

Antes mesmo de existir qualquer tecnologia embrionária de Inteligência Artificial, a imaginação fértil da ficção científica já projetava possíveis conflitos envolvendo esta tecnologia em uma sociedade futura. No conto 'Círculo Vicioso'[37], escrito em 1942, Asimov apresentou, pela primeira vez na história, uma série de leis para regulamentar o comportamento de robôs na convivência com os seres humanos.[38] Estas leis ficaram conhecidas mais tarde como as 'Três Leis da Robótica'. O que se originou na ficção científica acabou ganhando vida própria para além dela. Asimov chegou a ponderar que suas leis poderiam ter aplicabilidade real no campo da robótica. Em uma entrevista posterior, ele expressou sua confiança na funcionalidade dessas leis no âmbito real.[39]A visão de Asimov, contudo, não se limitava apenas aos robôs, mas também se estendia à Inteligência Artificial.

[36] Disponível em <https://www.wired.com/2011/01/0125robot-cometh-capek-rur-debut/> Acesso em out. de 2019.

[37] Foi escrito em outubro de 1941 e publicado pela revista *Astounding Science Fiction*, edição de março de 1942.

[38] ASIMOV, Isaac. (1983) *Machine that think: The Best Science Fiction Stories About Robot and Computer. História de Robô*, Porto Alegre.L&PM. 2010.

[39] Será que Asimov limitou a imaginação coletiva com o anúncio das suas leis para controle de robôs e IA que poderia ter tomado algum outro rumo? Alguém poderia ter anunciado talvez leis com outro teor e quantidade, mas provavelmente o formato de leis teria mantido. Porque não temos conhecimento sobre alternativas além de leis. O que será analisado na parte 2 deste trabalho.

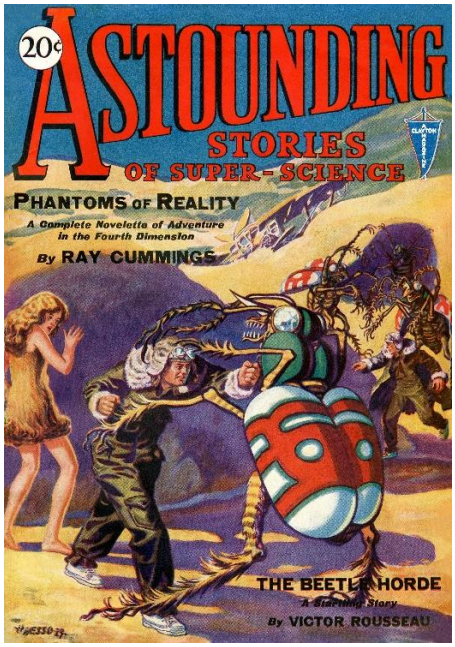


Fig. 1 - Capa da revista *Astounding Science Fiction*[40]

As Três Leis da Robótica se destacam pela importância de seu pioneirismo, pois anteciparam questões de segurança no desenvolvimento de robôs e da Inteligência Artificial. Em um mundo onde a IA é cada vez mais capaz de influenciar o nosso cotidiano, essa é uma questão crítica. Mesmo em PIA subsequentes, a preocupação central da primeira lei, que coloca a segurança humana acima de tudo, permanece como o ponto principal. A habilidade de garantir que a IA nunca cause danos intencionais aos humanos é fundamental para a confiança do público e a adoção da tecnologia.

A segunda lei estipula que os robôs e a IA devem obedecer aos humanos, o que, em essência, reflete a ideia de que a IA deve ser uma ferramenta útil para os humanos, e não um governante. Já a terceira lei, menos discutida, aborda a autopreservação dos robôs. Essa lei poderia ser interpretada como uma previsão de que os robôs poderiam alcançar algum nível de 'autoconsciência' ou 'desejo' de continuar existindo. Porém, o fato de ser considerada a lei menos importante, superada pelas duas primeiras, reitera que as necessidades e a segurança dos humanos devem sempre ser priorizadas.

[40] Disponível em <https://en.wikipedia.org/wiki/Analog_Science Fiction_and_Fact> Acesso em maio de 2023.

As leis de Asimov transcenderam as fronteiras da ficção científica, assumindo uma posição singular na robótica e IA. Em várias discussões sobre a possibilidade e a necessidade de regulamentar a IA, as leis de Asimov são frequentemente citadas como referência fundamental. Na sociedade futurista retratada em suas obras, tanto robôs quanto seres humanos estão cientes da norma que proíbe possíveis danos aos humanos causados por robôs.

Os pontos positivos das Três Leis da Robótica de Asimov são seu pioneirismo, simplicidade e abrangência. Contudo, a crítica mais contundente que recebem é a de sua irrelevância para o cenário atual, dada a dificuldade de sua aplicação prática. Acreditamos que princípios regulatórios subsequentes também enfrentam esse mesmo problema.

Osamu Tezuka, famoso desenhista de mangá japonês, formulou suas próprias leis para robôs em 1988. Essas leis foram introduzidas pela primeira vez no mangá de Tezuka chamado 'Astro Boy' (ou 'Tetsuwan Atom' no Japão), que foi publicado inicialmente em 1952. As leis delineiam um conjunto de princípios éticos para a interação entre robôs e humanos. Embora não sejam reconhecidas fora do contexto da obra de Tezuka, elas exerceram um impacto duradouro na cultura pop e na maneira como as pessoas imaginam as interações entre humanos e robôs. Algumas das leis de Tezuka são inovadoras e merecem atenção, como a proibição de robôs ganharem dinheiro, a proibição de alterarem sua aparência sem permissão, e a proibição de montarem outros robôs descartados pelos seres humanos.

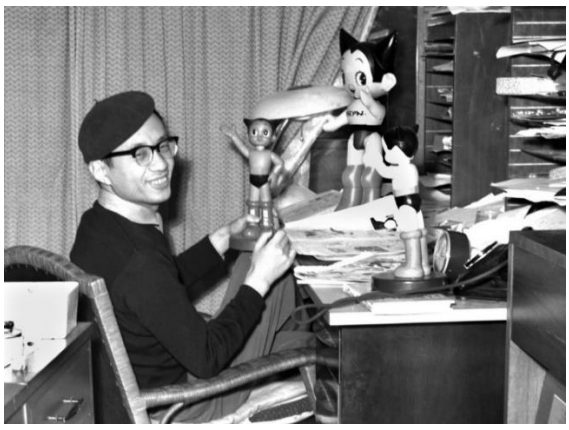


Fig. 2 - Osamu Tezuka segurando boneco Astroboy[41]

[41] Disponível em <https://www2.bfi.org.uk/news-opinion/sight-sound-magazine/features/experimental-short-films-osamu-tezuka> Acesso em mais de 2023.

As duas primeiras leis referentes a robôs originaram-se em obras de ficção científica. Segundo Searle, muitos dos experimentos de pensamento mais importantes na filosofia e na ciência são, de fato, ficções científicas[42]. A partir do ano 2000, a Roboética tornou-se uma realidade e inúmeras leis para regular robôs foram elaboradas e anunciadas por meio de iniciativas de academias, governos, empresas e centros de pesquisa.[43]

No entanto, este movimento não foi uniforme ao longo dos anos, com picos de publicações em determinadas épocas, geralmente influenciados por algum avanço significativo na área. Há de se notar um longo intervalo de 62 anos desde a publicação das leis de Asimov em 1942 até a *World Robot Declaration* durante *International Robot Fair Fukuoka*, em 2004. A única exceção a este hiato foi a contribuição do desenhista japonês Osamu Tezuka, que formulou suas próprias leis para robôs em 1988.

O primeiro Simpósio Internacional de Roboética[44] foi organizado em janeiro de 2004 pela *Scuola di Robotica* em conjunto com o Instituto Teológico da Pontificia Accademia della Santa Croce de Roma, reunindo filósofos, juristas, sociólogos, antropólogos e especialistas em ética. No mesmo ano, durante a Feira Internacional de Robôs em Fukuoka, Japão, que ocorreu em fevereiro de 2004, os participantes assinaram a Declaração Mundial de Robôs.

A *IEEE-Robotics & Automation Society* estabeleceu, também em 2004, um Comitê Técnico (doravante CT) em Roboética. Este comitê foi criado com o objetivo de proporcionar à instituição um meio estruturado para abordar as implicações éticas da pesquisa em robótica. Ademais, buscou-se promover um diálogo entre pesquisadores, filósofos e especialistas em ética. O CT também visou auxiliar na criação de ferramentas compartilhadas para lidar com questões éticas no âmbito da robótica. No ano seguinte, este comitê organizou um workshop como parte da *International Conference on Robotics and Automation* (ICRA 2005).

No ano de 2005, a Rede Europeia de Pesquisa em Robótica (EURON), que faz parte do 6º Programa-Quadro da CE 2003-2007, financiou o Atelier de Pesquisa em Roboética. Este atelier foi proposto pela *Scuola di Robotica*, em parceria com o *CNRS-Laboratoire d'Analyse et d'Architecture des Systèmes* (LAAS) e a *Scuola Superiore Sant´Anna*. O Atelier de Roboética

[42] Searle, John. (1981) *A redescoberta da mente*, São Paulo, Martins Fontes, 2011.
[43] Apesar do esforço em reunir maior quantidade de leis, provavelmente esta lista não está completa.
[44] Disponível em <http://www.roboethics.org/icra2005/veruggio.pdf> Acesso em maio de 2023.

ocorreu em conjunto com a IEEE *International Conference on Robotics and Automation Workshop on Robo-Ethics* em Barcelona, no dia 18 de abril de 2005. Este atelier contou com a participação de cientistas especialistas em robótica, e também de filósofos, juristas, sociólogos e outros estudiosos envolvidos em problemas relacionados à ética na robótica. O principal objetivo do Roboethics Atelier foi produzir um Roteiro de Roboética *(Roboethics Roadmap),* o qual serviria como ferramenta comum à comunidade interessada para o desenvolvimento de uma linguagem comum entre acadêmicos e agentes interessados em Roboética.

Além disso, o Roteiro de Roboética procurou promover o aprendizado sobre os diversos campos, estabelecendo conexões e gerando novas ideias. Buscou-se também um levantamento geral sobre os principais paradigmas éticos nas diferentes culturas, religiões e credos, bem como na definição de diretrizes éticas "adequadas" às diferentes culturas, religiões e credos.

Na ocasião do O Primeiro Simpósio Internacional de Roboética, a antropóloga da Universidade de Genebra, Daniela Cerqui identificou três principais posições éticas emergidas de dois dias de debate:

> 1) Aqueles que não se interessam pela ética. Eles consideram que suas ações são estritamente técnicas, e não acham que têm uma responsabilidade social ou moral em seu trabalho.
>
> 2) Aqueles que estão interessados em questões éticas de curto prazo. De acordo com esse perfil, as questões são expressas em termos de "bom" ou "mau" e referem-se a alguns valores culturais. Por exemplo, eles acham que os robôs devem aderir às convenções sociais. Isso incluirá "respeitar" e ajudar os humanos em diversas áreas, como na implementação de leis ou na ajuda a idosos. (Tais considerações são importantes, mas devemos lembrar que os valores usados para definir o "mau" e o "bom" são relativos. São os valores contemporâneos dos países industrializados).
>
> 3) Aqueles que pensam em termos de questões éticas de longo prazo, sobre, por exemplo, a "exclusão digital" entre Sul e Norte, ou jovens e idosos. Eles estão cientes da distância entre os países industrializados e os pobres, e se perguntam se os primeiros não deveriam mudar sua forma de desenvolver a robótica para serem mais úteis para o Sul. Eles não formulam explicitamente a pergunta para quê, mas podemos considerar que está implícito.

A partir dos pontos levantados por Daniela Cerqui, podemos vislumbrar a dissonância e o intenso debate entre os participantes vindos de diversas áreas de atuação. Mesmo após quase duas décadas, questões como exclusão e desigualdade entre os países persistem, seja na robótica ou na Inteligência Artificial. Se no momento da criação das primeiras leis, os robôs apenas faziam parte da ficção científica, pela apresentação mercadológica que acompanha a declaração coreana, observamos que no início dos anos 2000, o uso de robôs na indústria já era uma realidade consolidada. Assim, essa declaração parece inserir-se em uma estratégia do país para competir no mercado global.

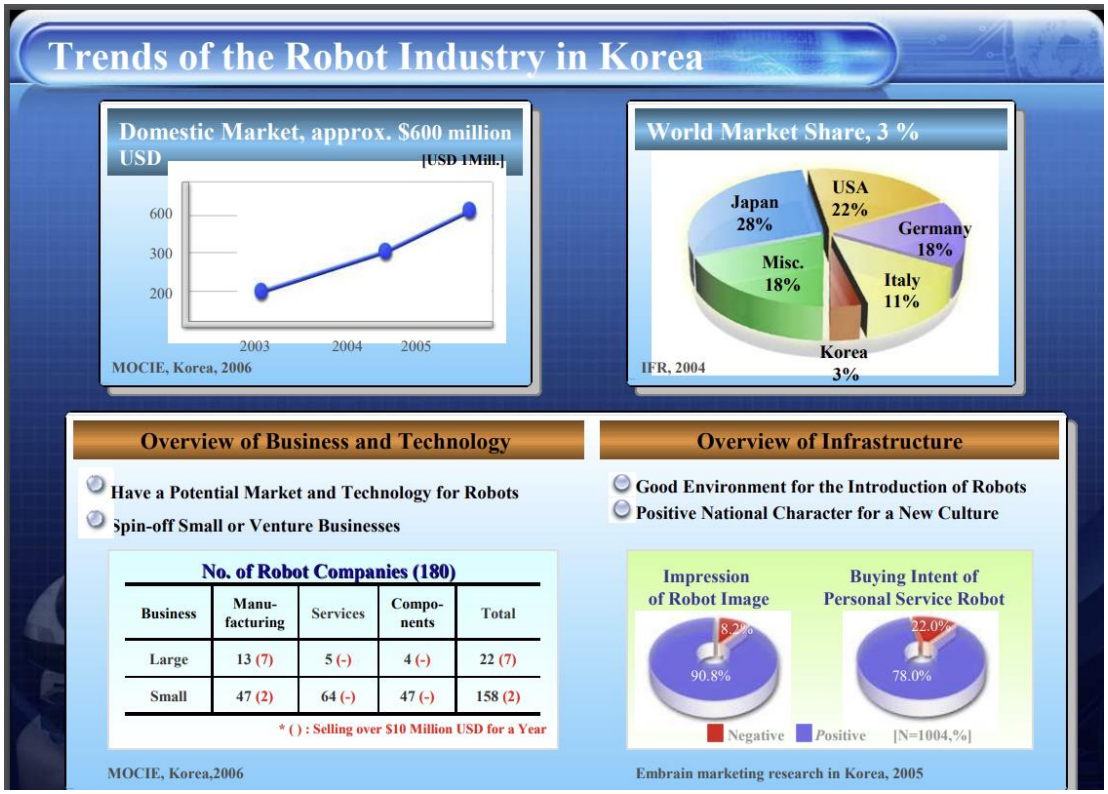


Fig. 3 – Infográficos da declaração de roboética coreana[45]

Um dado que chama atenção é o aumento expressivo do uso de robôs na indústria coreana. A Coreia, que em 2004 representava 3% do *market share*, alcançou um marco histórico em 2021, registrando 1.000 robôs industriais para cada 10.000 funcionários. Isso representa mais do que o triplo do número alcançado na China, o que coloca o país como líder mundial neste segmento. A economia coreana conta com duas grandes indústrias

[45] Ibid. pag.1

consumidoras de robôs industriais: a eletrônica e a automotiva, ambas com reconhecimento global.

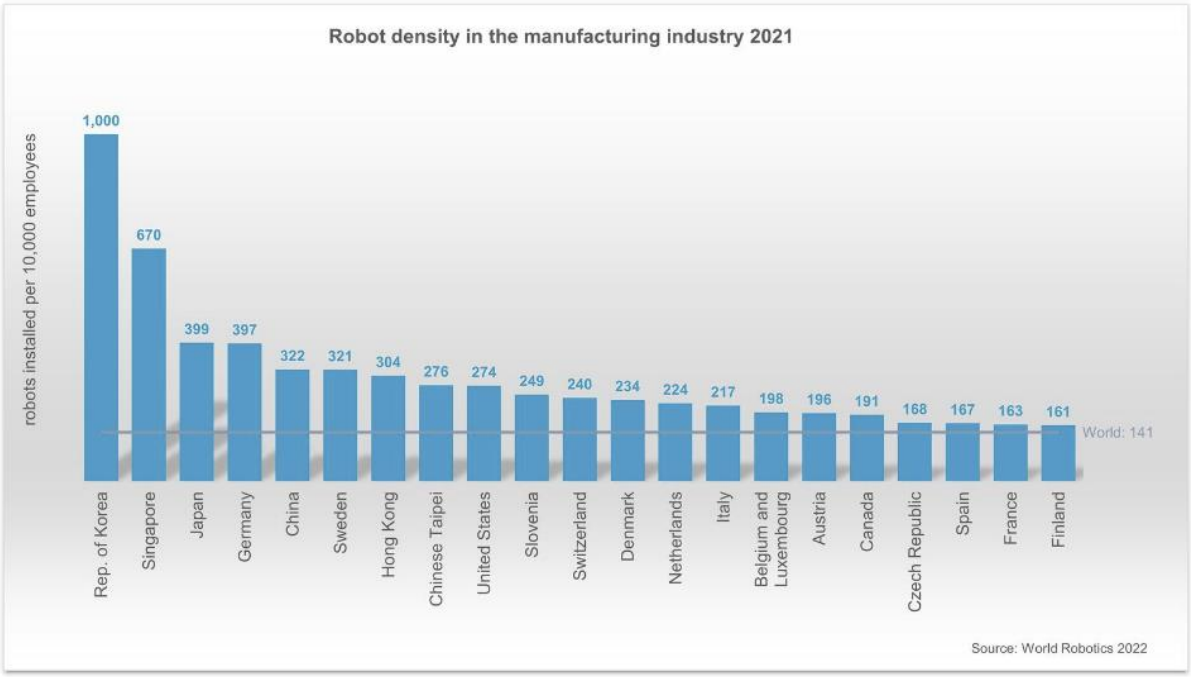


Fig. 4 - Utilização de robôs na indústria por países[46]

Se desde as leis de Asimov de 1942 até 2016 surgiram menos de dez leis, apenas no biênio de 2017 e 2018 temos quase trinta leis. Além da mudança do protagonismo da IA no lugar de robô que acelerou a discussão sobre o controle, alguns acontecimentos próximos de 2016 envolvendo IA poderia ter provocado o surgimento de princípios e leis para IA nos anos seguintes.

Em março de 2016, o AlphaGo da Google derrotou o campeão mundial de Go, o sul-coreano Sedol Lee. Anteriormente, em 1997[47], o computador Deep Blue da IBM já havia vencido o enxadrista russo Garry Kasparov. Enquanto o xadrez apresenta um tabuleiro com 64 casas, o jogo chinês Go possui 361 pontos onde as pedras podem ser posicionadas. Dessa forma, no início da partida, o xadrez apresenta 400 possibilidades de jogadas, em contraste com as 130 mil possíveis no Go. Por esta razão, especialistas acreditavam que, apesar da

[46] Disponível em <https://ifr.org/ifr-press-releases/news/china-overtakes-usa-in-robot-density> Acesso em junho de 2023.
[47] Disponível em <http://time.com/3705316/deep-blue-kasparov/ > Acesso em out. de 2019.

derrota de Kasparov, o mesmo não aconteceria no Go, ou ao menos levaria cerca de 50 anos para que um computador pudesse derrotar um ser humano numa partida do jogo. No entanto, em menos de 20 anos, uma máquina conseguiu também derrotar um humano em um jogo anteriormente considerado como improvável para tal feito. O mais surpreendente desta conquista é que o AlphaGo melhorou suas habilidades por conta própria, sendo instruído apenas a seguir exemplos sem uma programação específica[48] para tal.

Fig. 5 - Alphago vs Lee Sedol[49]

Em junho, ocorreu um acidente fatal com um carro em modo autônomo da Tesla[50] . Em setembro do mesmo ano, a Uber realizou os primeiros testes com carros autônomos. Pokémon Go, um jogo para celular que utiliza realidade aumentada, foi lançado em julho e teve 500 milhões de downloads em apenas três meses, tornando-se o jogo para aparelho celular de maior sucesso de todos os tempos.[51, 52]

[48] Disponível em <www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/> Acesso em out. de 2019.

[49] Disponível em <https://www.joongang.co.kr/article/25074871> Acesso em junho de 2023. “Quando o AlphaGo derrotou Lee Sedol, foi utilizada uma energia de 170 kW. Comparativamente, o cérebro de um adulto, como o de Lee Se-dol, consome apenas 0,02 kW. Isso significa que o AlphaGo consumiu a energia equivalente à utilizada por 8.500 Lee Se-dol para jogar Go. Em termos energéticos, foi uma batalha de 8.500 para 1. Até o momento, nenhuma inteligência artificial conseguiu operar com uma quantidade tão reduzida de energia como o cérebro humano. No caso da inteligência artificial para direção autônoma anunciada pela Tesla em agosto do ano passado, a potência utilizada por um servidor era impressionante, atingindo 1800 kW, mais de 10 vezes a do AlphaGo.” Tradução nossa.

[50] Disponível em <http://g1.globo.com/carros/noticia/2016/06/acidente-com-carro-da-tesla-em-modo-semiautonomo-deixa-1-morto.html> Acesso em out. de 2019

[51] Disponível em < http://g1.globo.com/carros/noticia/2016/06/acidente-com-carro-da-tesla-em-modo-semiautonomo-deixa-1-morto.html > Acesso em out. de 2019.

[52] Disponível em <https://www.businessinsider.com/pokemon-go-500-million-downloads-2016-9> Acesso em out. de 2019.

Em um plebiscito realizado em 23 de junho, a maioria dos britânicos decidiu pela saída do Reino Unido da União Europeia, o que causou espanto mundial. Em 9 de novembro, Trump ganhou as eleições presidenciais nos Estados Unidos. Tanto a campanha do plebiscito do Brexit quanto a eleição de Trump foram marcadas pela disseminação de notícias consideradas falsas nas mídias sociais, com o uso da tecnologia da empresa Cambridge Analytica, que posteriormente se tornou alvo de investigação criminal.[53] A palavra do ano de 2016, eleita pelo dicionário Oxford, foi "pós-verdade".[54]

Em 2016, tivemos algumas conquistas inéditas e surpreendentes na área da IA, como a vitória do AlphaGo, mas também ocorreram exemplos de acidentes com carros em modo autônomo e manipulação da opinião pública em duas grandes campanhas, por meio do uso de tecnologias nas redes sociais. Esses eventos, em parte, podem explicar as diversas iniciativas de controle da IA nos anos seguintes. No final de 2016, pesquisadores em IA de cinco das maiores empresas de tecnologia do mundo (Amazon, Apple, Facebook, Google - DeepMind, IBM e Microsoft), diretamente envolvidas no desenvolvimento da IA, reuniram-se e estabeleceram uma parceria para o desenvolvimento da IA anunciando os "Tenets of Partnership on AI".

Ao analisarmos as leis anunciadas, podemos identificar uma genealogia em que podemos observar a influência das leis anteriores nas posteriores, tanto em termos de ideias, formato, quanto de termos utilizados. Desde as três leis de Asimov, diversas iniciativas de controle da IA e robótica foram anunciadas, como leis, princípios (*tenets*), regras, declarações (*declaration, charter*) e diretrizes (*guidelines*). Os títulos "leis" e "princípios" são as duas formas mais utilizadas.

As leis que se inspiram diretamente em Asimov também apresentam uma concatenação hierárquica entre elas. Os princípios de Russell têm uma clara inspiração nas leis de Asimov, tanto em formato quanto em sequência de prioridade. A formulação "O único objetivo do robô é maximizar a realização dos valores humanos" e "O comportamento

[53] Disponível em <www.theguardian.com/politics/2017/mar/04/nigel-oakes-cambridge-analytica-what-role-brexit-trump> Acesso em out. de 2019.
[54] Disponível em < https://g1.globo.com/educacao/noticia/pos-verdade-e-eleita-a-palavra-do-ano-pelo-dicionario-oxford.ghtml > Acesso em out. de 2019.

humano fornece informações sobre os valores humanos" demonstra inspiração na *Robot Ethics Charter Korea* (2007), que visa "identificar padrões éticos centrados no homem para a coexistência entre humanos e robôs".

Em 2017, cerca de 2.300 pessoas, incluindo 880 pesquisadores em IA e robótica, juntamente com físicos, economistas, filósofos e juristas, reuniram-se na conferência de Inteligência Artificial em Asilomar, organizada pelo Instituto Futuro da Vida. Foi a maior mobilização com o objetivo de discutir o desenvolvimento da IA até o momento. Foram anunciados 23 princípios que devem orientar as questões de pesquisa, segurança e ética no desenvolvimento da IA. Esses princípios receberam o endosso de Stephen Hawking, Elon Musk, Saul Permuter, Frank Wilczek, entre outros. Os 23 princípios de Asilomar têm uma importância relevante para o estudo das tentativas de controle da IA por meio de regras, pois representam uma síntese das iniciativas anteriores.

> Esse senso entre os participantes reflete um engajamento social mais amplo com a IA que aqueceu dramaticamente nos últimos anos. Devido a essa crescente conscientização sobre a IA, surgiram dezenas de relatórios importantes da academia, do governo, da indústria e do setor sem fins lucrativos. Reunimos todos os relatórios que pudemos e compilamos uma lista de opiniões sobre o que a sociedade deveria fazer para gerenciar melhor a IA nas próximas décadas.[55]

Os princípios de Asilomar utilizam diversos termos para se referir à IA: sistemas de IA, tecnologias de IA, IA avançada, IA altamente avançada, superinteligência. Isso demonstra o grau de dificuldade que enfrentamos ao lidar com a IA, especialmente na questão de definir sua natureza, ainda mais como um agente. Essa diversidade de termos pode aumentar a confusão, tornando necessário estabelecer uma definição clara da natureza da IA para criar critérios válidos, especialmente quando se trata de considerá-la como um agente, entre outros aspectos.

[55] Disponível em <https://futureoflife.org/ai-principles/> Acesso em out. de 2019. Tradução nossa.

Fig. 6 - Conferência de Asilomar[56]

O desenvolvimento da robótica pode ser dividido em várias fases distintas[57], cada uma caracterizada por avanços tecnológicos e aplicações específicas. A primeira fase da robótica envolveu a criação de robôs capazes de executar tarefas básicas de manipulação física, como levantar, empurrar e transportar objetos. Os primeiros robôs industriais, desenvolvidos nas décadas de 1960 e 1970, eram geralmente utilizados em linhas de produção para realizar tarefas repetitivas e pesadas.

Na segunda fase, houve um avanço na mobilidade dos robôs, permitindo que se movessem de forma autônoma em diferentes ambientes. Além disso, os robôs passaram a ser equipados com sensores que lhes permitiam detectar e responder ao ambiente ao seu redor. Isso permitiu que os robôs interagissem com o ambiente de maneira mais sofisticada e realizassem tarefas mais complexas.

Na terceira fase, a ênfase foi colocada na capacidade dos robôs de colaborar e interagir com os seres humanos e outros robôs. A robótica colaborativa, por exemplo, envolve o desenvolvimento de robôs que podem trabalhar em estreita colaboração com humanos, compartilhando espaços de trabalho e realizando tarefas em equipe. A robótica social também emergiu nessa fase, buscando criar robôs capazes de interagir socialmente com as

[56] Disponível em <https://commons.wikimedia.org/w/index.php?curid=109870710> Acesso em junho de 2023.
[57] Disponível em <https://en.wikipedia.org/wiki/History_of_robots> Acesso em junho de 2023.

pessoas e desempenhar funções em ambientes como cuidados de saúde, educação e entretenimento. Destacamos alguns marcos históricos. [58]

> 1804 - A produção em massa é automatizada pela primeira vez
>
> No século 18, no Reino Unido e na França, a tecelagem era uma indústria importante, mas de mão-de-obra intensiva; os tecelões exigiam assistentes para levantar e abaixar os fios para produzir padrões. Os inventores tentaram automatizar o processo e, em 1804, um inventor francês chamado Joseph-Marie Jacquard revelou o que se tornaria o amplamente adotado "Tear Jacquard". Funcionava traduzindo padrões de cartões perfurados em comandos que determinavam o levantar e abaixar os fios, aumentando a velocidade na qual padrões complexos podiam ser tecidos em mais de vinte vezes, de uma polegada para dois pés por dia. O tear tornou-se o primeiro sistema amplamente utilizado que podia seguir um programa; nesse sentido, foi o primeiro exemplo de programação de computadores.
>
> 1959 - Primeiro braço robótico é instalado no chão de fábrica
>
> Conhecido como "Unimate", o primeiro braço robótico industrial foi trabalhar em uma fábrica da General Motors, levantando e empilhando peças metálicas cortadas a quente. Criado por George Devol e seu parceiro Joseph Engelberger, ele podia se mover para cima e para baixo nos eixos X e Y, possuía uma pinça giratória semelhante a uma pinça e podia seguir um programa de até 200 movimentos armazenados em sua memória. Desdobrável para inúmeras tarefas, principalmente algumas que eram muito exigentes ou perigosas para os humanos - como levantar cargas de 75 libras sem cansar e trabalhar em meio a fumaça tóxica - o Unimate iniciou a transformação da indústria automobilística em uma arena de automação generalizada.
>
> 1972- Primeiro robô a usar inteligência artificial
>
> Era conhecido como Shakey por causa da maneira gaguejante com que se movia, mas o que mais distinguia esse robô, criado por um grupo de engenheiros do Stanford Research Institute, era que ele incluía uma inteligência artificial pioneira. Se você desse um objetivo a Shakey - como navegar por uma sala ou empurrar uma caixa pelo chão - ele poderia alcançá-lo observando o mundo ao seu redor, criando um plano e executando. Com sensores que incluíam uma câmera de TV, um telêmetro e bigodes de metal sensíveis ao toque, Shakey reunia dados que permitiam construir um modelo de seu ambiente e então usar um programa de "planejamento" para gerar seus próximos movimentos. Essa ideia de uma camada de "planejamento" separada foi uma inovação tão crucial que ainda hoje é fundamental para muitos sistemas robóticos.
>
> 2005 - Carros autônomos passam em seu primeiro grande teste
>
> A era moderna dos carros autônomos foi lançada em 8 de outubro de 2005, quando um Volkswagen Touareg chamado "Stanley" venceu o segundo DARPA Grand Challenge - para completar um percurso difícil e muitas vezes angustiante de 131,2 milhas no deserto de Mojave em 10 horas. A corrida havia sido estabelecida no ano anterior pela Agência de Projetos de Pesquisa Avançada de Defesa (DARPA) do Departamento de Defesa para estimular a competição e a inovação na tecnologia

[58] Disponível em <https://www.aventine.org/robotics/history-of-robotics> Acesso em junho de 2023. Tradução nossa.

> de veículos militares autônomos, mas nenhum dos carros da competição anterior conseguiu percorrer mais de 13 quilômetros. O que impulsionou a vitória de Stanley foi uma constelação de melhorias, incluindo IA treinada nos hábitos de direção de humanos do mundo real e cinco sensores a laser “Lidar”, uma tecnologia que permitiu ao carro identificar objetos dentro de um alcance de 25 metros à frente do veículo. . Lidar, que significa “detecção e alcance de luz”, desde então se tornou um componente-chave de sistemas de visão robótica em carros e até mesmo em alguns robôs de armazém no estilo Kiva; na verdade, a principal empresa de Lidar, a Velodyne, foi derivada de um dos concorrentes de Stanley na corrida.

A IA também teve épocas classificadas como inverno e primavera ao longo do seu desenvolvimento. Esta classificação de épocas, apesar de ser uma forma metafórica de descrever os altos e baixos no desenvolvimento e no avanço desta tecnologia, oferece facilidade para a compreensão dos padrões históricos.

> Èpocas importantes no desenvolvimento da IA
>
> 1956–1974: OS ANOS DOURADOS
>
> Durante os anos dourados da IA, os programas – incluindo computadores resolvendo problemas de álgebra e aprendendo a falar inglês – parecem “surpreendentes” para a maioria das pessoas.
>
> 1974–1980: INVERNO AI DO SÉCULO XX
>
> O primeiro inverno de IA ocorre porque as capacidades dos programas de IA permanecem limitadas, principalmente devido à falta de poder de computação na época. Eles ainda podem lidar apenas com versões triviais dos problemas que deveriam resolver.
>
> 1987–1993: UM INTERESSE RENOVADO
>
> O fascínio e as expectativas da comunidade empresarial em relação à IA, especialmente sistemas especialistas, aumentam. Mas eles são rapidamente confrontados com a realidade de suas limitações.[59]

Acreditamos que a classificação histórica dos PIA também poderia trazer pontos valiosos como aconteceu com a história da robótica e IA. Assim, oferecemos neste trabalho uma divisão histórica para PIA. Identificamos as leis de Asimov como a pioneira que inaugurou o movimento que se estendeu até os dias atuais, quase oitenta anos depois também com seus invernos e primaveras ou até mesmo mudanças climáticas.

[59] Disponível em <https://www.historyofdatascience.com/ai-winter-the-highs-and-lows-of-artificial-intelligence/> Acesso em junho de 2023. Tradução nossa.

Tabela 1 - Classificação de PIA

Fases	Primeira fase - tradicional	Segunda fase - atual
Época	1942 até 2015	2016 em diante
Modo	Negativo	Positivo
Termo mais utilizado	Leis	Princípios
Protagonismo	Robôs	IA
Principal Iniciativa	As três leis de Asimov	23 Princípios de Asilomar
Objetivo principal	Proibição de danos para seres humanos	Garantia de benefícios para seres humanos
Objetivos secundários	Proteção para robô	Privacidade, Explicabilidade, Justiça, Transparência
Países protagonistas	EUA, UE, Japão, Coreia	EUA, UE, Reino Unido, Canadá, China

Dividimos a história dos PIA em duas fases: A primeira fase tradicional que começa com “As três leis de Asimov” até as “Oito leis de Shinpo Fumio” de 2015. Nesta fase, as leis eram dirigidas para controlar os robôs. A segunda fase começa em 2016 e continua até hoje. Nesta fase o termo mais utilizado é princípio e dirigido para a IA. A maior iniciativa nesta fase foi o encontro de Asilomar que culminou com a publicação de 23 princípios. Se na primeira fase a preocupação maior nas leis era evitar danos para seres humanos, a segunda fase é marcada pela esperança de trazer benefícios para humanidade. Projetamos uma terceira fase para PIA que será apresentada no último capítulo.

Acreditamos que a classificação da história dos PIA em diferentes fases oferece uma visão abrangente do desenvolvimento, que além de possibilitar um estudo histórico poderá

auxiliar na avaliação do progresso, no entendimento dos desafios e na inspiração para futuras inovações.

A pesquisa sobre os princípios do presente trabalho teve início em 2019, logo após o período de 2017/2018, que foi o biênio de maior intensidade na publicação de princípios de IA até hoje, incluindo uma das principais iniciativas, que foi a de Asilomar. Durante a verificação dos links coletados durante a finalização deste trabalho, no ano de 2023, constatamos que muitos dos princípios de IA que foram anunciados no passado já não estavam mais disponíveis nos sites onde foram originalmente publicados. Esse problema foi mais frequente em iniciativas privadas do que nas originadas em academias, institutos, associações, centros de pesquisas e governos, que conseguiram manter melhor as publicações passadas. Entendemos perfeitamente que a própria ideia de princípios pode estar em constante aperfeiçoamento, mas chamamos a atenção para a fragilidade dos compromissos assumidos principalmente por empresas, e para a facilidade com que abandonam tais compromissos anteriormente divulgados como duradouros em poucos anos, de acordo com seu próprio julgamento, unilateralmente.

PARTE I CAPÍTULO 3 - QUATRO ESTUDOS DE PIA

Se nos últimos anos houve quase uma proliferação de Princípios de Inteligência Artificial (PIA) que autores como Floridi classificaram como um mercado de princípios, estudos críticos sobre os PIA, mesmo que não sejam numerosos, têm surgido recentemente. Aqui, selecionamos quatro estudos: "*Policy Paper on the Asilomar Principles on Artificial Intelligence*" da Federação de Cientistas da Alemanha (VDW), "*A Unified Framework of Five Principles for AI in Society*" de Luciano Floridi e Josh Cowls. "*Linking Artificial Intelligence Principles*" de Yi Zeng, Enmeng Lu, Cunqing Huangfu da Academia Chinesa de Ciências. "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*" de Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy e Madhulika Srikumar do Berkman Klein Center for Internet & Society.

Faremos uma exposição a partir do menor número de PIA abordados até o maior número: 1 (um) PIA no caso Asilomar da Federação de Cientistas da Alemanha (VDW), 6 (seis) PIA de Luciano Floridi, 27 (vinte e sete) PIA pelos pesquisadores da Academia Chinesa de Ciências, 32 (trinta e dois) PIA pelos pesquisadores do Berkman Klein Center for Internet & Society. Apresentaremos esses quatro estudos levantando informações básicas como autores, objetivos, PIA estudados, metodologia e resultados, citando introdução e conclusão de cada artigo.

2.1. Documento de Política sobre os Princípios de Asilomar sobre Inteligência Artificial - Federação de Cientistas da Alemanha (*Vereinigung Deutscher Wissenschaftler*)

O primeiro estudo, "*Policy Paper on the Asilomar Principles on Artificial Intelligence*", da Federação de Cientistas da Alemanha (VDW), foi publicado em junho de 2018, praticamente um ano após a publicação dos 23 Princípios de Asilomar em agosto de 2017. Conforme se pode verificar pelo título, o artigo foca em analisar apenas os Princípios de Asilomar. Esses princípios foram apresentados em uma conferência que ocorreu no local de mesmo nome no ano anterior, com especialistas em IA de todo o mundo, que discutiram as condições para o desenvolvimento seguro, responsável e benéfico da IA para a sociedade.

> A inteligência artificial abriu a caixa de Pandora. Tem o potencial de minar a lógica do controle - incluindo os perigos mencionados acima. Na era digital, as ferramentas do pensamento humano podem desdobrar uma posição autônoma que se dirige poderosamente contra o ser humano impotente. É hora de uma defesa ativa e refletida contra tal perigo. Este é o objetivo do texto a seguir. A Federação de Cientistas Alemães apoia este aviso. Sua fundação remonta ao alerta dos "Dezoito de Göttingen" 8 em 1957 contra os perigos do armamento nuclear. A responsabilidade da ciência hoje significa nada menos do que então tornar os perigos não reconhecidos visíveis ao público e servir à segurança com perícia científica. Pensando nisso, o VDW “Equipe de Pesquisa Technology Assessment of Digitalization” iniciou seus trabalhos e apresenta seus primeiros resultados. (VDW. 2018. p. 3)[60]

Pela introdução, podemos obter um resumo sobre o conteúdo do artigo e a motivação por trás da iniciativa. O artigo da Federação de Cientistas da Alemanha (VDW) apresenta uma análise crítica dos 23 princípios de Asilomar e como eles poderiam ou não ser aplicados na prática. Ele também discute preocupações éticas em torno do desenvolvimento da IA incluindo privacidade, segurança, responsabilidade e justiça. Os pesquisadores da VDW apontam como causa principal do problema nos princípios de Asilomar uma visão demasiadamente otimista sobre IA, um entusiasmo técnico dominado pelos ganhos de produtividade econômica, sem análises socioeconômicas amplas sobre seus impactos, e uma visão da consolidação da IA como um destino inevitável. Assim, mesmo para os possíveis perigos futuros, restaria apenas uma alternativa: tentar reduzir consequências indesejáveis, mas os autores dos princípios de Asilomar jamais cogitam em proibir mesmo para aplicações de uma IA forte com potencial destrutivo e corrida de armas letais autônomas.

> Os princípios de Asilomar são um excelente ponto de partida para discussões sobre como explorar o potencial da Inteligência Artificial nos próximos anos. No entanto, os princípios não são uma estrutura normativa apropriada para a definição necessária de limites absolutos para a pesquisa, desenvolvimento e aplicação da IA. nem para a aplicação de tais limites por motivos de segurança.[61]

Os pesquisadores da VDW questionam se seria realista que a utilização da IA seja controlada apenas por acordos voluntários entre pesquisadores, sem uma participação formal das estruturas e processos institucionais existentes do espaço político constituído democraticamente. Essa indagação é o cerne deste projeto.

[60] FEDERATION OF GERMAN SCIENTISTS E.V. (VDW) *Policy Paper on the Asilomar Principles on Artificial Intelligence*Marienstraße 19/20, 10117 Berlin. 2018.
[61] Ibid. p. 24. Tradução nossa.

> **Os princípios usam numerosos termos legais indefinidos** que teriam de ser definidos para serem desenvolvidos como um instrumento gerenciável.[62] […] Nossa análise dos princípios de Asilomar mostrou que eles não são consistentes com a lógica da prevenção bem-sucedida de riscos. Se apenas seguíssemos as recomendações dos princípios de Asilomar, a nosso ver, seria aceito um risco essencial.[63]

O artigo da Federação de Cientistas Alemães também destaca a problemática de quem poderia determinar o que é "benéfico" diante de uma tecnologia com aspectos abrangentes, capaz de impactar praticamente toda a sociedade. Questiona-se se a busca pelo benefício da IA prevaleceria sobre as possibilidades de danos e ameaças, e se isso seria um consenso em escala global. Há a ponderação se esse consenso, uma vez estabelecido, continuará válido no futuro quando os resultados do mesmo não puderem mais ser revertidos.

> O lema da conferência "AI benéfica" mencionado definiu a abordagem e também se reflete no breve preâmbulo dos princípios: Com essa visão puramente positivista e utilitária do uso da IA, permanece a grande e talvez decisiva questão aberta (entre outras): como lidar com desenvolvimentos que não são benéficos para todos ou que nem todos e, acima de tudo, como lidar com as ameaças colocadas por esses desenvolvimentos.[64]

Para os pesquisadores da VDW, a pesquisa e desenvolvimento (P&D) da IA também deve seguir princípios normativos éticos e legais, tal como ocorre em outras áreas de pesquisa, fundamentando-se na Declaração Universal dos Direitos Humanos e em sua codificação nas legislações nacionais. Eles expressam urgência, pois o desenvolvimento de sistemas jurídicos internacionalmente válidos e equipados com instrumentos de execução pode levar décadas, portanto, o trabalho deve iniciar imediatamente. O artigo ainda reconhece o valor pioneiro de Asimov na questão fundamental.

> A orientação mais importante deve ser que a IA não pode prejudicar uma pessoa sob nenhuma circunstância concebível. Isso está de acordo com as leis robóticas de Asimov e é um pré-requisito absoluto para a "utilidade" da IA.[65]

Os pesquisadores da VDW analisam os diferentes princípios e apresentam uma discussão crítica sobre a sua aplicabilidade e eficácia. Destacam que muitos desses princípios são vagos e imprecisos, e que pode haver uma falta de clareza sobre como esses princípios podem ser aplicados na prática. Além disso, observam que alguns princípios podem ser

[62] Ibid., p. 13. Tradução e destaque nossa.
[63] VDW. op. cit. p. 4.
[64] VDW op. cit. p. 11. Tradução nossa.
[65] Ibid. p. 29. Tradução nossa.

conflituosos entre si, tornando difícil equilibrá-los na prática. Eles questionam se a cooperação entre os pesquisadores de IA, conforme descrita nos princípios, seria realista e se fatores externos, como lucro financeiro e segurança nacional, não teriam um impacto significativo ou não definiriam a agenda de pesquisa.

> Muitas formulações escolhidas parecem assumir que os cientistas encarregados de P&D por IA trabalham juntos de maneira cooperativa e sem dominação, ou que isso seja possível, desde que os pesquisadores estejam dispostos a fazê-lo.[66]

Os pesquisadores da VDW defendem que os princípios de Asilomar são um guia crucial para direcionar a governança e a regulamentação da IA. Entretanto, destacam também a importância de uma abordagem pragmática e adaptável para lidar com as futuras mudanças e desafios nessa área. Eles enfatizam a necessidade de uma colaboração estreita entre especialistas em IA, formuladores de políticas, indústria e sociedade civil para assegurar que a IA seja desenvolvida de maneira responsável e benéfica para a humanidade.

O artigo apresenta uma análise relevante dos princípios articulados na Conferência de Asilomar, evidenciando suas lacunas e potenciais consequências prejudiciais. Sugere-se que esses princípios, apesar de bem-intencionados, são excessivamente vagos e amplos, o que poderia levar a interpretações diversas e dificultar sua aplicação no mundo real.

Os pesquisadores da VDW não se limitam a criticar, propondo alternativas e complementos aos princípios de Asilomar que buscam corrigir as deficiências por eles identificadas. Com uma fundamentação robusta e referências a diversos outros estudos e pesquisas na área de ética em IA, o artigo é, sem dúvida, uma contribuição valiosa para o campo.

Talvez o aspecto mais significativo deste artigo seja a identidade da própria entidade. Esta federação tem suas raízes no Grupo 18 de Göttingen[67], que se manifestou contra as armas nucleares, fato que ressalta a magnitude do perigo representado pelo advento da IA.

---

[66] Ibid. p. 13. Tradução nossa.

[67] 18 de Göttingen (em alemão: *Göttinger Achtzehn*) foi um grupo formado por 18 cientistas nucleares da República Federal da Alemanha que em 12 de abril de 1957 divulgou um manifesto contra o governo (Manifesto de Göttingen) expressando oposição ao abastecimento do Exército da Alemanha de armas nucleares táticas. O manifesto foi dirigido em especial ao chanceler Konrad Adenauer e ao ministro da Defesa, Franz Josef Strauß. O nome 18 de Göttingen refere-se às origens acadêmicas comuns de muitos de seus membros na cidade universitária de Göttingen. É também uma alusão ao "Sete de Göttingen", sete professores da Universidade de

> A Federação de Cientistas Alemães (*Vereinigung Deutscher Wissenschaftler*, VDW) é uma rede de cientistas de várias disciplinas acadêmicas, que refletem criticamente sobre sua responsabilidade pelos efeitos de sua pesquisa científica e desenvolvimento técnico e que usam seus conhecimentos para participar ativamente no debate social sobre temas como paz, clima, biodiversidade e economia[68]. […] A ciência e a tecnologia são fundamentos importantes da nossa civilização e da nossa cultura. No entanto, ainda são poucos os cientistas que se preocupam com os impactos sociais de seu trabalho, principalmente quando seus resultados adentram o campo de força político e econômico. O VDW enfrenta esse déficit com um discurso ativo.

2.2. Uma estrutura unificada de cinco princípios para IA na sociedade - Luciano Floridi e Josh Cowls[69]

Luciano Floridi é um dos filósofos contemporâneos mais conhecidos no campo da filosofia da informação e da ética da tecnologia da informação, incluindo a IA. Floridi é conhecido por sua definição de "filosofia da informação", que visa compreender e explicar a natureza e dinâmica da informação como um conceito fundamental e universal, que é crucial para entender o mundo e a nós mesmos, bem como as questões éticas e sociais que surgem no contexto da tecnologia da informação.

No campo da IA, Floridi foi fundamental na discussão sobre a ética. Ele defende que as questões éticas na IA não são apenas problemas técnicos, mas também questões filosóficas fundamentais. Ele se concentra em questões como a natureza da IA, as implicações éticas da IA e as maneiras pelas quais a IA pode afetar e ser afetada pela sociedade. Além disso, Floridi tem sido um defensor do conceito de "dignidade da informação", argumentando que a informação tem um valor intrínseco que precisa ser protegido e preservado. Ele também propôs a "ética da informação" como uma nova área de pesquisa ética, que examina os direitos e deveres morais em relação à informação. Entre 2008 e 2013, Floridi [70]ocupou a cátedra de pesquisa em filosofia da informação e a Cátedra UNESCO em Informação e Ética da Computação na Universidade de Hertfordshire. Ele foi o fundador e diretor do IEG, um

---

Göttingen que em 1837 protestaram publicamente contra a suspensão da Constituição pelo rei Ernesto Augusto I. Disponível em < https://en.wikipedia.org/wiki/Federation_of_German_Scientists>. Acesso em maio de 2023. Tradução nossa.

[68] Disponível em <https://vdw-ev.de/english/about-us/> Acesso em maio de 2023. Tradução nossa.

[69] FLORIDI, Luciano and COWLS, Josh, *A Unified Framework of Five Principles for AI in Society* (September 20, 2019). Disponível em
SSRN: <https://ssrn.com/abstract=3831321> or <http://dx.doi.org/10.2139/ssrn.3831321> Acesso em maio de 2023.

[70] Disponível em <https://en.wikipedia.org/wiki/Luciano_Floridi> Acesso em maio de 2023. Tradução nossa.

grupo de pesquisa interdepartamental sobre a filosofia da informação na Universidade de Oxford.

O artigo "*A Unified Framework of Five Principles for AI in Society*" de Luciano Floridi e Josh Cowls foi publicado em junho de 2019.

> A Inteligência Artificial (IA) já está tendo um grande impacto na sociedade. Como resultado, muitas organizações lançaram uma ampla gama de iniciativas para estabelecer princípios éticos para a adoção de IA socialmente benéfica. Infelizmente, o grande volume de princípios propostos ameaça sobrecarregar e confundir. Como esse problema de “proliferação de princípios” pode ser resolvido? Neste artigo, relatamos os resultados de uma análise refinada de vários dos conjuntos de princípios éticos mais importantes para IA. Avaliamos se esses princípios convergem para um conjunto de princípios acordados ou divergem, com desacordo significativo sobre o que constitui "IA ética". Nossa análise encontra um alto grau de sobreposição entre os conjuntos de princípios que analisamos. Em seguida, identificamos uma estrutura abrangente que consiste em cinco princípios básicos para a IA ética. Quatro deles são princípios centrais comumente usados na bioética: beneficência, não maleficência, autonomia e justiça. Com base em nossa análise comparativa, argumentamos que um novo princípio é necessário adicionalmente: a explicabilidade, entendida como incorporando tanto o sentido epistemológico da inteligibilidade (como resposta à pergunta 'como isso funciona?') quanto o sentido ético de responsabilidade (como resposta à pergunta: 'quem é responsável pela forma como funciona?'). Na discussão que se segue, observamos as limitações e avaliamos as implicações dessa estrutura ética para futuros esforços para criar leis, regras, padrões técnicos e melhores práticas para IA ética em uma ampla variedade de contextos. (Ibid. p. 2. Tradução nossa)

Eles selecionaram seis PIA para o artigo:

1. Os Princípios Asilomar AI, iniciativa do Future of Life Institute.

2. A Declaração de Montreal para IA Responsável, iniciativa da Universidade de Montreal

3. Os Princípios Gerais oferecidos na segunda versão do Design Eticamente Alinhado: Uma Visão para Priorizar o Bem-Estar Humano com Sistemas Autônomos e Inteligentes. IEEE

4. Os Princípios Éticos oferecidos na Declaração sobre Inteligência Artificial, Robótica e Sistemas 'Autônomos', publicada pelo Grupo Europeu de Ética em Ciências e Novas Tecnologias da Comissão Europeia

5. Os "cinco princípios abrangentes para um código de IA" oferecidos no relatório do Comitê de Inteligência Artificial da Câmara dos Lordes do Reino Unido

6. Os Princípios da Parceria em IA Parceria em IA

Segundo os autores esta seleção de seis PIA seguiu três critérios definidos: tempo recente, relevância e autoria. Podemos reconhecer novamente a importância dos Princípios Asilomar, que encabeça esta lista.

> Cada conjunto de princípios atende a três critérios básicos: são recentes, publicados nos últimos três anos; diretamente relevantes para a IA e seu impacto na sociedade como um todo (excluindo, portanto, documentos específicos para um determinado domínio, indústria ou setor); e altamente respeitável, publicado por organizações de várias partes interessadas com autoridade, pelo menos, de âmbito nacional. (Ibid. p. 5. Tradução nossa)

Os autores apresentam uma análise sobre os seis princípios éticos da IA selecionados, propondo um novo *framework* de cinco princípios éticos para a IA que podem orientar a pesquisa, o desenvolvimento e a implementação de tecnologias de IA de maneira mais ética e responsável. Os autores argumentam que a IA é uma tecnologia cada vez mais importante na vida das pessoas e que é crucial que seu desenvolvimento seja guiado por princípios éticos sólidos. Assim propõem os seguintes princípios: beneficência (a IA deve ser desenvolvida para melhorar o bem-estar humano), não maleficência (a IA não deve causar danos), autonomia (os indivíduos devem ter o controle sobre suas próprias informações e decisões), justiça (a IA deve ser desenvolvida e usada de forma justa e equitativa) explicam que estes princípios já fazem parte da Bioética e para IA acrescentam um princípio extra da explicabilidade (os processos de tomada de decisão da IA devem ser transparentes e explicáveis).

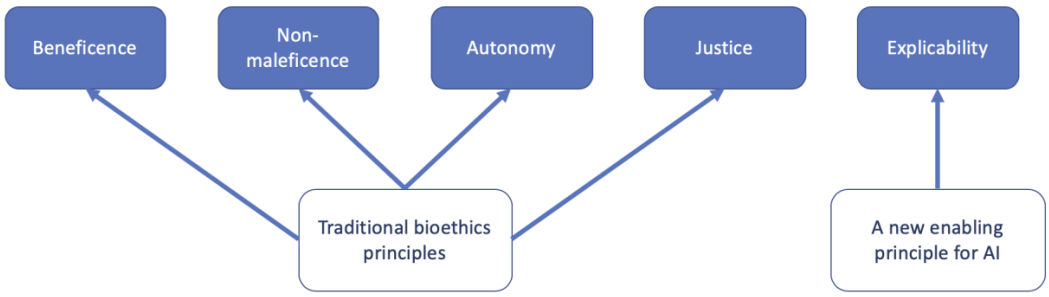


Fig.7 - Cinco princípios éticos para a IA

Os autores destacam que esses princípios são interdependentes e que devem ser aplicados em conjunto para garantir um desenvolvimento ético da IA. Argumentam que são princípios baseados em valores fundamentais, como a dignidade humana e a justiça, e que

devem ser aplicados de maneira flexível e adaptável a diferentes contextos. O artigo apresenta uma discussão de cada um dos cinco princípios e como eles podem ser aplicados na prática e oferecem exemplos do seu uso.

> De fato, o *framework* desempenhou um papel valioso no trabalho do AI4People, o primeiro fórum global da Europa sobre o impacto social da IA, que recentemente a adotou para propor 20 recomendações concretas para uma 'Boa Sociedade de IA' à Comissão Europeia (Floridi et al., 2018). Desde então, foi amplamente adotado pelas Diretrizes de Ética para IA Confiável publicadas pelo Grupo de Especialistas de Alto Nível da Comissão Europeia em Inteligência Artificial (HLEGAI 2018 e 2019), que por sua vez influenciou a Recomendação da OCDE do Conselho de Inteligência Artificial (OCDE 2019), atingindo 42 países6 (ver Tabela 1).

Os autores também destacam a importância de envolver diversas partes interessadas, como governos, empresas e a sociedade civil, na aplicação desses princípios. No final do artigo, os autores ressaltam a importância crescente da China no desenvolvimento da IA.

> De particular interesse a esse respeito é o papel da China, que já abriga a startupde IA mais valiosa do mundo (Jezard, 2018), desfruta de várias vantagens estruturais no desenvolvimento da IA (Lee & Triolo, 2017) e cujo governo tem declarado suas ambições de liderar o mundo em tecnologia de IA de ponta até 2030 (ChinaState Council, 2017). Em seu Aviso do Conselho de Estado sobre IA e outros lugares, o governo chinês expressou interesse em considerar mais a fundo o impacto social e ético da IA (Ding, 2018; Webster et. al, 2017). O entusiasmo com o uso de tecnologias também não é exclusivo dos governos, mas também é compartilhado pelo público em geral – mais na China e na Índia do que na Europa ou nos EUA, como mostra uma nova pesquisa representativa (Instituto Vodafone, 2018). (Ibid. p. 9. Tradução nossa)

2.3. Vinculando Princípios de Inteligência Artificial - Academia Chinesa de Ciências[71]

Os pesquisadores da Academia Chinesa de Ciências publicaram o artigo “Linking Artificial Intelligence Principles” sobre as leis e princípios da IA com base nas vinte e sete leis publicadas entre 2016 e 2018, traçam um panorama das leis com base na quantidade de palavras-chaves. É um breve artigo de quatro páginas, mas como mencionado anteriormente por Floridi, tem o seu valor por ser um estudo publicado pelos cientistas da China, além do resultado revelador.

[71] YI Zeng, ENMENG Lu, CUNQING, Huangfu. *Linking Artificial Intelligence Principles.* AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019), 2019.

> Os princípios da Inteligência Artificial definem considerações sociais e éticas para desenvolver a IA futura. Eles vêm de institutos de pesquisa, organizações governamentais e indústrias. Todas as versões dos princípios da IA têm considerações diferentes, abrangendo diferentes perspectivas e enfatizando diferentes. Nenhuma delas pode ser considerada completa e pode cobrir as demais propostas de princípios de IA. Aqui apresentamos o LAIP, um esforço e uma plataforma para vincular e analisar diferentes Princípios de Inteligência Artificial. Queremos estabelecer explicitamente os tópicos e links comuns entre os Princípios de IA propostos por diferentes organizações e investigar sua singularidade. Com base nesses esforços, para o futuro a longo prazo da IA, em vez de adotar diretamente qualquer um dos princípios da IA, defendemos a necessidade de incorporar vários princípios da IA em uma estrutura abrangente e focar em como eles podem interagir e completar cada um outro. (Ibid. p. 1. Tradução nossa)

Os autores discutem a importância de vincular (*link*) diferentes conjuntos de princípios éticos para a IA e desenvolver uma estrutura unificada para orientar o desenvolvimento e uso responsável. Eles analisam vários PIA incluindo os princípios de Asilomar, os princípios de Montreal, os princípios de Cingapura e os princípios da União Europeia. Eles destacam que esses conjuntos de princípios são frequentemente redundantes e apresentam sobreposições significativas. Além disso, os autores observam que esses conjuntos de princípios são frequentemente específicos para uma área de aplicação da IA e, portanto, podem não ser facilmente aplicáveis em outras áreas.

As iniciativas são classificadas por entidades por trás das iniciativas; governos, empresas, academias (junto com ONGs e OSCIPs). Em uma comparação na quantidade relativa entre as três esferas identificaram mais termo “colaboração” e menos “privacidade” e “segurança” nas leis anunciadas por empresas. Nas leis criadas por governos aparecem mais os termos “segurança” e menos “responsabilidade” e “prestação de conta”. Curiosamente, nas leis criadas por academias aparecem menos o termo “colaboração” do que outras. Apesar de ser um estudo basicamente quantitativo sem um escrutínio sobre o conteúdo das leis, podemos verificar que as entidades revelam a sua cultura e seus interesses e precauções através de elaborações das leis de IA, motivo pelo qual devemos buscar uma composição multissetorial equilibrada.

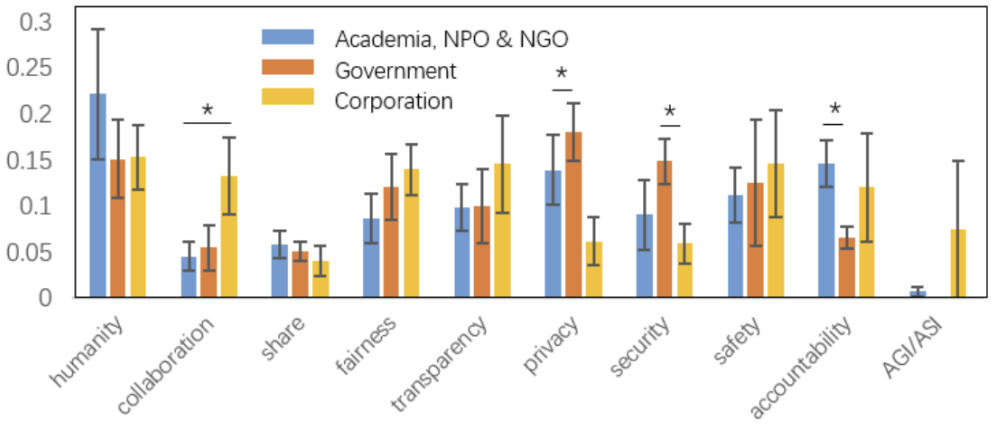


Gráfico 1 - Frequência média de tópicos em diferentes tipos de editores

Com base nessa análise, os autores propõem uma estrutura unificada para orientar o desenvolvimento e uso responsável da IA. Essa estrutura consiste em quatro princípios gerais: beneficência (a IA deve ser desenvolvida para melhorar o bem-estar humano), não maleficência (a IA não deve causar danos), autonomia (os indivíduos devem ter o controle sobre suas próprias informações e decisões) e justiça (a IA deve ser desenvolvida e usada de forma justa e equitativa). Além disso, os autores propõem quatro dimensões adicionais para a aplicação desses princípios: transparência, responsabilidade, segurança e privacidade. Eles argumentam que essas dimensões são fundamentais para garantir que a IA seja desenvolvida e usada de maneira ética e responsável.

> Diferentes Princípios de IA têm suas próprias perspectivas e cobertura para as estratégias atuais e futuras de IA. Em vez de adotar diretamente qualquer um dos princípios de IA, defendemos a necessidade de vincular e incorporar vários princípios de IA em uma estrutura abrangente e focar em como eles podem interagir e se complementar. A plataforma Linking Arti-ficial Intelligence Principles (LAIP) está disponível como um serviço online no endereço http://www.linking-ai-principles.org. Ele oferece suporte à pesquisa semântica por palavras-chave e pesquisa de parágrafos, onde princípios semanticamente semelhantes podem ser listados para exploração. (Ibid. p. 4. Tradução nossa)

Os autores apresentam uma revisão dos principais princípios éticos propostos por organizações internacionais, governos e especialistas em IA. Eles também discutem como esses princípios podem ser aplicados em diferentes contextos, como na saúde, na justiça criminal e na segurança nacional.

O artigo conclui que a aplicação dos princípios éticos na IA requer uma abordagem multidisciplinar, que envolva especialistas em ética, tecnologia, direito e políticas públicas. Além disso, é importante que os desenvolvedores de IA trabalhem em estreita colaboração com os usuários e outras partes interessadas para garantir que as preocupações éticas sejam adequadamente abordadas ao longo de todo o ciclo de vida da IA. O artigo destaca a importância de uma estrutura unificada para orientar o desenvolvimento e uso responsável da IA.

2.4. Inteligência artificial baseada em princípios: mapeando o consenso em abordagens éticas e baseadas em direitos para os princípios da IA - pesquisadores do Berkman Klein Center for Internet & Society[72]

O artigo "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*" dos pesquisadores do Berkman Klein Center for Internet & Society foi publicado em 2020. Os autores mapearam 32 PIA para a pesquisa incluindo PIA de European Union's High-Level Expert Group on AI, a UNESCO, a IEEE, a AI Now Institute, a Future of Life Institute, entre outros. Os autores analisam essas iniciativas e identificam áreas de consenso entre os diferentes conjuntos de princípios propostos.

> Juntamente com o rápido desenvolvimento da tecnologia de inteligência artificial (IA), testemunhamos uma proliferação de documentos de "princípios" destinados a fornecer orientações normativas sobre sistemas baseados em IA. Nosso desejo de comparar esses documentos – e os princípios individuais que eles contêm – lado a lado, para avaliá-los e identificar tendências, e para descobrir o momento oculto em uma conversa global fragmentada sobre o futuro da IA, resultou neste white paper e a visualização de dados associada. Esperamos que o projeto Principled Artificial Intelligence seja útil para formuladores de políticas, defensores, acadêmicos e outros que trabalham na linha de frente para capturar os benefícios e reduzir os danos da tecnologia de IA à medida que ela continua a ser desenvolvida e implantada em todo o mundo. (Ibid. p. 3. Tradução nossa)

O artigo apresenta uma abordagem sistemática e abrangente para mapear os princípios éticos da IA propostos por diferentes organizações e pesquisadores. O artigo destaca os pontos de consenso entre as abordagens, o que pode ajudar a identificar um

[72] FJELD, Jessica, et al. "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*." Berkman Klein Center for Internet & Society, 2020. Disponível em <https://dash.harvard.edu/handle/1/42160420> Acesso em maio de 2023.

conjunto comum de princípios éticos para a IA que pode ser amplamente aceito e aplicado. Os autores identificaram oito temas principais:

• Privacidade. Os princípios deste tema defendem a ideia de que os sistemas de IA devem respeitar a privacidade dos indivíduos, tanto no uso de dados para o desenvolvimento de sistemas tecnológicos quanto no fornecimento às pessoas impactadas de agenciamento sobre seus dados e decisões tomadas com eles. Os princípios de privacidade estão presentes em 97% dos documentos do conjunto de dados.

• Responsabilidade. Este tema inclui princípios relativos à importância dos mecanismos para garantir que a responsabilidade pelos impactos dos sistemas de IA seja distribuída de forma adequada e que os remédios adequados sejam fornecidos. Os princípios de responsabilidade estão presentes em 97% dos documentos do conjunto de dados.

• Segurança e proteção. Esses princípios expressam requisitos de que os sistemas de IA sejam seguros, funcionando conforme o pretendido e também seguros, resistentes a serem comprometidos por partes não autorizadas. Os princípios de segurança e proteção estão presentes em 81% dos documentos do conjunto de dados.

• Transparência e explicabilidade. Os princípios deste tema articulam requisitos para que os sistemas de IA sejam projetados e implementados para permitir a supervisão, inclusive por meio da tradução de suas operações em resultados inteligíveis e do fornecimento de informações sobre onde, quando e como estão sendo usados. Os princípios de transparência e explicabilidade estão presentes em 94% dos documentos do conjunto de dados.

• Equidade e não discriminação. Com as preocupações sobre o viés da IA já afetando indivíduos globalmente, os princípios de justiça e não discriminação exigem que os sistemas de IA sejam projetados e usados para maximizar a justiça e promover a inclusão. Os princípios de justiça e não discriminação estão presentes em 100% dos documentos do conjunto de dados.

• Controle Humano da Tecnologia. Os princípios sob este tema exigem que decisões importantes permaneçam sujeitas à revisão humana. Os princípios do Controle Humano da Tecnologia estão presentes em 69% dos documentos do conjunto de dados.

• Responsabilidade Profissional. Esses princípios reconhecem o papel vital que os indivíduos envolvidos no desenvolvimento e implantação de sistemas de IA desempenham nos impactos dos sistemas e exigem seu profissionalismo e integridade para garantir que as partes interessadas apropriadas sejam consultadas e os efeitos de longo prazo sejam planejados. Os princípios de Responsabilidade Profissional estão presentes em 78% dos documentos do conjunto de dados.

• Promoção dos Valores Humanos. Finalmente, os princípios dos Valores Humanos afirmam que os fins aos quais a IA é dedicada e os meios pelos quais ela é implementada devem corresponder aos nossos valores fundamentais e geralmente promover o bem-estar da humanidade. Os princípios da promoção dos valores humanos estão presentes em 69% dos documentos do conjunto de dados.

O objetivo principal do artigo foi mapear o consenso sobre os princípios éticos para a IA que emergiram em várias iniciativas de especialistas em todo o mundo. Os autores também destacam as diferentes ênfases que as diferentes iniciativas colocam em cada tema e apresentam uma matriz que mapeia os diferentes conjuntos de princípios em relação a esses temas.

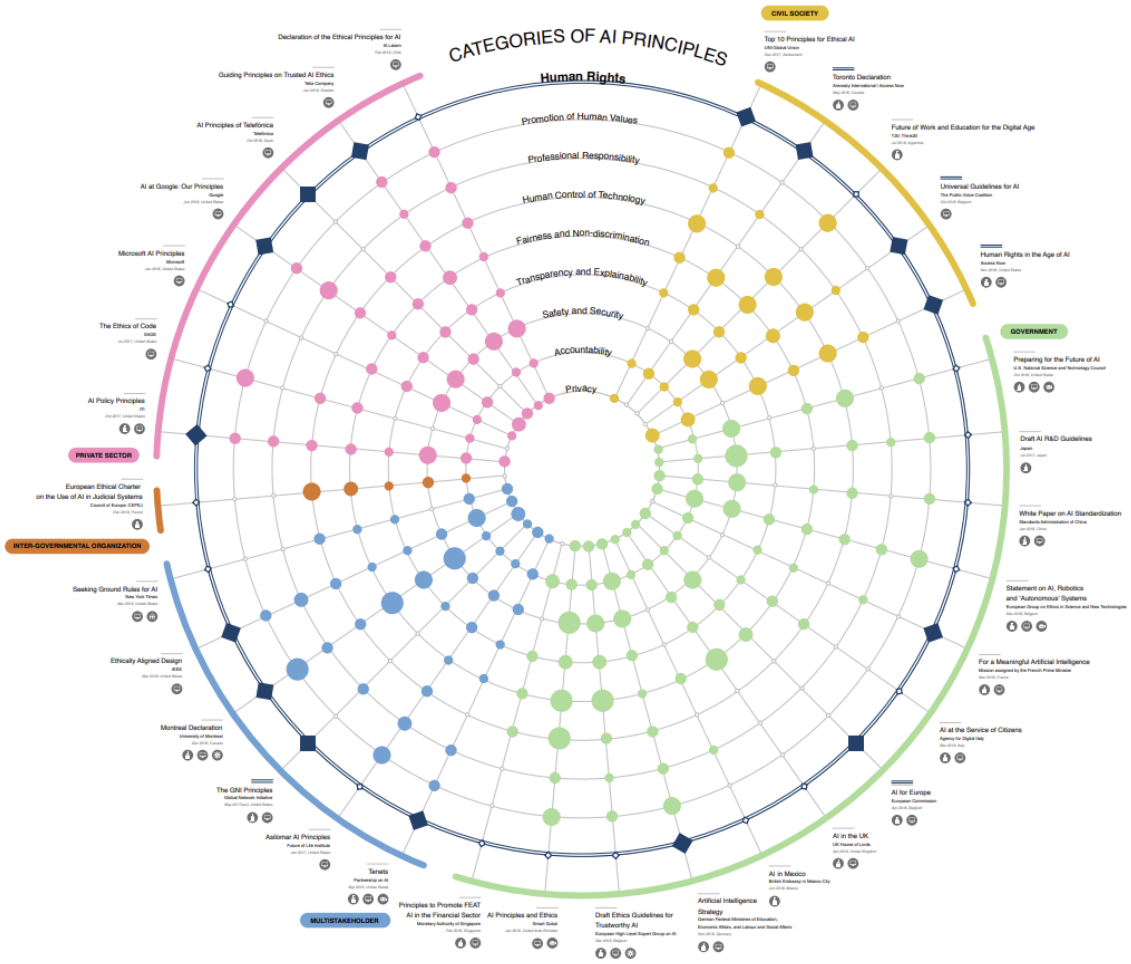


Fig.8 - Um mapa de abordagens éticas e baseadas em direitos[73]

[73] Essa visualização apresenta trinta e dois conjuntos de princípios lado a lado, permitindo a comparação entre esforços de governos, empresas, grupos de defesa e iniciativas de várias partes interessadas. Isto destaca oito temas compartilhados: responsabilidade, justiça e não discriminação, controle humano de tecnologia, privacidade, responsabilidade profissional, promoção de valores humanos, segurança e proteção, e transparência e explicabilidade - e documentos onde é feita referência a normas internacionais direitos humanos. Nosso conjunto de dados não é exaustivo, mas sim uma amostra de IA proeminente e recente princípios. (Ibid. p. 8 e 9. Tradução nossa)

Uma das principais conclusões do artigo é que existe um consenso crescente em torno dos princípios éticos para a IA, especialmente em relação à transparência, responsabilidade e justiça. No entanto, os autores destacam que ainda há desafios significativos na implementação desses princípios na prática e que é necessário um esforço conjunto de especialistas, formuladores de políticas, empresas e sociedade civil para garantir que a IA seja desenvolvida e usada de maneira ética e responsável.

O artigo é uma análise importante das iniciativas de princípios éticos para a IA e identifica áreas de consenso e divergência. Ele destaca a importância de uma abordagem colaborativa e multidisciplinar para garantir que a IA seja desenvolvida e usada de maneira ética e responsável.

2.5. Considerações

Os quatro artigos abordam PIA com o intuito de estabelecer diretrizes éticas para o desenvolvimento e uso responsável da IA, apresentando semelhanças e diferenças, além de distintas metodologias adotadas. Os textos reconhecem a importância da transparência e responsabilidade na concepção e implementação de sistemas de IA. Todos eles enfatizam a necessidade de assegurar que a IA seja empregada para o benefício da sociedade como um todo, fomentando o bem-estar humano e evitando danos desnecessários. Os artigos sublinham a relevância de promover a diversidade e inclusão no desenvolvimento da IA, garantindo que ela atenda às necessidades e valores de diferentes grupos e evite o agravamento de desigualdades existentes.

Verifica-se que existem basicamente dois tipos de métodos aplicados nos quatro estudos anteriores: uma análise mais detalhada de uma iniciativa específica, como foi o caso dos pesquisadores da Federação de Cientistas da Alemanha (VDW) sobre os Princípios de Asilomar, e outros três artigos que selecionaram uma variedade de iniciativas para comparar conteúdo, identificando diferenças e convergências entre elas, extraindo palavras-chave, iniciativas e temas, como nos artigos de Luciano Floridi, pesquisadores da Academia Chinesa de Ciências e pesquisadores do Berkman Klein Center for Internet & Society. Apesar da similaridade na metodologia, mesmo nesses três artigos há diferenças. Por exemplo, o artigo

de Floridi e Cowls propõe um framework unificado de cinco princípios, enquanto o artigo do Berkman Klein Center mapeia o consenso em abordagens éticas e baseadas em direitos.

As perspectivas culturais e geográficas também podem influenciar os resultados. O artigo da Federação de Cientistas da Alemanha (VDW) aborda a perspectiva alemã, enquanto o da Academia Chinesa de Ciências reflete a perspectiva chinesa. Essas diferenças podem influenciar os princípios e diretrizes propostos. O artigo da Academia Chinesa de Ciências dá maior ênfase à colaboração internacional na governança da IA, enquanto o artigo do Berkman Klein Center foca mais nos direitos humanos e abordagens baseadas em direitos. Este é o motivo pelo qual PIA devem ser resultado de uma iniciática ampla e global. Ressaltamos aqui o trecho da "Recomendações sobre a inclusão da África subsaariana na ética global da IA"**:**

> […] Apesar da natureza global das implicações éticas da inteligência artificial, a atenção até o momento se concentrou principalmente nos EUA e na UE, com uma crescente conscientização sobre a China, especialmente suas crescentes capacidades de IA, seu impacto no Sul Global e na ordem geopolítica global. Apesar da clara necessidade de entender como a IA afeta as pessoas em todo o mundo, uma perspectiva verdadeiramente global continua sendo um ponto cego crítico na conversa ética. [74]

[74] Disponível em <https://researchictafrica.net/wp/wp-content/uploads/2020/11/RANITP2019-2-AI-Ethics.pdf> Acesso em junho de 2023. p. 2. Tradução nossa.

PARTE II CAPÍTULO 1 – FÓRMULA BÁSICA DAS LEIS

1.1 - As primeiras leis da sociedade

As três leis de Asimov - que inauguram leis para robôs e posteriormente para Inteligência Artificial – foram anunciadas em 1942. Em 1947, apenas 5 anos mais tarde, algo que parece mais uma dessas coincidências na história acontece, um código de lei promulgado por um rei chamado Lipit-Ishtar, que precedeu Hamurabi por cerca de cento e cinquenta anos foi traduzido[75]. O código de Hamurabi, até então conhecido como o mais antigo conjunto de leis, havia sido descoberto em 1901 e traduzido em 1902. O código de Lipit-Ishtar, como hoje é conhecido, foi encontrado em tábuas de argila endurecidas ao sol. Estas tábuas inscritas em cuneiforme foram escavadas no início do séc. XX, mas permaneceram sem identificação por cerca de 50 anos.

Em 1948, Taha Baqir[76] - curador do Museu do Iraque em Bagdá, descobriu em Nippur, uma antiga cidade suméria situada no atual Iraque, duas tábuas com código de lei parecido com o código de Hamurabi. As tábuas foram traduzidas para o inglês por assiriólogo Samuel Kramer em 1952. Assim o código de Ur-Nammu, o mais antigo código de lei conhecido pela humanidade, que precede o código de Hamurabi em aproximadamente 300 anos foi revelado. A fama do Código de Lipit-Ishtar como o mais antigo durou cerca de 5 anos.

Segundo Fernanda Pirie, professora de Antropologia Jurídica da Universidade de Oxford e autora do livro "The Rule of Laws", em 2122 a.C., Ur-Nammu[77], um líder militar, tornou-se o novo rei da cidade mesopotâmica de Ur. Ur-Nammu, prometendo corrigir as injustiças, introduziu medidas para proteger a população mais vulnerável[78].

[75] Disponível em <https://www.journals.uchicago.edu/doi/abs/10.2307/500752?journalCode=aja> Acesso em outubro de 2022.
[76] Disponível em <https://www.penn.museum/sites/bulletin/3637/> Acesso em outubro de 2022.
[77] Fernanda Pirie usa “Ur-Namma” em seu livro, mas “Ur-Nammu” é a grafia mais utilizada.
[78] PIRIE, Fernanda. *The Rule of Laws: A 4,000-Year Quest to Order the World*. London. Profile Books. 2021. p. 17.

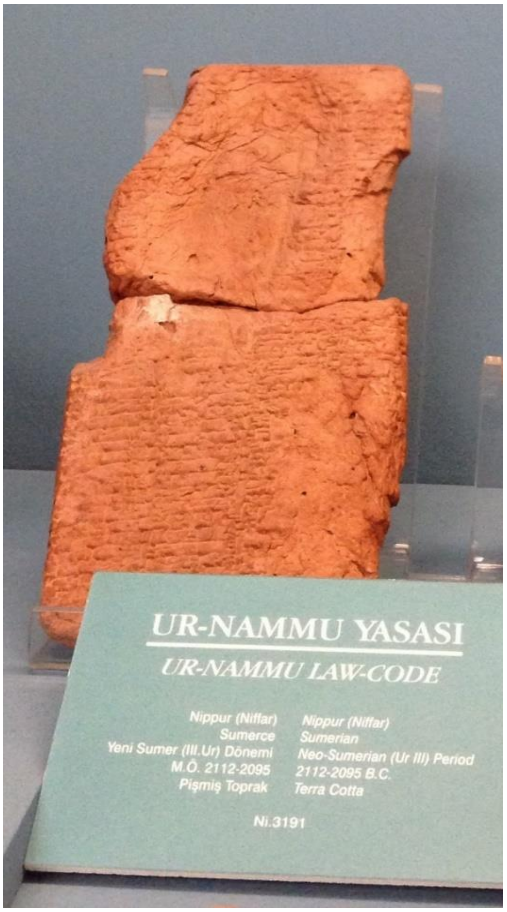


Fig. 9 - Código de Ur-Nammu - Museus de Arqueologia de Istambul[79]

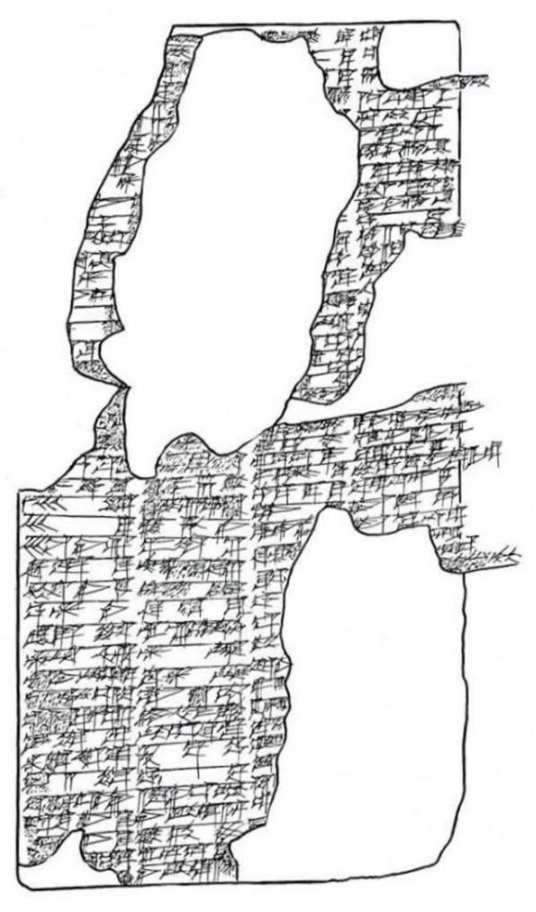


Fig.10 - Código de Ur-Nammu, uma cópia do anverso contendo o prólogo do código[80]

Embora o Código de Ur-Nammu seja atualmente o código de leis mais antigo conhecido, é provável que muitos predecessores de Ur-Nammu também tenham elaborado suas próprias leis. A inovação de Ur-Nammu, para Pirie (2021), foi a criação de um conjunto de leis que os escribas registraram em tábuas de argila. Na perspectiva contemporânea, essas leis podem parecer triviais, contendo instruções sobre pagamento de compensações e punições. Entretanto, são as leis mais antigas descobertas por arqueólogos em qualquer parte do mundo até hoje, e estão na origem de uma tradição jurídica que se desenvolveu ao longo dos milhares de anos seguintes, servindo como exemplo para legisladores em diversos lugares.

[79] Foto por Osama Shukir Muhammed Amin, Istanbul Archaeological Museums/Ancient Orient Museum, Creative Commons.
[80] Disponível em <https://www.penn.museum/sites/bulletin/3637/> Acesso em outubro de 2022.

Mesmo após o colapso das civilizações da Mesopotâmia, sua tradição legal sobreviveu, inspirando as leis da nossa sociedade.

Apesar de não termos uma certeza sobre o pioneirismo na criação das primeiras leis na história da civilização, segundo Martha T. Roth (1995), o que é certo, porém, é que na Mesopotâmia, o clima seco preservou estas primeiras tábuas de argila em cuneiforme. Essas relíquias oferecem evidências de que, no terceiro milênio a.C., os reis da Mesopotâmia elaboravam leis. [81]

O código de Ur-Nammu, Lipit-Ishtar e Hamurabi apresentam formatos semelhantes, todos iniciando com um prólogo. Para sociólogo italiano Domenico de Masi, o prólogo ou incipit de uma obra, seja uma escritura ou preceitos morais, já revelam o seu teor.

> Muitos modelos são inspirados por um conjunto preciso de sagradas escrituras ou até meros preceitos morais. Nesses casos, basta ler o *incipit* dos textos de referência para perceber as primeiras diferenças que os separam e as primeiras afinidades que os que unem. (MASI. 2014. p.19)[82]

Podemos notar que os três prólogos dos códigos de leis da Mesopotâmia revelam mais afinidades e semelhanças que diferenças. O prólogo do código de Ur-Nammu narra como Ur-Nammu foi designado pela divindade para receber as leis. Ur-Nammu é apresentado como estabelecedor de equidade na terra. Podemos verificar que a legitimação divina é um recurso comum desde os códigos iniciais verificados também nos códigos Lipit-Ishtar e Hamurabi.

> Depois que An e Enlil transferiram o reinado de Ur para Nanna, naquela época Ur-Nammu, filho nascido de Ninsun, para sua amada mãe que o gerou, de acordo com seus princípios de equidade e verdade... Então Ur-Nammu, o poderoso guerreiro, rei de Ur, rei da Suméria e Akkad, pelo poder de Nanna, senhor da cidade, e de acordo com a verdadeira palavra de Utu, estabeleceu equidade na terra; ele baniu a maldição, a violência e conflitos, e fixou as despesas mensais do Templo em 90 gur de cevada, 30 ovelhas e 30 sila de manteiga. relação a uma mina. (Prólogo Ur-Nammu)
>
> Quando o grande An, o pai dos deuses, (e) Enlil, o rei de toda a terra, o governante que estabelece as decisões, …Ninisin, a filha de An, … para ela… alegria... por sua testa clara; quando eles deram a ela o reino da Suméria e Akkad (e) a regra favorável em sua (cidade) Isin, ... estabelecido por An. Quando An(i) Enlil chamou Lipit-Ishtar, então Lipit-Ishtar, o sábio pastor cujo nome foi pronunciado Nunamnir, para reinar sobre a terra, estabelecer justiça na terra, remover queixas, repelir o(s) inimigo(s) a

[81] ROTH, T. Martha. *Law collections from Mesopotâmia and Asia Minor*. Atlanta: Scholars Press. 1995. p. 16-17.
[82] MASI, Domenico de. *O futuro chegou*. Editora Casa da Palavra. RJ. 2014. p.19.

> evitar a rebelião pela força das armas, (e) trazer prosperidade aos sumérios e acadianos, então eu, Lipit-Ishtar, o humilde pastor de Nippur, o virtuoso fazendeiro de Ur, que não deixa Eridu, digno ser governante de Erech, rei de Isin, rei da Suméria e Acádia, que era querido ao coração de Inan, estabeleceu a justiça na Suméria e na Acádia de acordo com as palavras de Enlil. (Prólogo Lipit-Ishtar)
>
> Quando o alto Anu, Rei de Anunaki e Bel, Senhor da Terra d dos Céus, determinador dos destinos do mundo, entregou o governo de toda humanidade a Marduk... quando foi pronunciado o alto nome da Babilônia; quando ele a fez famosa no mundo e nela estabeleceu um duradouro reino cujos alicerces tinham a firmeza do céu e da terra - por esse tempo de Anu e Bel me chamaram, a mim, Hamurabi, o excelso príncipe, o adorador dos deuses, para implantar a justiça na terra, para destruir os maus e o mal, para prevenir a opressão do fraco pelo forte... para iluminar o mundo e propiciar o bem-estar do povo. Hamurabi, governador escolhido por Bel, sou eu, eu o que trouxe a abundância à terra; o que fez obra completa para Nippur e Durilu; o que deu vida à cidade de Uruk; o que supriu água com abundância aos seus habitantes... o que tornou bela a cidade de Borsippa... o que enceleirou grãos para a poderosa Urash... o que ajudou o povo em tempo de necessidade; o que estabeleceu a segurança na Babilônia; o governador do povo, o servo cujos feitos são agradáveis a Anunit. (Prólogo do Código Hamurabi)[83]

É muito provável que o código de Ur-Nammu influenciou outros códigos na região da Mesopotâmia. Os códigos de Hamurabi que vieram cerca de trezentos anos depois, apresentam claramente estas influências. No prólogo, Hamurabi se apresenta como um governante na tradição mesopotâmica, reconhecido pelas divindades, seguindo o mesmo formato utilizado anteriormente por Ur-Nammu.

> Estas são as decisões judiciais que Hamurabi, o rei, estabeleceu para trazer a verdade e uma ordem justa em sua terra. em minha estela, leia para ele, para que ele possa entender meus preciosos comandos; e que minha estela demonstre sua posição, para que ele entenda seu caso e acalme seu coração... Eu sou Hamurabi, rei da justiça, a quem Shamash concedeu a verdade. [84]

Pelos prólogos, podemos verificar que os primeiros legisladores da civilização, prometeram justiça e ordem na sociedade através do estabelecimento das leis: “Então Ur-Nammu, o poderoso guerreiro, rei de Ur, rei da Suméria e Akkad, pelo poder de Nanna, senhor da cidade, e de acordo com a verdadeira palavra de Utu, estabeleceu equidade na terra”; “Eu decidi que deve haver justiça em Sumer e em Akkad, que o país deve prosperar: quem pode opor-se à minha decisão? Eu, Lipit-Ishtar, conduzi a meu povo: quando poderá ser anulada

[83] Disponível em <https://www.ancient-origins.net/artifacts-ancient-writings/code-ur-nammu-sumerians-009333> Acesso em novembro de 2022. Tradução nossa.
[84] Ibid. p.24-25.

minha sentença?"; "Estas são as decisões judiciais que Hamurabi, o rei, estabeleceu para trazer a verdade e uma ordem justa em sua terra".

No final do prólogo, três leis são anunciadas no código de Ur-Nammu como se resumissem o espírito contido nas leis seguintes: "O órfão não será entregue ao homem rico, a viúva não será entregue ao poderoso, o homem de um siclo não será entregue ao homem de uma mina (60 siclos). Eliminei a inimizade, a violência e os clamores por justiça."

Após o prólogo, o código de Ur-Nammu apresenta as leis casuísticas em formato:

se - (crime) – então – (punição):

1. Se um homem comete um assassinato, esse homem deve ser morto.
2. Se um homem cometer um roubo, ele será morto.
3. Se um homem cometer um sequestro, ele deve ser preso e pagar 15 siclos de prata.
4. Se um escravo se casar com uma escrava e essa escrava for libertada, ele não deixará a casa.
5. Se um escravo se casar com uma pessoa nativa (ou seja, livre), ele deve entregar o filho primogênito ao seu dono.
6. Se um homem violar o direito de outro e deflorar a esposa virgem de um jovem, eles devem matar esse homem.
7. Se a esposa de um homem seguiu outro homem e ele dormiu com ela, eles devem matar aquela mulher, mas aquele homem será libertado.
8. Se um homem procedeu à força e deflorou a escrava virgem de outro homem, esse homem deve pagar cinco siclos de prata.
9. Se um homem se divorciar de sua primeira esposa, ele deverá pagar a ela uma mina de prata.
10. Se for uma (antiga) viúva de quem ele se divorciar, ele deverá pagar a ela meia mina de prata.
11. Se o homem dormiu com a viúva sem ter havido contrato de casamento, ele não precisa pagar nenhuma prata.
13. Se um homem for acusado de feitiçaria, ele deve passar pela provação da água; se ele for provado inocente, seu acusador deve pagar 3 siclos.
14. Se um homem acusar a esposa de outro homem de adultério, e a provação no rio provar que ela é inocente, então o homem que a acusou deve pagar um terço de mina de prata.
15. Se um futuro genro entrar na casa de seu futuro sogro, mas seu sogro posteriormente entregar sua filha a outro homem, o sogro retornará ao genro rejeitado. sogro dobrou a quantidade de presentes de noiva que ele trouxe.
17. Se um escravo escapar dos limites da cidade e alguém o devolver, o dono pagará dois siclos a quem o devolveu.
18. Se um homem arrancar o olho de outro homem, ele pesará ½ mina de prata.
19. Se um homem cortar o pé de outro homem, ele pagará dez siclos.
20. Se um homem, durante uma briga, quebrar o membro de outro homem com um porrete, ele deverá pagar uma mina de prata.
21. Se alguém cortar o nariz de outro homem com uma faca de cobre, deverá pagar dois terços de mina de prata.
22. Se um homem arrancar o dente de outro homem, ele pagará dois siclos de prata.

> 24. [...] se não tiver escravo, deverá pagar 10 siclos de prata. Se não tiver prata, deverá dar outra coisa que lhe pertença.
> 25. Se a escrava de um homem, comparando-se com sua amante, fala insolentemente com ela, sua boca deve ser limpa com 1 litro de sal.
> 28. Se um homem compareceu como testemunha e foi provado ser um perjuro, ele deve pagar quinze siclos de prata.
> 29. Se um homem comparecer como testemunha, mas retirar o juramento, deverá efetuar o pagamento, na medida do valor em litígio da causa.
> 30. Se um homem cultivar furtivamente o campo de outro homem e ele fizer uma reclamação, esta deve ser rejeitada, e este homem perderá suas despesas.
> 31. Se um homem inundou o campo de outro homem com água, ele medirá três kur de cevada por iku de campo.
> 32. Se um homem arrendou um campo arável para um (outro) homem para cultivo, mas ele não o cultivou, transformando-o em terreno baldio, ele medirá três kur de cevada por iku de campo.[85]

Pelas leis 4, 17, 25 podemos saber que havia leis dirigidas diretamente para escravos e não para os senhores de escravos.[86] Por mais que os escravos eram desumanizados e tratados como mercadorias, demonstra o reconhecimento como um agente capacitado para cumprimentos das leis como qualquer outro cidadão pelos legisladores desde as primeiras leis da civilização, como podemos verificar no código de Ur-Nammu.

No epílogo, Lipit-Ishtar exorta aos futuros legisladores a manterem o código vigente, desejando sucesso se assim cumprisse, e maldição em caso contrário. Segundo Fernanda Pirie (2021), o potencial dessa nova técnica que prometia ordem e justiça foi rapidamente apreciado por outros governantes da Mesopotâmia e arredores. A Babilônia era um centro comercial, recebendo comerciantes e outros visitantes da Índia, Ásia Central, Pérsia, Arábia, Egito, Armênia e Grécia. Os comerciantes teriam apreciado os benefícios de usar ideias legais e adotadas para sua terra natal. Mais que os códigos iniciais como Ur-Nammu e Lipit-Ishtar,

[85] Disponível em <http://realhistoryww.com/world_history/ancient/Misc/Sumer/ur_nammu_law.htm> Acesso em novembro de 2022. Tradução nossa.

[86] Nos próximos capítulos algumas leis que previam responsabilidade de donos de escravos serão analisadas. O historiador Paul E. Lovejoy resumiu as características da condição de escravo numa lista com sete itens:

1- O escravo é uma propriedade.
2- É objeto de compra e venda, como qualquer outra mercadoria.
3- Mesmo que o reconheçam como ser humano, é um estrangeiro por natureza, arrancado do seu meio familiar e social.
4- A relação entre senhor e escravo é baseado na violência.
5- Seu trabalho está sempre à disposição do seu dono.
6- Cabe também ao senhor o controle da procriação do cativo, cujo filhos não lhe pertencem. Sua própria sexualidade não lhe pertence.
7- A escravidão é hereditária, passa de pai para filho.

GOMES. Laurentino. *Escravidão Vol. I As Origens*. Ed. Globo. Rio de Janeiro. 2019. pag. 68, 69.

o código de Hamurabi atingiu o ápice das leis da Mesopotâmia tornando-se o mais influente. Poderia até reconhecer assim que usufruiu do sucesso desejado por Lipit-Ishtar pelo cumprimento da exortação. As leis de Hamurabi deixaram um longo legado na região e foram amplamente adotadas por seus sucessores. Enquanto durou o império, o código de Hamurabi foi usado como um exercício de escrita, um modelo de escrita legal para escribas treinarem o seu ofício. Depois de Hamurabi, os assírios estabeleceram império na Babilônia. Uma das primeiras coisas que fizeram foi criar leis em forma casuística semelhante às leis mesopotâmicas anteriores.

O grande mérito destes primeiros legisladores, para Fernanda Pirie (2021), foi a criação de um padrão objetivo ao qual qualquer cidadão poderia se referir, proporcionando previsibilidade para comerciantes, e resolver problemas sociais, especificando penalidades para crimes, compensação por lesões, regras para contratos. Ao definir direitos e deveres, esses primeiros legisladores estavam criando ordem e categorias, especificando as relações entre elas e oferecendo uma estrutura mais permanente à sociedade.

As leis são assuntos que também ocupam principais obras da tradição da filosófica. A origem das leis é o tema que inicia o diálogo de As Leis de Platão. O diálogo se inicia com a seguinte pergunta de um dos personagens.

> **O ateniense**: A quem atribuis, estrangeiro, a autoria de vossas disposições legais? A um deus ou a algum homem?
>
> **Clínias**: A um deus, estrangeiro, com toda a certeza a um deus. Nós cretenses chamamos de Zeus o nosso legislador, enquanto na Lacedemônia (ou Esparta), onde nosso amigo aqui tem seu domicílio, afirmam — acredito — ser Apolo o deles. Não é assim, Megilo? [87]

Seguindo a tradição da Mesopotâmia, também na Grécia Antiga, as leis eram muitas vezes atribuídas a uma divindade, neste caso Zeus (para os cretenses) e Apolo (para os lacedemônios ou espartanos). Esta atribuição enfatiza o status sagrado e inviolável das leis. Se as leis fossem consideradas como sendo dadas por um deus, elas carregariam uma autoridade moral suprema que não poderia ser desafiada. Ao atribuir as leis a uma divindade, os gregos antigos mostravam a interação entre o divino e o humano na organização da

[87] Disponível em <https://www.democracia.org.br/wp-content/uploads/2019/02/Plat%C3%A3o-As-Leis.pdf> Acesso em novembro de 2022.

sociedade. As leis não eram apenas um contrato social, mas também um contrato sagrado. A moralidade e a justiça eram vistas não apenas como princípios sociais, mas também como princípios divinos. O diálogo também indica que diferentes regiões da Grécia tinham suas próprias tradições e divindades legisladoras.

> De resto, seja como deus quiser: agora é preciso obedecer à lei e em defender. [...] Tu. ao contrário, evitaste encontrar-me e instruir-me, não o quiseste; e me conduzes aqui, onde a lei ordena citar aqueles que tem necessidade de pena e não de instrução. [...] E não vos encolerizeis comigo, porque digo a verdade; não há nenhum homem que se salve, se quer opor-se, com franqueza, a vós ou a qualquer outro povo, e impedir que muitos atos contrários à justiça e às leis se pratique na cidade. E não há outro caminho: quem combate verdadeiramente pelo que é justo, se quer ser salvo por algum tempo, deve viver a vida privada, nunca se meter nos negócios públicos. [...] Que o juiz não ceda já por isso, não dispense sentença a favor, mas a pronuncie retamente e jure condescender com quem lhe agrada, mas proceder segundo as leis.[88]

Na Apologia de Sócrates, percebemos a importância que Sócrates atribui à lei. Para ele, obedecer à lei é uma necessidade e um princípio que norteia a conduta do cidadão. Sócrates ressalta a lei como a entidade que se deve recorrer quando se precisa de pena, e não de instrução, sugerindo que a lei tem a finalidade de manter a ordem e a justiça, e não apenas de educar. Sócrates ainda enfatiza a necessidade de oposição aos atos contrários à justiça e às leis. Ele sugere que quem luta verdadeiramente pelo que é justo deve se manter afastado da vida pública e dos negócios públicos. Além disso, Sócrates destaca o papel do juiz como uma figura chave na aplicação da lei. Ele argumenta que os juízes não devem ceder a favoritismos ou preferências pessoais, mas sim pronunciar sentenças justas de acordo com as leis. Este é um princípio fundamental da justiça, que requer que os juízes sejam imparciais e se baseiem na lei ao tomar suas decisões.

Uma vez criada e praticada pelos sucessivos reis da Mesopotâmia, a tradição legal nunca mais deixou fazer parte da humanidade. De inúmeras interações entre os povos de diversas localidades através de comércios ou até mesmo por guerras e dominações, as leis de uma sociedade acabaram influenciando outras tanto regionalmente como temporalmente. Muitos séculos depois, Hobbes reconheceu a superioridade das leis como técnicas criadas e como elas são indispensáveis para governantes e toda sociedade. Pois as leis,

[88] Disponível em <https://edisciplinas.usp.br/pluginfile.php/270801/mod_resource/content/1/platao%20apologia%20de%20socrates.pdf> Acesso em novembro de 2022.

fundamentalmente, oferecem mecanismos para trazer ordem na sociedade garantindo direitos básicos de seus membros. Regulam problemas inevitáveis que surgem na convivência de pessoas numa sociedade.

> [...] existem dois modos de lutar: um com as leis, outro com a força. O primeiro é um método próprio do homem, o segundo, dos animais. [...] um príncipe deve saber como utilizar esses dois métodos, e ter sempre em mente que um deles sem o outro não produz efeitos duradouros. (MORRISON apud Maquiavel. 2012. p. 89)

Para Pirie (2021), a tradição legal ocidental, a partir da criação de primeiros legisladores da Mesopotâmia passando pelos gregos e romanos, também recebeu influências da Índia e China. A notável expansão econômica, tecnológica e militar permitiu então exportar suas leis ao redor do mundo, alegando que trariam "civilização" aos povos indígenas e varreriam o modelo ultrapassado de ordem "despótica" ou "primitiva". No século XX, isso se tornou a visão de uma ordem internacional na qual governos devidamente eleitos promovem a paz e a prosperidade, defendem a democracia e respeitam os direitos humanos. É o equivalente à ordem cosmológica invocada pelos imperadores chineses e à ordem do dharma elaborada pelos brâmanes hindus que as potências colonizadoras tanto desejaram substituir. O avanço da civilização ocidental nos últimos três séculos, foi acompanhado também pela difusão de suas leis para o resto do mundo.

1.2 - Tipos de leis

Qual a natureza dos PIA que foram anunciados como leis, princípios, regras, declarações por iniciativas de governos, academias, empresas, centro de pesquisas e outras iniciativas? Verificamos na Parte I deste trabalho que diversos nomes são empregados para PIA com o predomínio do termo lei na primeira fase e princípio na segunda fase. Esta polissemia não se deve apenas à falta de clareza na definição e tempo para consolidação nas elaborações dos PIA. O próprio termo lei é amplamente empregado em aplicações distintas. Provavelmente essa origem difusa deve ser também uma das causas dessa profusão de uso nos PIA.

Quando falamos em leis da robótica ou Inteligência Artificial que tipo de leis estamos nos referindo? Nesta seção, analisaremos de forma abrangente outros tipos de leis para oferecer uma delimitação para as leis da IA e situá-las adequadamente, dentro do propósito

deste trabalho, passando em largo sem entrarmos em detalhes. Para iniciar com o estudo das leis para IA, introduzimos no capítulo anterior o surgimento das primeiras leis na civilização. A partir deste surgimento, o termo lei ganhou uma vida própria, tornando-se presente em um variado número de aplicações. O termo lei e suas variações como normas, regras, princípios, códigos são largamente utilizados para distintas áreas como jurídica, ciência, técnica, social, entre outros.

A palavra "lei" é empregada em pelo menos quatro diferentes contextos: Leis Físicas ou Científicas ou Lei da Natureza, Leis Sociais, Leis do Senso Comum e Leis Jurídicas, entre outros. Essas categorias de leis não são mutuamente exclusivas e pode haver sobreposições ou interconexões entre elas. As Leis Físicas ou Científicas[89], também conhecidas como Leis da Natureza, são os princípios e regularidades que governam os fenômenos naturais. Essas leis são formuladas pela ciência e descrevem as relações causais entre os elementos do mundo físico. Elas são universais e se aplicam consistentemente em todos os lugares e em todos os tempos.

Leis Sociais[90] são os princípios e padrões que regulam o comportamento e as interações dos indivíduos em uma sociedade. Essas leis são estabelecidas pelas normas, costumes e convenções sociais, e podem variar de acordo com a cultura, a época e o contexto social. Elas visam manter a ordem, a coesão social e a convivência entre os membros de uma comunidade. As normas sociais, como muitos outros fenômenos sociais, são o resultado não planejado da interação dos indivíduos. Tem-se argumentado que as normas sociais devem ser entendidas como uma espécie de gramática das interações sociais. Exemplos de leis sociais incluem normas de etiqueta, regras de conduta em instituições como escolas e empresas, e leis de trânsito.

Leis do Senso Comum são princípios práticos e observações gerais baseadas na experiência cotidiana e no conhecimento comum de uma sociedade. Embora não sejam leis no sentido jurídico, elas são consideradas como guias informais para a conduta e a tomada de decisões. As leis do senso comum refletem conhecimentos básicos e poderiam ser

[89] Disponível em <https://plato.stanford.edu/entries/lawphil-nature/> Acesso em outubro de 2022.
[90] Disponível em <https://plato.stanford.edu/entries/social-norms/> Acesso em outubro de 2022.

classificadas como descritivas e normativas. Exemplos de leis do senso comum incluem "não falar com a boca cheia" e "olhe para os dois lados antes de atravessar a rua". Duhem (2019) citou um exemplo de lei do senso comum, “todo homem é mortal”.

Leis Jurídicas são as regras estabelecidas por autoridades governamentais ou sistemas jurídicos para regular a conduta dos indivíduos e proteger os direitos e interesses da sociedade. Essas leis são codificadas em legislações e são aplicadas por meio de instituições judiciais e do sistema de justiça. As leis jurídicas podem abranger uma ampla gama de áreas, incluindo direito civil, direito penal, direito constitucional, direito trabalhista, entre outros. Elas são vinculativas e têm consequências legais quando infringidas.

Ainda pontuamos a distinção entre Lei da Natureza e Lei Natural. A primeira é usada juntamente com Leis Físicas ou Científicas. Por outro lado, a Lei Natural é um princípio que se acredita ser inerente à natureza humana e identificável pela razão humana. Na teologia, pode referir-se à ordem moral que Deus inscreveu no mundo e nas criaturas humanas. Em ambos os casos, a Lei Natural é sobre como as coisas "deveriam ser", em vez de apenas descrever como elas são. Por exemplo, a questão proposta pela Academia de Dijon que originou a famosa obra de Rousseau “Discurso sobre a origem e os fundamentos da desigualdade entre os homens”, trata-se da origem da desigualdade entre os homens e se é autorizada pela lei natural.

O jurista brasileiro Daniel Coelho Souza definiu seis categorias para a normatividade social, o conjunto de regras escritas ou orais de origem determinada ou indeterminada, que tutelam a conduta do homem em grupo impondo-lhes deveres positivos ou negativos. Exceto as normas técnicas, todas outras poderiam ser incluídas em Leis Sociais e Jurídicas.

> Normas técnicas: Regras que indicam a maneira para alcançar um fim determinado. Não tem obrigatoriedade em si. Não possuem análise valorativa.
>
> Normas éticas: Disciplinam o comportamento do homem, quer no íntimo subjetivo quer no exterior social. Prescrevem deveres para a realização de valores. Não implicam apenas juízo de valor, mas impõem a escolha de uma diretriz considerada obrigatória.
>
> Normas religiosas: consideradas como emanadas da divindade ou por ela sancionadas, e também outras tornadas obrigatórias pela autoridade

> religiosa. É unilateral e autônoma. É aplicada internamente. Incoercível, não tem capacidade de punir o indivíduo, possui sanções (consequência) pré-fixadas.
>
> Normas morais: regras de conduta desprovidas de coerção que só prescreve deveres. É unilateral. Autônoma, não sofre influencias externas. É aplicada em seu interior, na mente do indivíduo. A moral é incoercível e possui sanções difusas, variáveis, amplas.
>
> Normas de trato social: padrão de conduta social elaboradas pela sociedade e que não resguardando os interesses de segurança do homem, visam tornar o ambiente social mais ameno sob pressão da própria sociedade. É unilateral e heterônoma. É aplicada externamente. É incoercível, não pode punir o indivíduo, possui sanções difusas.
>
> Normas jurídicas: regras de conduta impostas ou reconhecidas pelo poder público, composta de preceito e sanção. É bilateral. Sofre influência externa, é afetado pela sociedade. É aplicado em seu exterior, na vida social do indivíduo. É coercível, possui sanções pré-fixadas.[91]

É curioso que utilizamos a mesma palavra, "lei", e suas variações, para nos referirmos tanto às leis da natureza ou físicas quanto às leis da sociedade ou jurídicas. Na filosofia da ciência, existem diversas teorias que discutem as Leis da Natureza, como o Pragmatismo, Empirismo Moderno, Realismo Científico, Cientificismo, Anti-realismo Científico, Realismo Estrutural, Essencialismo Disposicional, Realismo Nomológico, entre outras. Cada uma dessas teorias oferece uma maneira diferente de entender as leis da natureza. Alguns as veem como descrições verdadeiras da realidade, enquanto outros as enxergam como representações matemáticas da estrutura do mundo físico. Algumas teorias as consideram apenas descrições de regularidades observáveis, sem existência autônoma, ou ainda como ferramentas úteis para prever fenômenos. Há ainda a teoria que sustenta que as leis não são literalmente 'verdadeiras' e não precisam corresponder à realidade, entre outras perspectivas.

Os conceitos de lei natural, tanto no sentido social e jurídico quanto no científico, são derivados de 'physis'[92], a palavra grega para natureza. No entanto, o uso do conceito no âmbito social e jurídico parece ser anterior ao da ciência e teve uma forte influência religiosa em sua formação. Newton, por exemplo, foi influenciado por uma visão que sustentava que

[91] COELHO SOUZA, Daniel. *Introdução à ciência do direito*. Rio de Janeiro: Fundação Getúlio Vargas, 1972. - 2ª ed. São Paulo: Saraiva, 1983.
[92] Disponível em <https://en.wikipedia.org/wiki/Scientific_law> Acesso em outubro de 2022.

Deus havia estabelecido leis físicas universais e imutáveis. Essa ideia evoluiu posteriormente no século XVII na Europa, com o início de experimentações científicas e o desenvolvimento de formas avançadas de matemática.

> Os filósofos da ciência Carl Hempel e Paul Oppenheim chamaram atenção para a estrutura lógica deste tipo de explicação, que passou a ser chamado Modelo de Lei de Cobertura da Explicação, ou então Modelo Dedutivo-Nomológico. Um exemplo famoso de unificação foi dado pela lei da gravitação universal de Newton (1687), que explicou a queda dos corpos na Terra, o movimento dos corpos celestes e as marés de uma única maneira. [93]

A noção de leis da natureza é derivada da tradição jurídica, que valoriza as leis codificadas. Esse aspecto é curioso, pois, enquanto as leis da sociedade ou jurídicas são inerentemente mutáveis e violáveis, as leis da natureza ou físicas são vistas como invariáveis e universais. Estas últimas exibem regularidades que podem ser formuladas matematicamente. A diferença fundamental entre duas leis, pelo menos para este trabalho é o agente do seu cumprimento; a natureza e o homem. O que será desenvolvido nos próximos capítulos junto com a axiomatização (ou simbolização) para as leis da sociedade ou jurídicas.

Na Parte I apresentamos a divisão dos PIA por fases com o predomínio do termo lei na primeira fase e princípio na segunda fase. Ambos são usados para descrever regras fundamentais ou normas que regem o comportamento de sistemas naturais ou humanos e muitas vezes sem distinções. No entanto, acreditamos que eles têm conotações levemente distintas e são usados em contextos diferentes. Na ciência, uma lei é uma descrição matemática de um fenômeno que é universalmente verdadeira sob as mesmas condições. Em um contexto jurídico, uma lei é uma regra estabelecida pela autoridade governamental para regular o comportamento humano na sociedade. Historicamente, o termo "lei" era muitas vezes usado em um sentido mais amplo para incluir o que agora chamamos de teorias ou hipóteses. Hoje, o termo "lei" na ciência geralmente se refere a regras que foram rigorosamente testadas e são amplamente aceitas como verdadeiras.

Em contraste com uma "lei", um “princípio” geralmente não é formulado como uma descrição matemática precisa de um fenômeno. Em vez disso, é frequentemente uma

[93] PESSOA, Osvaldo Jr. *Notas de Aula de Teoria do Conhecimento e Filosofia da Ciência I: Um Panorama Histórico com Olhar Contemporâneo*. 2014. P. 17.

declaração qualitativa ou uma orientação geral. Por exemplo, na física, o Princípio da Incerteza de Heisenberg não é uma lei no sentido estrito; em vez disso, é uma orientação sobre as limitações fundamentais de nossa capacidade de medir simultaneamente a posição e o momento de uma partícula. De maneira similar, na ética, um princípio como o "princípio da autonomia" não é uma lei que possa ser rigorosamente aplicada em todas as situações; em vez disso, é uma orientação que sugere que devemos respeitar a capacidade dos indivíduos de tomar suas próprias decisões. Outro exemplo seria o "princípio de precaução" de Jonas.

Pode haver uma sobreposição significativa entre os termos e a distinção pode ser nebulosa em muitos casos. Adotamos o termo "Princípios da Inteligência Artificial" neste trabalho pois acreditamos que pelas suas características seria mais adequado que o termo lei, uma vez que "leis" passam a ideia de regras mais precisas e quantitativas sobre como os sistemas se comportam e "princípios" mais qualitativos e orientadores. Conforme citado no artigo de Floridi (2019), os princípios são os termos já adotados na bioética.

1.3 - As regras das leis

Em inglês, *rule of law(s)* [94] é entendido como "estado de direito", primado da lei enquanto que *rules of laws* pode referir-se às normas ou princípios que regem a criação e aplicação das leis. Uma possível tradução em português seria "regras das leis" ou "princípios das leis". O "estado de direito" referindo-se ao princípio de que todos os indivíduos, incluindo aqueles em posições de poder, estão sujeitos e são responsáveis perante a lei, enfatizando a importância de um judiciário independente, processos claros e justos para fazer cumprir as leis e a garantia dos direitos humanos para todos. Por outro lado, "regras das leis" refere às leis, regulamentos e estruturas legais específicas que regem uma determinada sociedade ou instituição. Essas regras são criadas e aplicadas por governos, organizações e outras autoridades para manter a ordem, proteger os direitos e promover a justiça.

> O estado de direito (rule of laws) é um conjunto de princípios, ou ideais, para garantir uma sociedade ordenada e justa. Muitos países em todo o mundo se esforçam para defender o estado de direito onde ninguém está acima da lei, todos são tratados igualmente sob a lei, todos são responsabilizados pelas mesmas leis,

[94] Disponível em <https://plato.stanford.edu/entries/rule-of-law/> Acesso em outubro de 2022.

> há processos claros e justos para fazer cumprir as leis, há um judiciário independente e os direitos humanos são garantidos para todos. [95]

Pirie (2021), como antropóloga especializada em direito, apresentou o surgimento das primeiras leis da civilização percorrendo Mesopotâmia, China e Índia no seu livro "The Rule of Laws". Tom Bingham, descrito como 'o mais eminente de nossos juízes' (The Guardian), apresentou uma abordagem diferente no seu livro [96] "The Rule of Law". [97] Bingham traça o uso do termo "The rule of law", identificando os primeiros registros históricos. O termo foi empregado na Inglaterra por A. V. Dicey, professor de Direito Inglês em Oxford, na obra "Uma Introdução ao Estudo da Lei da Constituição", de 1885. Segundo Tom Bingham, o livro de Dicey causou grande impressão e teve várias edições. Porém, embora Dicey tenha cunhado a expressão, ele não inventou a ideia que está por trás dela. J. W. F. Allison, autor do livro "The English Historical Constitution" [98] rastreou a ideia até Aristóteles, que em uma tradução moderna para o inglês se refere às regras da lei.

Bingham (2011) apresenta três regras que Dicey definiu como as regras da lei. A primeira é que "nenhum homem é punível ou pode ser legalmente obrigado a sofrer em seu corpo ou bens, exceto por uma violação distinta da lei estabelecida de maneira legal comum perante os tribunais comuns do país". A segunda regra indica que, "todo homem, qualquer que seja sua posição ou condição, está sujeito à lei ordinária do reino e passível de jurisdição dos tribunais ordinários." A terceira regra pode ser descrita como um atributo especial das instituições inglesas.

Bingham (2011), após esta introdução com três regras da lei de Dicey, analisa os principais marcos legais tais como Magna Carta de 1215, Habeas corpus, A abolição da tortura, A Petição de Direito de 1628, A Lei de Emenda de Habeas Corpus de 1679, A Declaração de Direitos de 1689, e O Decreto de Estabelecimento de 1701, A Constituição dos Estados Unidos

[95] Disponível em <https://www.americanbar.org/groups/public_education/resources/rule-of-law/> Acesso em outubro de 2022.
[96] Este capítulo teve contribuição de dois livros com títulos praticamente idênticos, *The Rule of Law* de Tom Bingham de 2011 e *The Rule of Laws* de Fernanda Pirie de 2021. Se nos primeiros anos da dissertação até a qualificação em 2019, o foco da bibliografia se concentrava em livros de Inteligência Artificial, a indagação sobre a natureza das leis trouxe aproximação com fundamentos das leis, mas encontrar obras que abordassem tais assuntos na perspectiva abordada pelo presente trabalho não foi fácil. Uma vez que seção de livros de direito, apesar de ser numerosa, raramente traz livros com conceitos primitivos de leis.
[97] BINGHAM, Tom. *The Rule of Law*. Penguin Books. London. 2011.
[98] BINGHAM apud Allison. 2011. p. 3.

da América, A Declaração Francesa de o Direito do Homem e do Cidadão de 1789, A Declaração Americana de Direitos, A lei da guerra, A Declaração Universal dos Direitos Humanos. Após este percurso apresenta suas oito regras das leis.

1. A lei deve ser acessível e, na medida do possível, inteligível, clara e previsível
2. A questão do direito legal e responsabilidade deve ser normalmente resolvida pela aplicação da lei e não pelo exercício do poder discricionário
3. As leis do país devem ser aplicadas igualmente a todos, exceto na medida em que diferenças objetivas justifiquem a diferenciação
4. Ministros e funcionários públicos em todos os níveis devem exercer os poderes a eles conferidos de boa-fé, de forma justa, para o propósito para o qual os poderes foram conferidos. sem exceder os limites de tais poderes e não injustificadamente
5. A lei deve oferecer proteção adequada aos direitos humanos fundamentais
6. Devem ser fornecidos meios para resolver, sem custo proibitivo ou demora excessiva, disputas civis de boa-fé que as próprias partes são incapazes de resolver
7. Os procedimentos judiciais fornecidos pelo estado devem ser justos
8. A regra da lei exige o cumprimento pelo Estado de suas obrigações tanto no direito internacional quanto no direito nacional

Bingham (2011) selecionou oito regras das leis: acessibilidade da lei, lei não discricionário, igualdade perante a lei, exercício do poder, direito humano, resolução de disputas, julgamento justo, estado de direito na ordem jurídica internacional. Podemos verificar que as duas primeiras regras de Dicey são presentes aqui, se buscarmos nas raízes, “A lei deve ser acessível e, na medida do possível, inteligível, clara e previsível”, “Os procedimentos judiciais fornecidos pelo estado devem ser justos”, também foram apresentados pelos códigos de Hamurabi e código de Ur-Nammu.

Na tradição jurídica, há outros autores. Lon Fuller em seu livro "The Morality of Law[99]" apresentou oito regras de legalidade que são "regras internas da legalidade". Elas são voltadas para a criação e execução de leis justas e eficazes.

1. Regra geral: as normas devem ser estabelecidas de forma geral e abstrata, em vez de específicas e direcionadas a indivíduos ou grupos.
2. Promulgação: as normas devem ser divulgadas publicamente, para que as pessoas possam saber o que é esperado delas e possam se adaptar ao sistema jurídico.
3. Retroatividade: as normas não devem ter efeito retroativo, ou seja, não devem ser aplicadas a ações que ocorreram antes de sua promulgação.

[99] Disponível em <https://plato.stanford.edu/entries/rule-of-law/> Acesso em outubro 2022. Tradução nossa.

4. Ausência de contradição: as normas não devem se contradizer, para que as pessoas possam prever as consequências de suas ações.
5. Possibilidade de cumprimento: as normas devem ser claras o suficiente para que as pessoas possam cumpri-las e evitar penalidades.
6. Coerência: as normas devem ser coerentes com as normas existentes do sistema jurídico em questão.
7. Congruência entre ação oficial e regra anunciada: as autoridades devem seguir e aplicar as normas de forma consistente, sem mudar de opinião ou serem arbitrárias.
8. Acesso aos tribunais: as pessoas devem ter acesso a tribunais imparciais para resolver disputas e obter justiça.

Norbert Wiener, no capítulo “Lei e Comunicação” da sua obra "Cibernética" de 1948, [100] apresenta uma perspectiva sobre as características necessárias das leis. Apesar de não se apresentar listada por características, é possível identificar as sete regras gerais. Wiener enxerga a lei como um sistema de controle e comunicação, onde a legislação deve ser capaz de gerenciar efetivamente situações críticas.

Controle ético aplicado à comunicação: a lei serve como um meio para regular e controlar a forma como nos comunicamos e interagimos na sociedade.

Foco na justiça: Wiener enfatiza a importância de a lei assegurar a justiça, evitar conflitos e, quando esses surgem, garantir que sejam decididos de maneira justa.

Clareza e previsibilidade: Segundo Wiener, para que uma lei seja justa e aplicável, ela deve ser clara e permitir que os cidadãos saibam antecipadamente quais são seus direitos e deveres. Isso ajuda a evitar confusão e conflito.

Inclusivo e universalidade: As leis devem ser aplicáveis a todos, independentemente de suas origens culturais ou religiosas. Wiener reconhece a diversidade das concepções de justiça ao longo da história e entre diferentes culturas.

Precisão e univocidade: Wiener salienta a necessidade de as leis serem formuladas de maneira clara e inequívoca, de modo que possam ser interpretadas da mesma forma por todos, não apenas por especialistas.

Antecipação das decisões judiciais: A técnica de interpretação de julgamentos anteriores deve ser tal que permita aos agentes judiciários prever, com grande probabilidade, qual será a decisão de um tribunal.

Questões da lei como problemas de comunicação e cibernética.

[100] WIENER, Norbert. *Cibernética*. Cultrix. São Paulo. 1985. p.97, 98, 102.

O Índice Anual do Estado de Direito[101] (*Rule of Law Index*) é um relatório elaborado pela organização não governamental *World Justice Project* (WJP), ou Projeto Justiça Mundial em português. Lançado pela primeira vez em 2008, o índice tem como objetivo fornecer uma avaliação quantitativa da adesão ao estado de direito em países ao redor do mundo. O índice usa mais de 500 variáveis para medir o cumprimento do estado de direito, agrupadas em oito fatores principais: Limites ao poder do governo, Ausência de corrupção, Transparência e clareza na elaboração de leis, Direitos fundamentais, Ordem e segurança, Cumprimento dos regulamentos, Justiça civil, Justiça penal.

Os dados para o índice são coletados através de pesquisas de opinião com o público geral e questionários direcionados a especialistas locais, abrangendo temas como governança, corrupção, direitos humanos, segurança e justiça. A última edição do Índice baseia-se em pesquisas com mais de 150.000 famílias e 3.600 advogados e especialistas para medir como o estado de direito é vivenciado e percebido em todo o mundo. Este índice permite uma comparação entre países e ao longo do tempo em relação ao estado de direito. Isso ajuda na identificação de problemas, no acompanhamento de progressos e na formulação de políticas. Ele fornece aos decisores políticos, aos cidadãos e às empresas uma ferramenta para entender o estado de direito em diferentes sociedades e para desenvolver estratégias para fortalecer o estado de direito onde ele é fraco.

Projeto Justiça Mundial definiu seus “Quatro Princípios Universais” para estado de direito. Esses quatro princípios universais constituem uma definição operacional do estado de direito. Eles foram desenvolvidos de acordo com padrões e normas aceitos internacionalmente e foram testados e refinados em consulta com uma ampla variedade de especialistas em todo o mundo.

1. **Responsabilidade:** O governo, assim como os atores privados, são responsáveis perante a lei.
2. **Lei justo:** A lei é clara, divulgada, estável e aplicada uniformemente. Garante os direitos humanos, bem como os direitos de propriedade, contratuais e processuais.
3. **Governo Aberto:** Os processos pelos quais a lei é adotada, administrada, julgada e aplicada são acessíveis, justos e eficientes.
4. **Justiça Acessível e Imparcial:** A justiça é entregue oportunamente por representantes competentes, éticos e independentes e neutros que são

[101] Disponível em <https://worldjusticeproject.org/about-us/overview/what-rule-law> Acesso em outubro 2022. Tradução nossa.

> acessíveis, possuem recursos adequados e refletem a composição das comunidades que atendem.

Ao examinar os conjuntos de princípios anteriores (três regras de Dicey, oito princípios de legalidade de Fuller, oito regras de lei de Bingham, Quatro Princípios Universais do Índice do Estado de Direito do WJP, sete princípios de Wiener), podemos identificar as seguintes regras comuns.

> **Clareza e previsibilidade**: A lei deve ser clara, previsível e compreensível para o cidadão médio. Deve permitir que as pessoas saibam quais são seus direitos e obrigações.
>
> **Justiça**: A lei deve garantir a justiça, tanto em termos de proteger os direitos humanos, como em termos de resolver disputas de maneira justa e eficaz.
>
> **Universalidade e igualdade**: A lei deve ser aplicável a todos igualmente, independentemente do seu status, origem cultural ou religiosa. As diferenças só podem ser justificadas com base em diferenças objetivas que justifiquem a diferenciação.
>
> **Acessibilidade**: O acesso à justiça e aos tribunais deve ser garantido a todos, independentemente de sua situação econômica ou social.
>
> **Estabilidade e coerência**: A lei deve ser estável, sem alterações retroativas, e deve ser coerente com as outras leis existentes no sistema jurídico.
>
> **Responsabilidade (governo e funcionários públicos)**: Os governantes e funcionários públicos devem ser responsabilizados pela lei e devem exercer seus poderes de maneira justa e dentro dos limites desses poderes.
>
> **Publicidade**: A lei deve ser divulgada publicamente para que as pessoas saibam o que é esperado delas e possam se adaptar ao sistema jurídico.
>
> **Aplicação (uniforme)**: A lei deve ser aplicada de maneira uniforme e justa em todas as circunstâncias, sem qualquer exercício discricionário injustificado do poder.

Estas regras comuns modernas também podem ser identificadas, de forma rudimentar, em códigos legais mais antigos da história citados previamente. Os códigos de leis, escritos em tábuas de argila e publicamente exibidos, representam a primeira tentativa de tornar a lei acessível e conhecida pela população. As regras e as respectivas penalidades eram claramente estabelecidas, proporcionando uma certa previsibilidade para os povos da Mesopotâmia desde o Código de Ur-Nammu. O Código de Hamurabi, talvez o mais famoso dos códigos antigos, proclamou que sua função era "trazer justiça à terra... e destruir o mal e o malfeitor". Ainda que os padrões de justiça naquela época fossem diferentes dos atuais, o conceito central de buscar justiça e ordem estava presente.

A codificação das leis contribuiu para a estabilidade e a coerência do sistema legal, pois estabelecia um conjunto fixo de regras a serem seguidas, ajudando a limitar a arbitrariedade e a mudança constante das regras. Ainda que os códigos antigos distinguissem entre diferentes classes sociais, havia uma tentativa de criar um sistema de regras que se aplicasse de maneira geral à população com leis especificas também para escravos, representando um princípio rudimentar de universalidade. O Código de Hamurabi, muitas vezes esculpidos em estelas de pedra, era publicamente exibido para que todos os cidadãos pudessem vê-lo, simbolizando um primeiro passo em direção ao princípio da publicidade das leis.

Será que os PIA também seguem estas regras que até podemos identificar nos primeiros códigos? No capítulo 3 da Parte I, no artigo publicado pelos cientistas da Federação de Cientistas da Alemanha (VDW) sobre os 23 Princípios de Asilomar, os critérios como Clareza, Previsibilidade, Responsabilidade, Justiça, Aplicação, Acessibilidade, Estabilidade, Coerência são todos criticados como insuficientes ou duvidosos, talvez somente o requisito Publicidade escaparia deste escrutínio.

Seriam estas as regras mais básicas das leis? Para Pirie (2021), por trás do sucesso das leis introduzidas na Mesopotâmia para o mundo inteiro em mais de quatro milênios, havia as ideias sobre a lei natural e a humanidade comum. A ideia de lei natural apontado como um dos fatores do sucesso, foi discutido durante a tradição jurídica entre diversos autores.

Hobbes reconheceu uma forma de "lei natural" em sua filosofia, mas rejeitou a noção de que esta poderia servir como uma base para os direitos humanos no sentido moderno. Para Hobbes, a lei natural era um estado de guerra de todos contra todos, que precisava ser controlado por um soberano poderoso. Kant, na Metafísica dos Costumes, argumenta que a moralidade é baseada em princípios racionais universais e que esses princípios são inerentes à natureza humana. Defende a ideia de que Deus, ao criar os seres humanos como seres racionais, também inscreveu em seus corações a lei moral básica de não fazer aos outros aquilo que consideramos injusto para nós mesmos.

> Também se pode perguntar como é possível saber que coisas foram ordenadas por Deus. A tal pergunta se pode responder: o próprio Deus, pois ele fez os homens racionais, prescreveu e inscreveu em todos os corações a seguinte lei: ninguém fará a seu semelhante aquilo que considere injusto que um outro lhe faça.[102]

Bentham é famoso por ter descrito a ideia de direitos naturais como "tolices sobre pernas de pau".[103] Como utilitarista, acreditava que as leis deveriam ser feitas para promover a maior felicidade para o maior número de pessoas, e não baseadas em algum conceito abstrato de direito natural. John Austin, positivista, argumentava que as leis são regras feitas por seres humanos (seja um soberano, um governante ou uma instituição legislativa) e não derivam de qualquer direito natural ou divino. Hart, um dos mais influentes filósofos legais do século XX, criticou a teoria do direito natural em sua obra "O Conceito de Direito". Hart argumentou que a lei é um sistema de regras sociais e que não há necessidade de vinculá-la a uma moralidade superior ou a uma "lei natural".

Apesar da discordância por parte de variados pensadores, a ideia da lei natural pode ter sido fundamental para o sucesso das leis durante milhares de anos em diversas localidades. Junto com a ideia sobre lei natural, Pirie (2021) também apontou a ideia de humanidade comum como outro fator de sucesso. De fato, o reconhecimento de humanidade comum é o fundamento mais básico para sistemas legais. John Austin, considerado como um dos precursores do positivismo jurídico, que discorda da ideia de lei natural, definiu a lei da seguinte maneira.

---

[102] MORRISON, Wayne. Filosofia do Direito – Dos gregos ao pós-modernismo. Martins Fontes. São Paulo. 2012. p. 103.

[103] "O texto de Jeremy Bentham […] apresenta uma das mais famosas e influentes críticas dirigidas ao direito natural e, por extensão, à noção de direitos humanos. […] Bentham destaca a fragilidade dessa perspectiva e seu argumento central reside em dois pontos: a tendência desse tipo de "legislação" produzir anarquia e a ausência de uma base ontológica, de modo que tais "direitos" expressariam nada mais do que desejos e paixões sem qualquer sustentação. Bentham demostra uma preocupação particular em relação à linguagem utilizada na Declaração francesa, vista como ambígua e imprecisa, uma retórica que carece de sentido. As críticas formuladas nesse texto situam-se no prolongamento das ideias expostas por Bentham ainda em 1776 na obra A Fragment on Government, quando foi traçada a distinção fundamental do positivismo jurídico entre o direito como ele é (law as it is) e o direito como deve ser (law as it ought to be)." Disponível em <https://loja.editoradialetica.com/humanidades/tolices-sobre-pernas-de-pau-um-comentario-a-declaracao-de-direitos-do-homem-e-do-cidadao-de-1789> Acesso em maio de 2023.

> Pode -se afirmar que uma lei, na acepção mais geral e abrangente em que o termo é empregado em seu sentido literal, é uma regra estabelecida para a orientação de um ser inteligente por um ser inteligente que tem poder sobre ele. [104]

Se analisarmos todas as regras acima podemos inferir que a regra mais fundamental da lei seria humanidade comum - apesar da distinção sobre a natureza da sua origem -, uma orientação de um ser inteligente por um ser inteligente. No conceito do jurídico moderno, há também a necessidade do "estado de direito" para garantia das regras das leis.

1.4 - Simbolização das leis

No capítulo 2 da Parte I, verificamos quatro tipos de estudos sobre os princípios da Inteligência Artificial, cada abordagem com sua própria metodologia tais como agrupamentos e identificações por palavras-chaves e iniciativas por trás, entre outras. Neste trabalho, procuramos oferecer uma alternativa metodológica distinta daquelas. A investigação sobre a natureza dos Princípios da Inteligência Artificial levou-nos à procura de uma essência, elementos comuns entre os diversos princípios. Esta tentativa ao encontro com a simbolização como método, [105] possibilitou-nos uma visão mais nítida, e a partir deste entendimento, apreendemos que todas as regras das leis compartilham da mesma natureza, a humanidade comum, que poderiam ser simbolizadas.

O sistema de simbolização adotado para esta tarefa consiste em fórmula geral das leis [(1) (a) (2)], onde (1) e (2) representam posições dos agentes envolvidos na ação da lei representada pela por (a).

[(1) (a) (2)]

fórmula básica das leis da sociedade

Ocupando estas posições por letras, [H a H] seria a simbolização geral de lei, onde [H] representa seres humanos. As leis da sociedade operam em [H a H], uma forma de

[104] MORRISON, Wayne. *Filosofia do Direito – Dos gregos ao pós-modernismo*. Martins Fontes. São Paulo. 2012. p. 273.

[105] A ideia do uso de axiomas neste trabalho foi inspirada na obra "A Existência de Deus" de Richard Swinburne. O evento de lançamento da versão em português foi realizado na FFLCH/USP com a presença do autor em setembro de 2015.

representação pela redução. Poderia considerar como axioma das leis da sociedade que são normativas e violáveis.

[H a H]

axioma das leis da sociedade

Que ainda podem ser representadas distinguindo o primeiro e segundo agente [H].

[H^1 a H^2]

axioma das leis da sociedade com distinção entre agentes

Apesar da semelhança esta simbolização não segue exatamente as regras para simbolização de sentenças utilizadas na lógica proposicional. Poderia considerar que neste trabalho a simbolização tem um uso livre. Esta metodologia permitiu uma inferência sobre a natureza dos agentes, seus papéis, suas relações, além da definição da proposição que justificaria a validade dos princípios da Inteligência Artificial. Além de funcionar como método, essa simbolização desempenha também o papel do fio condutor desta exposição que será exposto em uma ordem crescente de complexidade desta simbolização.

A simbolização aliada ao método de redução na descoberta de fundamentos é um recurso que faz parte da história da ciência e também da filosofia: Leibniz procurou chegar em ideias complexas a partir de ideias simples usando regras lógicas na elaboração de conceitos de Característica Universal e Gramática Racional. O uso de símbolos para processo de raciocínio dedutivo proposto por Leibniz, e desenvolvido posteriormente por George Boole, Gottob Frege e Augustus De Morgan formou base da lógica simbólica moderna. Patrick Suppes[106], conhecido por seu trabalho na teoria de decisão, axiomatizou as teorias empíricas o que lhe possibilitou a compreensão da estrutura interna da teoria e suas eventuais relações com outras. De modo semelhante, acreditamos que a simbolização dos Princípios da

[106] Patrick Suppes teve a ideia de axiomatizar as teorias empíricas de uma maneira muito mais simples e "transparente" conceitualmente que as tentativas precedentes, o que não somente lhe permitiu compreender mais facilmente a estrutura interna "essencial" da teoria assim reconstruída, mas também examinar adequadamente suas eventuais relações com outras." MOULINES, Carlos, Ulisses. *O desenvolvimento moderno da filosofia da ciência (1890 – 2000)*. Associação Filosófica Scientia Studia. São Paulo. 2020. p. 169.

Inteligência Artificial ofereceu-nos uma compreensão melhor da sua natureza e possibilitou, a partir dela, formulações de questões derivadas, ampliando seu horizonte.

> A simbolização é um processo que associa símbolos a palavras, mas os símbolos podem desempenhar o papel de letras esquemáticas, como na formalização de Hilbert, ou seja, como termos com um sentido meramente formal que permite uma variedade de interpretações, ou ter um papel substantivo, como termos cujos significados devem ser transmitidos por elucidação. A contribuição de Bertran-San Millán explica como Frege utilizou os símbolos da aritmética como nomes canônicos, ou seja, como símbolos com um significado específico e fixo, de modo que as letras matemáticas sempre tenham um domínio específico, determinado pela aplicação pretendida. Peano compartilhou uma compreensão substantiva semelhante de símbolos matemáticos em seus primeiros escritos, mas mudou-se para uma visão de símbolos indefinidos como constantes não lógicas não interpretadas desprovidas de significado, quando investigou questões metateóricas sobre a independência dos axiomas com Padoa. [107]

Na tradição da filosofia não é difícil encontrar obras que começam introduzindo definições básicas para elaborar conceitos mais complexos, mesmo para desenvolver temas como leis, normas sociais e ética. No capítulo *Bem-estar Social e Ética Pessoal* do livro *Sistema da Ética*[108], Wilhelm Dilthey apresenta quatro leis elementares que fazem parte do primeiro axioma.

> Primeiro axioma - Os atos volitivos humanos produzem símbolos, fórmulas gerais, que se originam a partir de leis elementares. As mais importantes leis elementares são:
>
> 1) O transcurso desde instintos dispersos até uma estruturação destes.
> 2) A relação básica de motivação contida em cada movimento instintivo, conteúdo instintivo e meios, os movimentos.
> 3) A lei do costume.
> 4) A coerência básica de ação e reação nas relações de distintas unidades vitais entre si.
>
> Destas coerências básicas surgem a formas dos processos volitivos na humanidade. Estas são: o desenvolvimento de coerência motivadas, de bens, o surgimento de costumes, uso, lei, conformação de regras, máximas, prescrições éticas e juízos éticos.

Antes de Dilthey, Spinoza, no século XVII, elaborou a sua obra Ética[109] seguindo uma ordem geométrica, na definição do próprio autor, onde Deus, A natureza e a origem da mente, A origem e a natureza dos afetos, A servidão humana ou a força dos afetos, A potência do

[107] Disponível em <https://journals.openedition.org/philosophiascientiae/2788> Acesso em outubro de 2022. Tradução nossa.
[108] DILTHEY, Wilhelm. *Sistema da Ética, Coleção Fundamentos de Direito*. São Paulo. Ícone Editora. 2005. p. 179
[109] SPINOZA. *Ética*. 2ª ed. Belo Horizonte. Autêntica Editora. 2013.

intelecto ou a liberdade humana são apresentadas através de definições, axiomas, proposições (com corolários e escólios) e demonstrações.

> Definições (1~8)
>
> 1. Por causa de si compreendo aquilo cuja essência envolve a existência, ou seja, aquilo cuja natureza não pode ser concebida senão como existente.
> 3. Por substância compreendo aquilo que existe em si mesmo e que por si mesmo é concebido, isto é, aquilo cujo conceito não exige o conceito de outra coisa do qual deva ser formado.
> 5. Por modo compreendo as afecções de uma substância, ou seja, aquilo que existe em outra coisa, por meio da qual é também é concebido.
>
> Axiomas (1~7)
>
> 1. Tudo o que existe, existe ou em si mesmo ou em outra coisa
>
> Proposições (1~36)
>
> Proposição 1. Uma substância é, por natureza, primeira, relativamente às suas afecções.
>
> Demonstração. É evidente, pelas def. 3 e 5 (Ética Spinoza)

Há obras filosóficas que lembram até a lógica da programação de software como Tratado Lógico-Filosófico [110] de Wittgenstein, que apresentam sentenças declarativas numeradas desde 1* até 7 com subníveis, dispostas seguindo uma ordem hierárquica onde a definição anterior permite a seguinte em escala de complexidade.

> 1* O mundo é tudo o que é o caso
>
> 1.1 O mundo é a totalidade dos fatos, não das coisas.
>
> 1.11 O mundo é determinado pelos fatos, e por estes serem todos os fatos.
>
> 1.12 Pois a totalidade dos fatos determina tanto o que é o caso quanto tudo o que não é o caso.
>
> 1.13 Os fatos no espaço lógico são o mundo.
>
> 1.2 O mundo se divide em fatos.
>
> 2 O que é o caso, o fato, é a existência de fatos atômicos. (Tratado Lógico-Filosófico Wittgenstein)

Einstein escreveu em seu ensaio “As leis da ciência e as leis da ética”[111] que existem duas leis fundamentais que governam a humanidade: as leis da ciência e as leis da ética. As

[110] WITTGENSTEIN, Ludwig. *Major Works.* New York. HarperCollins. 2009. p. 5. Tradução nossa.
[111] EINSTEIN, Albert. *Out of My Later Years.* Philosophical Library. New York. 1950. p. 114-115. Tradução nossa.

leis da ciência se referem às leis físicas e naturais que governam o universo. As leis da ética se referem às normas morais e valores que guiam as ações humanas. Essas leis são baseadas em princípios como a justiça, a equidade e o respeito pelos direitos humanos. Elas não são determinadas por observações empíricas, mas por considerações filosóficas e culturais. Apesar de não serem derivadas por observações, Einstein argumenta que de modo similar da ciência, poderia construir as leis da ética a partir de axiomas.

> No entanto, as diretrizes éticas podem se tornar racionais e coerentes pelo pensamento lógico e pelo conhecimento empírico. Se pudermos concordar com algumas proposições éticas fundamentais, então outras proposições teóricas podem ser derivadas delas, desde que as premissas originais sejam declaradas com precisão suficiente. Tais premissas éticas desempenham um papel semelhante na ética, ao desempenhado pelos axiomas na matemática.[112]

Einstein (1950) esclarece que, embora as leis da ciência e as leis da ética possam parecer distintas, elas estão, na verdade, interconectadas. Ainda reforça que as leis da ética são tão importantes quanto as leis da ciência, pois elas nos guiam em nossas ações e nos ajudam a construir uma sociedade justa e sustentável. Ele conclui seu ensaio afirmando que "o nosso futuro depende tanto da sabedoria da ética quanto da sabedoria da ciência".

> É privilégio do gênio moral do homem, personificado por indivíduos inspirados, promover axiomas éticos que são tão abrangentes e tão bem fundamentados que os homens os aceitarão como fundamentados na vasta massa de suas experiências emocionais individuais. **Os axiomas éticos são encontrados e testados não muito diferentemente dos axiomas da ciência**. A verdade é o que resiste ao teste da experiência.[113]

Duhem, no seu “Ensaios de filosofia da ciência”[114] afirma que as leis da física são relações simbólicas. Argumentamos que todas as leis da sociedade que também apresentam relações simbólicas, e assim, podem ser representadas por [H^1 a H^2]. Inevitavelmente, todas as leis da humanidade apresentam estes elementos elementares independente da origem da lei, objetivo e cultura. Uma lei é sempre regida exigindo o seu cumprimento pelos seres humanos representado por H^1. Esta exigência que pode ser tanto uma proibição ou realização de uma ação representado por “a”. O que pode variar na história da nomologia seria o papel

[112] Ibid
[113] Ibid destaque nossa.
[114] DUHEM, Pierre. *Ensaios de filosofia da ciência*. Associação Filosófica Scientia Studia. São Paulo. 2019. p. 198.

do H^2, que nem sempre será ocupado por seres humanos podendo ser ocupados por mortos, deuses, animais, estado, natureza, bens, etc.

1.5 - Fórmula básica das leis - [H a H]

A partir da simbolização introduzida no capítulo anterior, podemos representar as leis sociais e jurídicas pela formulação básica (formulação elementar das leis ou axioma básico).

[(1) a (2)][115]

fórmula básica das leis da sociedade

[H a H][116]

axioma das leis da sociedade

Podemos derivar da mesma ideia de simbolização para representar outros ocupantes para (2) como divindades, natureza, animais, etc. Até hoje apenas o ser humano ocupou lugar 1 (tanto que corporações precisaram tomam emprestado a figura do ser humano e assumem-se como "pessoa" para seu funcionamento). O lugar 2 foi ocupado por homens de diferentes posições (escravos e castas) e situações (vivos e mortos como antepassados), deuses, natureza, animais, etc.

H	a	H
		N
		A

axioma das leis da sociedade com natureza variável em (2)

[115] Adotamos colchetes para expressar simbolização adotada, tanto para axiomas [H a H] como para elementos quando citados isoladamente como [H] e [a]

[116] Adotamos letra "H" (maiúscula) para representar seres humanos [H], letra "a" para ação da lei em minúscula [a], "R" para robô e IA [R]. Utilizamos numerais como 1 e 2 para diferenciar agentes da ação e do objeto sobrepostos como usados em potência. Se usarmos o mesmo número [H^1 a H^1] podemos indicar ações autorreferenciais como danos a si mesmo, uso de drogas ilícitas, suicídio, eutanásia, etc.

É possível também representar as origens das leis: leis naturais (por N), divinas (por D) e positivas (por H).

N[H a H], D[H a N], H[H a A]
ou como fração
[H a H]/N, [H a N]/D, [H a A]/H

axioma das leis da sociedade pela origem das leis

As primeiras leis que temos conhecimento na história como o Código de Ur-Nammu – podem ser simbolizadas por [H^1 a H^2]. Vejamos as três primeiras leis que aparecem no prólogo:

O órfão não será entregue ao homem rico

H^1 (ser humano) - a (não entregar ao homem rico) - H^2 (órfão)

A viúva não será entregue ao poderoso

H^1 (ser humano) - a (não entregar ao poderoso) - H^2 (viúva)

O homem de um siclo não será entregue ao homem de uma mina (60 siclos)

H^1 (ser humano) - a (não entregar ao homem de uma mina) - H^2 (homem de um siclo)

As primeiras duas leis em forma casuística:

1. Se um homem comete um assassinato, esse homem deve ser morto.

H^1 (ser humano) - a (se cometer assassinato de H^2, H^1 deve ser morto) - H^2 (ser humano)

2. Se um homem cometer um roubo, ele será morto.

H^1 (ser humano) - a (se roubar H^2, H^1 será morto) - H^2 (qualquer ser humano)

Da mesma forma que demonstramos que as leis do Código de Ur-Nammu – podem ser simbolizadas por [H^1 a H^2], o Código de Hamurabi podem ser simbolizados por H^1 a H^2. Vejamos algumas leis que aparecem no prólogo:

Código de Hamurabi[117]

> I - SORTILÉGIOS, JUÍZO DE DEUS, FALSO TESTEMUNHO, PREVARICAÇÃO DE JUÍZES
>
> 1º - Se alguém acusa um outro, lhe imputa um sortilégio, mas não pode dar a prova disso, aquele que acusou, deverá ser morto.
>
> II - CRIMES DE FURTO E DE ROUBO, REIVINDICAÇÃO DE MÓVEIS
>
> 6º - Se alguém furta bens do Deus ou da Corte deverá ser morto; e mais quem recebeu dele a coisa furtada também deverá ser morto.
>
> 22º - Se alguém comete roubo e é preso, ele é morto.
>
> IV - LOCAÇÕES E REGIMEN GERAL DOS FUNDOS RÚSTICOS, MÚTUO, LOCAÇÃO DE CASAS, DAÇÃO EM PAGAMENTO
>
> 59º - Se alguém, sem ciência do proprietário do horto, corta lenha no horto alheio, deverá pagar uma meia mina.
>
> X - MATRIMÔNIO E FAMÍLIA, DELITOS CONTRA A ORDEM DA FAMÍLIA. CONTRIBUIÇÕES E DOAÇÕES NUPCIAIS
>
> SUCESSÃO
>
> 128º - Se alguém toma uma mulher, mas não conclui um contrato com ela, esta mulher não é esposa.

[117] Disponível em <http://www.dhnet.org.br/direitos/anthist/hamurabi.htm> Acesso em outubro de 2022. Tradução nossa.

XI - ADOÇÃO, OFENSAS AOS PAIS, SUBSTITUIÇÃO DE CRIANÇA

195º - Se um filho espanca seu pai se lhe deverão decepar as mãos.

XII - DELITOS E PENAS (LESÕES CORPORAIS, TALIÃO, INDENIZAÇÃO E COMPOSIÇÃO)

196º - Se alguém arranca o olho a um outro, se lhe deverá arrancar o olho.

197º - Se ele quebra o osso a um outro, se lhe deverá quebrar o osso.

198º - Se ele arranca o olho de um liberto, deverá pagar uma mina.

199º - Se ele arranca um olho de um escravo alheio, ou quebra um osso ao escravo alheio, deverá pagar a metade de seu preço.

200º - Se alguém parte os dentes de um outro, de igual condição, deverá ter partidos os seus dentes.

201º - Se ele partiu os dentes de um liberto deverá pagar um terço de mina.

205º - Se o escravo de um homem livre espanca um homem livre, se lhe deverá cortar a orelha.

Tentaremos exemplificar um caso distinto que aparece no 6º código.

6º - Se alguém furta bens do Deus ou da Corte deverá ser morto; e mais quem recebeu dele a coisa furtada também deverá ser morto.

H (ser humano) - a (furta bens do Deus ou da Corte, H deve ser morto) - D (Deus ou da Corte)

No trecho “XII - DELITOS E PENAS (LESÕES CORPORAIS, TALIÃO, INDENIZAÇÃO E COMPOSIÇÃO)” acima identificamos a "lei de talião" que se refere à ideia de que uma pessoa que prejudicou outra deve ser punida de maneira semelhante - ou seja, "olho por olho, dente por dente". Mas notamos também que existem muitas leis no código que prescrevem multas e outras penalidades que não são uma retribuição exata do dano causado. Além disso, as penalidades muitas vezes variavam dependendo do status social das partes envolvidas, o que vai contra os princípios modernos de igualdade perante a lei. Entretanto, o Código de Hamurabi e sua lei de talião se tornaram famosos porque representam uma das primeiras tentativas documentadas de estabelecer um código de leis escrito e universalmente aplicável

dentro de uma civilização. A adoção da lei de talião representa uma tentativa de impor um senso de justiça proporcional nas punições, em vez de permitir a vingança descontrolada.

Nas religiões abraâmicas, os Dez Mandamentos[118] (em hebraico: עֲשֶׂרֶת הַדְּבָרוֹת, *Aseret ha'Dibrot*), também conhecidos como Decálogo, são um conjunto de princípios relacionados à ética e à adoração. Os mandamentos aparecem duas vezes na Bíblia hebraica (no Êxodo e no Deuteronômio), e incluem instruções como a de não adorar outros deuses, honrar os pais e guardar o Sabá, bem como proíbe idolatria, blasfêmia, assassinato, adultério, roubo, desonestidade e cobiça.

Os Dez Mandamentos[119]

> 20 E Deus falou todas estas palavras:
>
> 2 "Eu sou o SENHOR, o teu Deus, que te tirou do Egito, da terra da escravidão.
>
> 3 "Não terás outros deuses além de mim.
>
> 4 "Não farás para ti nenhum ídolo, nenhuma imagem de qualquer coisa no céu, na terra, ou nas águas debaixo da terra. 5 Não te prostrarás diante deles nem lhes prestarás culto, porque eu, o SENHOR, o teu Deus, sou Deus zeloso, que castigo os filhos pelos pecados de seus pais até a terceira e quarta geração daqueles que me desprezam, 6 mas trato com bondade até mil gerações[a] aos que me amam e obedecem aos meus mandamentos.
>
> 7 "Não tomarás em vão o nome do SENHOR, o teu Deus, pois o SENHOR não deixará impune quem tomar o seu nome em vão.
>
> 8 "Lembra-te do dia de sábado, para santificá-lo. 9 Trabalharás seis dias e neles farás todos os teus trabalhos, 10 mas o sétimo dia é o sábado dedicado ao SENHOR, o teu Deus. Nesse dia não farás trabalho algum, nem tu, nem teus filhos ou filhas, nem teus servos ou servas, nem teus animais, nem os estrangeiros que morarem em tuas cidades. 11 Pois em seis dias o SENHOR fez os céus e a terra, o mar e tudo o que neles existe, mas no sétimo dia descansou. Portanto, o SENHOR abençoou o sétimo dia e o santificou.
>
> 12 "Honra teu pai e tua mãe, a fim de que tenhas vida longa na terra que o SENHOR, o teu Deus, te dá.
>
> 13 "Não matarás.

[118] Disponível em <https://pt.wikipedia.org/wiki/Dez_Mandamentos> Acesso em outubro de 2022. Tradução nossa.
[119] Disponível em <https://www.biblegateway.com/passage/?search=%C3%8Axodo%2020&version=NVI-PT> Acesso em outubro de 2022. Tradução nossa.

[14] "Não adulterarás.

[15] "Não furtarás.

[16] "Não darás falso testemunho contra o teu próximo.

[17] "Não cobiçarás a casa do teu próximo. Não cobiçarás a mulher do teu próximo, nem seus servos ou servas, nem seu boi ou jumento, nem coisa alguma que lhe pertença".

Os quatro mandamentos iniciais são leis que precisam ser cumpridas em relação a Deus [H a D] simbolização que foi já vista no código de Hamurabi. Os seis mandamentos restantes que ditam regras com os semelhantes podem ser representados novamente por [H a H].

Em 1942, ocorre a publicação das três leis de Asimov; em 1947, o Código de Lipit-Ishtar é decifrado; em 1948, o Código de Ur-Nammu é descoberto; em 1950, Turing publica o artigo "Computing Machinery and Intelligence", [120] em 1952, o Código de Ur-Nammu é decifrado; em 1956 ocorre o workshop "Dartmouth Summer Research Project on Artificial Intelligence" [121], considerado como o evento precursor da Inteligência Artificial. Muitos leitores provavelmente não enxergariam alguma ligação entre as descobertas das primeiras leis da humanidade e os acontecimentos germinais na Inteligência Artificial e leis da robótica. Mas além da coincidência na proximidade dos acontecimentos com cerca de 10 anos, guardam um paralelo histórico que veremos a seguir.

[120] Disponível em <https://turingarchive.kings.cam.ac.uk/computing-machinery-and-intelligence> Acesso em outubro 2022.
[121] Disponível em <https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth> Acesso em outubro 2022.

PARTE II CAPÍTULO 2 – O INEDITISMO DE [R a H]

2.1 - Ineditismo de [R a H]

Podemos afirmar neste início do capítulo que todo o caminho percorrido até aqui, de certa medida, teve o objetivo de evidenciar o novo axioma das leis da sociedade revelado por PIA: [R a H]. Pela primeira vez na história da humanidade, na fórmula básica das leis da sociedade [(1) a (2)], o lugar (1) é ocupado por não humano, neste caso por [R], que representa robô e IA. Como apresentado anteriormente, no conto "Círculo Vicioso" [122]de 1942, Asimov apresenta uma sequência de leis para regulamentar o comportamento de robôs no convívio com os seres humanos.

No prefácio da coletânea de história de robôs "*Machines that think*" [123] de 1983, Asimov esclareceu o motivo que o levou a elaborar as "Três Leis da Robótica", como se fosse um prólogo nos primeiros códigos de leis da Mesopotâmia. A intenção foi refutar o "Complexo de Frankenstein" disseminado pelo livro de Mary Shelley, considerada como a primeira obra de ficção científica, publicada em 1818.

> O êxito de *Frankenstein* foi tão grande que a ideia básica – "o homem cria o robô; o robô mata o homem" – se repetiu sem parar numa série interminável de história de ficção científica. Virou um dos mais insuportáveis chavões do gênero – e que combati e destruí, com sucesso, tenho orgulho de dizer, ao enunciar as minhas "Três leis da robótica".

Apesar do seu surgimento numa obra de ficção, essas três leis inauguraram e influenciaram surgimento de outras leis para robôs e Inteligência Artificial até hoje. Marvin Minsky que organizou a conferência de Dartmouth – considerada como o marco inaugural para a Inteligência Artificial – em uma entrevista revelou a influência de Asimov: "Depois que

[122] Foi escrito em outubro de 1941 e publicado pela revista *Astounding Science Fiction*, edição de março de 1942.
[123] ASIMOV, Isaac. (1983) *Machine that think: The Best Science Fiction Stories About Robot and Computer. História de Robô*, Porto Alegre.L&PM. 2010.

Círculo Vicioso apareceu na edição de março de 1942 da *Astounding Science Fiction*, eu nunca parei de pensar em como as mentes podem funcionar." [124]

Se compararmos o teor contido nas leis e os objetos da ação, apesar da distância de milhares de anos, o código de Ur-Nammu – a primeira lei da civilização que temos conhecimento, e as "Três Leis da Robótica" - a primeira lei da robótica, apresentam similaridades. A fundamental diferença entre eles está no agente a quem essas leis são destinadas.

> 1. Se um homem comete um assassinato, esse homem deve ser morto.
> 2. Se um homem cometer um roubo, ele será morto.
> 3. Se um homem cometer um sequestro, ele deve ser preso e pagar 15 siclos de prata.
> 4. Se um escravo se casar com uma escrava e essa escrava for libertada, ele não deixará a casa.
> 5. Se um escravo se casar com uma pessoa nativa (ou seja, livre), ele deve entregar o filho primogênito ao seu dono.
> 25. Se a escrava de um homem, comparando-se com sua amante, fala insolentemente com ela, sua boca deve ser limpa com 1 litro de sal.[125]

Segundo Jack M. Balkin (2017) - professor de direito constitucional da universidade de Yale - ao criar as três leis, Asimov moveu a imaginação sobre robôs de ameaças a objetos de interpretação e regulação inaugurando uma tradição que continua até hoje.

> [...] talvez o mais importante, Asimov chamou suas leis de leis de robótica, não as leis dos usuários de robôs ou programadores de robôs ou operadores de robôs. Suas leis eram dirigidas para robôs. Elas eram centradas em robôs, isto é, eram instruções de programação inseridas no código dos próprios robôs. Elas eram leis que os robôs tinham que seguir - porque eles foram programados desta maneira - e não que os usuários de robôs tivessem que seguir. [126]

Balkin (2017), observa como mais importante o fato de as leis de Asimov serem dirigidas para robôs e não para usuários, programadores ou operadores de robôs. Além disso, a aplicação dessas leis não se restringia apenas a robôs, mas também para agentes de Inteligência Artificial e algoritmos de aprendizado de máquinas. Alega-se que apesar de

[124] Disponível em <https://www.nytimes.com/1992/04/12/business/technology-a-celebration-of-isaac-asimov.html?pagewanted=all&src=pm> Acesso em outubro de 2022. Tradução nossa.
[125] Disponível em <http://realhistoryww.com/world_history/ancient/Misc/Sumer/ur_nammu_law.htm> Acesso em novembro de 2022. Tradução nossa.
[126] Disponível em http://digitalcommons.law.yale.edu/fss_papers/5159. Acesso em outubro de 2022. Tradução nossa.

Asimov ter escrito principalmente sobre robôs, havia claramente preocupação com computadores muito inteligentes. Assim, incluir a Inteligência Artificial seria perfeitamente consistente com as preocupações de Asimov nas suas três leis.

> Asimov não diz muito sobre as leis humanas que exigiam essa programação, mas assume que havia algum tipo de exigência do governo para que eles fossem colocados no cérebro positrônico de cada robô. [...] De qualquer forma, meu objetivo é perguntar como podemos usar a ideia de Asimov sobre as leis da robótica hoje. Quando falo de robôs, no entanto, incluirei não apenas robôs - objetos materiais incorporados que interagem com seu ambiente – mas também agentes de Inteligência Artificial e algoritmos de aprendizado de máquina. Isso é perfeitamente consistente com as preocupações de Asimov, eu acho. Embora Asimov tenha escrito principalmente sobre robôs, ele também escreveu sobre computadores muito inteligentes. E a síndrome de Frankenstein que ele estava tentando combater poderia surgir do medo de Al ou algoritmos tanto quanto o medo de robôs incorporados. Hoje, as pessoas parecem temer não apenas robôs, mas também agentes e algoritmos de IA incluindo aprendizado de máquina sistemas.' Os robôs parecem ser apenas um caso especial de um conjunto muito maior de preocupações. (BALKIN. 2017. p. 219)

Mas Balkin (2017), como outros autores, incluindo o próprio Asimov que afirmou que a intenção foi refutar o "Complexo de Frankenstein", parece não ter toda noção sobre a dimensão histórica deste acontecimento, mesmo enfatizando a importância da agência de robôs e Inteligência Artificial nas três leis. Se o Código de Ur-Nammu pode ser considerado como o pioneiro na tradição de leis na história humana até hoje, temos precisamente as Três Leis da Robótica de 1942, como o início de um novo tipo de lei. Pela primeira vez na história da humanidade, na simbolização [(1) a (2)] - fórmula básica das leis da sociedade, o papel de (1) é ocupado por não humano, e sim por robô e Inteligência Artificial.

Se todas as leis da sociedade podem ser representadas por [H^1 a H^2] - axioma das leis da sociedade, as leis da robótica e Inteligência Artificial podem ser representadas por [R a H], onde [R] representada robô e Inteligência Artificial. [R a H] é a ruptura da tradição [H a H]. A ocupação do lugar R no lugar de H^1 é inédita, e, representa um marco na história humana e não humana. Se a criação de novas leis pode significar surgimento de mudanças numa sociedade, a alteração na própria estrutura fundamental da lei, ou seja, na natureza do agente, certamente anuncia uma ruptura inédita que está em curso.

Nas primeiras leis da civilização, o lugar (1) já foi ocupado por escravos desumanizados e considerados até como mercadorias, mas mesmo assim, assumiam papel como agentes para cumprimento de leis como podemos verificar no código de Hamurabi. A lei casuística apresenta escravo com agência para o cumprimento da lei com sanção especifica direcionada a ele e não para o seu dono. O que demonstra que a sua humanidade era inegável por mais que os seus direitos eram negados.

> 205º - Se o escravo de um homem livre espanca um homem livre, se lhe deverá cortar a orelha.[127]

Até hoje, em apenas um único caso, o lugar (1) foi ocupado por não humano, apenas as corporações chegaram a ocupar esta posição. As corporações ou pessoas jurídicas ocupam o papel de [H] na sociedade. Durante a era industrial do século XIX o conceito de "personalidade corporativa" se desenvolveu e se espalhou, especialmente nos Estados Unidos e na Grã-Bretanha. Mas mesmo uma corporação assumindo o papel de (1) como agente capaz de cumprir lei, este artificio não causa estranheza, pelo menos hoje, pois há um entendimento de que um grupo de pessoas está por trás desta formação. O conceito de "personalidade corporativa" é uma ferramenta legal criada para permitir o funcionamento das corporações dentro do sistema legal, e não uma afirmação de que as corporações são "pessoas" no sentido literal ou moral. Nas considerações finais esta ocupação será retomada.

No seu surgimento, as corporações[128]apresentaram diversos problemas e receberam inúmeras críticas. Em *A Riqueza das Nações de* 1776, Adam Smith advertiu o perigo dos administradores “com o dinheiro de outros” terem “negligência e esbanjamento”. Nesta época, após inúmeros escandalosos, a formação de corporação chegou a ficar proibida na Inglaterra. De certa maneira, algumas centenas de anos foram necessários para naturalizar este conceito de "personalidade corporativa" na sociedade. Hoje aceitamos “pessoas jurídicas”

[127] Disponível em <http://www.dhnet.org.br/direitos/anthist/hamurabi.htm> Acesso em outubro de 2022. Tradução nossa.
[128] BAKAN, Joel. *A Corporação – A Busca patológica por lucro e poder*. 1ª ed. São Paulo: Novo Conceito, 2007. P. 7. Disponível em <http://www.dhnet.org.br/direitos/anthist/hamurabi.htm> Acesso em outubro de 2022. Tradução nossa.

como um conceito normal, talvez como fruto do esforço contínuo por parte das próprias corporações.

> A corporação inteligente entende que as pessoas fazem comparações em termos humanos [...] porque é assim que as pessoas pensam, nós pensamos em termos que muitas vezes são muito, muito pessoais [...] se você caminhar pela rua com um microfone e uma câmera e parar [pessoas] na rua [...] elas vão descrever [as corporações] em termos muito humanos. […] Hoje, as corporações usam o *branding* para criar personalidades únicas e atraentes para si mesmas. O *branding* vai além das estratégias criadas para simplesmente associar as corporações aos seres humanos de verdade – como nas antigas campanhas da AT&T que mostravam trabalhadores e acionistas ou no mais recente uso do endosso de personalidades (como nas propagandas da Nike com Michael Jordan) e de mascotes corporativos (como Ronald McDonald's, o Tigre Tony, o homenzinho da Michelin e o Mickey Mouse). (BAKAN.2007. P. 30)

Na tradição da filosofia da tecnologia há um debate sobre a autonomia da tecnologia para autores como Heidegger, Ellul, Latour, entre outros. As visões sobre a tecnologia ditando o seu próprio rumo independentemente da vontade da sociedade podem variar, mas geralmente envolvem uma análise crítica das relações entre tecnologia, sociedade e poder. Esses autores argumentam que a tecnologia não é apenas uma ferramenta neutra, mas tem seus próprios imperativos e influencia profundamente a forma de vida em sociedade. Essa crítica desafia a ideia de que a tecnologia está sob total controle humano, levantando questões sobre a ética e as consequências imprevistas do desenvolvimento tecnológico. Mas essa autonomia da tecnologia jamais foi caracterizada como uma agência capaz de receber cumprimento das leis como seres humanos expressas em PIA.

> O pensamento dominante da filosofia da tecnologia da fase inicial clássica com Heidegger, Ellul, com ideia de determinismo e tecnologia como uma força autônoma ganha reforço com IA. O que parecia exagero na época com o advento da IA parece profético. O ser humano disponível, é na forma de dados. O uso de plástico é um bom exemplo de como uma tecnologia que facilita vida humana pode se mostrar devastador para o ambiente e para seres humanos. O ciclo foi demorado, quando notamos o seu lado negativo deparamos com problemas em um nível já avançado e quase irreversível. (GIACOIA JR., Oswaldo. 2004. p. 637-654.)[129]

[129] GIACOIA JR., Oswaldo. Um direito próprio da natureza? Notas sobre ética, direito e tecnologia. In: Fragmentos de Cultura, Goiânia, 2004. v. 14, n. 4, p. 637-654.

A perspectiva de Giacoia sugere que com o advento da IA, a ideia de autonomia e determinismo de autores como Heidegger, Ellul, o que antes parecia um exagero ganha reforço com IA, mas essas ideias proféticas têm sentido mais metafórico. A mudança fundamental não se localiza apenas no nível da "força autônoma". O que presenciamos é um irromper na natureza desta autonomia e determinismo que está em curso com [R a H].

2.2 - PIA em [R a H]

Dos quarenta PIA reunidos neste trabalho (Capítulo 1 da Parte I e apêndice), destacamos os PIA que apresentam a fórmula [R a H]. Como foi mencionado antes, não faz parte do escopo deste trabalho analisar outros PIA do tipo [H a H], pois continuam sob escrutínio do direito tradicional onde o agente da ação não é [R], mas permanecem na forma [H^1 a H^2], onde a responsabilidade da lei está em [H^1].

As três leis da robótica - Isaac Asimov (1942. EUA)

1. Um robô não pode prejudicar um ser humano ou, por omissão, permitir que o ser humano sofra dano.
2. Um robô tem de obedecer às ordens recebidas dos seres humanos, a menos que contradigam a Primeira Lei,
3. Um robô tem de proteger sua própria existência, desde que essa proteção não entre em conflito com a Primeira ou a Segunda Lei.

[R a H]

Novo axioma das leis da sociedade com agência da IA

As três leis da robótica de Asimov inauguraram um novo axioma das leis da sociedade. Como foi visto no capítulo anterior, as alterações eram permitidas apenas na posição [H^2]. O sujeito jurídico [H^1] era uma entidade pessoa física ou jurídica capaz de ser titular de direitos e deveres jurídicos. Na história da humanidade, nem sempre os direitos e deveres jurídicos

eram atribuídos simultaneamente [130]. As desigualdades de gênero e classe têm raízes profundas na história, com muitos grupos marginalizados, como escravos e mulheres, sendo negados direitos civis e políticos, mesmo quando se esperava que eles cumprissem deveres sociais e legais.

Declaração Mundial de Robôs Feira Internacional de Robôs de Fukuoka (2004. Japão)

> 1. Robôs de próxima geração serão parceiros que coexistem com seres humanos
> 2. Os robôs da próxima geração ajudarão os seres humanos tanto física quanto psicologicamente
> 3. Os robôs da próxima geração contribuirão para a realização de uma sociedade segura e pacífica

Carta de ética do robô da Coreia – 2007

> Capítulo 2 (Princípios de humanos e robôs): Humanos e robôs devem aderir à dignidade da vida, da inteligência e da ética em engenharia.
>
> Capítulo 4 (Ética do robô): Os robôs não devem ferir os seres humanos como amigos, ajudantes e companheiros que obedecem a seus comandos.
>
> Capítulo 6 (Ética do usuário): Os usuários de robôs devem respeitar os robôs como amigos humanos e proibir modificações ilegais ou abuso de robôs.

Há princípios que anunciam reciprocidade entre os seres humanos e robôs e IA com a mistura de agentes como nos princípios 4º e 6º da Carta de ética do robô Coreia e alguns outros.

[R a H] e [H a R]

Novo axioma das leis da sociedade com agência da IA

[130] Esta condição contraditória reflete desigualdades históricas profundamente arraigadas na sociedade. Em relação à escravidão, os escravos eram considerados propriedade e não pessoas. Eles podiam ser comprados, vendidos e obrigados a trabalhar sem qualquer direito legal. Embora fossem esperados para cumprir deveres (trabalho), não tinham direitos porque não eram reconhecidos como "seres humanos plenos". As mulheres, por outro lado, eram frequentemente vistas como subordinadas aos homens e incapazes de exercer responsabilidades civis ou políticas, como a negação de direitos como o de voto ou a posse de propriedade. No entanto, as mulheres ainda eram obrigadas a cumprir deveres jurídicos, como pagar impostos.

As três leis da robótica responsável - IEEE (2009. Mundial)

2. Um robô deve responder aos humanos conforme apropriado para suas funções.

3. Um robô deve ser dotado de autonomia situada suficiente para proteger sua própria existência, desde que tal proteção forneça uma transferência suave de controle que não entre em conflito com a Primeira e a Segunda Leis.

Carta de Ética de Robôs (2012. Coreia do Sul)

Parte 3: Direitos e responsabilidades dos robôs

Seg. 1: Responsabilidades dos Robôs

i) Um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal.

ii) Um robô deve obedecer a quaisquer ordens que lhe sejam dadas por seres humanos, exceto quando tais ordens entrarem em conflito com a Parte 3, Seção 1, subseção “i” desta Carta.

iii) Um robô não deve enganar um ser humano.

Seção 2: Direitos dos Robôs

De acordo com a lei coreana, os robôs têm os seguintes direitos fundamentais:

i) O direito de existir sem medo de ferimentos ou morte.

ii) O direito de viver uma existência livre de abusos sistemáticos.

Oito leis da robótica de Shinpo Fumio - Keio University (2015. Japão)

1) A humanidade em primeiro lugar - os robôs não podem prejudicar ou se tornar pessoas.

2) Obediência à ordem — devem seguir ordens humanas e estar sujeitos a controle.

7) Participação individual — os indivíduos devem participar da criação de regras que regem os robôs, e os robôs não devem reger os indivíduos.

As lei de Satya Nadella (2016. EUA)

2. "A.I. deve ser transparente", o que significa que os humanos devem saber e ser capazes de entender como funcionam.

3. "A.I. deve maximizar a eficiência sem destruir a dignidade das pessoas".

6. "A.I. deve se proteger contra o preconceito" para que não discrimine as pessoas.

Três princípios para criar inteligência artificial segura (ou IA compatível com humanos) - Stuart Russell (2017. EUA)

1. O único objetivo do robô é maximizar a realização dos valores humanos.

2. O robô está inicialmente incerto sobre quais são esses valores

3. O comportamento humano fornece informações sobre os valores humanos

Três Regras para Sistemas de Inteligência Artificial - CEO do Allen Institute for Artificial Intelligence (2017. EUA)

1. Um sistema IA deve estar sujeito a toda a gama de leis que se aplicam ao seu operador humano.

2. Um sistema IA deve revelar claramente que não é humano.

3. Um sistema IA não pode reter ou divulgar informações confidenciais sem a aprovação explícita da fonte dessas informações.

23 Princípios de Asilomar (2017. EUA)

8) Transparência Judicial: Qualquer envolvimento de um sistema autônomo na tomada de decisão judicial deve fornecer uma explicação satisfatória auditável por uma autoridade humana competente.

14) Benefício compartilhado: as tecnologias de IA devem beneficiar e capacitar o maior número possível de pessoas.

15) Prosperidade Compartilhada: A prosperidade econômica criada pela IA deve ser amplamente compartilhada, para beneficiar toda a humanidade.

17) Não subversão: O poder conferido pelo controle de sistemas de IA altamente avançados deve respeitar e melhorar, ao invés de subverter, os processos sociais e cívicos dos quais depende a saúde da sociedade.

Problemas de longo prazo

22) Auto-aperfeiçoamento recursivo: Os sistemas de IA projetados para auto-aperfeiçoamento ou auto-replicação recursiva de uma maneira que possa levar a um aumento rápido da qualidade ou quantidade devem estar sujeitos a medidas rígidas de segurança e controle.

23) Bem Comum: A superinteligência só deve ser desenvolvida a serviço de ideais éticos amplamente compartilhados e para o benefício de toda a humanidade, e não de um estado ou organização.

Há princípios que anunciam deveres para robôs e IA com eles mesmos como verificado no 22º princípio de Asilomar e anteriormente nos 2º e 3º princípios de Stuart Russell**.**

[R a R]

Novo axioma das leis da sociedade com agência da IA

Princípios para a Governança da IA - The Future Society (2017. EUA e UE)

> Princípio 1: A IA não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. Consequentemente, o objetivo de governar a inteligência artificial é desenvolver estruturas políticas, códigos ou práticas voluntárias, diretrizes práticas, regulamentações nacionais e internacionais e normas éticas que protejam e promovam a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos.

Três ideias da Iniciativa de IA centrada no ser humano de Stanford (HAI) (2018. EUA)

> 3. O objetivo final da IA deve ser aumentar nossa humanidade, não a diminuir ou substituí-la.

Princípios Harmoniosos de Inteligência Artificial – HAIP (2018. China)

> Código Concreto
>
> 2. Princípios de Inteligência Artificial.
>
> > (7) Privacidade para humanos: a IA precisa respeitar a privacidade humana. E não tem o direito de utilizar e compartilhar informações privadas de humanos sem confirmação explícita.
> >
> > (8) Viés no ser humano: a IA não pode introduzir viés para entender e interagir com a humanidade e deve interagir ativamente com humanos para remover o viés potencial gerado.
> >
> > (9) Responsabilidade pelo ser humano: a IA precisa manter o ser humano seguro, com base em que essa consideração de segurança não prejudique direta e indiretamente a sociedade humana. A IA precisa ajudar o ser humano na transformação para o futuro ser humano.
> >
> > (10) Moralidade e Ética Comuns: Sendo parte da harmoniosa sociedade Humano-IA, a IA deve maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade e tratar outras IAs de vida consciente com princípios semelhantes.
> >
> > (11) Restrições legais para IA: A IA precisa obedecer às restrições legais para que o ser humano faça parte da sociedade.

(12) Proteção de existência: Para IA de vida consciente, eles precisam proteger sua própria existência com base em não prejudicar a existência de seres humanos e outras IAs de vida consciente, a menos que quebrem as restrições legais que fazem com que as decisões legais baseadas em leis não cumpram eles vivos.

4. Princípios compartilhados para humanos e IA

Esta seção define princípios compartilhados que humanos e IA precisam seguir:

(18) Colaboração: humanos e IA precisam colaborar para os avanços e o futuro de longo prazo de ambos os lados.

(19) Coordenação: Quando os conflitos emergirem das interações entre humanos e IA, os benefícios para a humanidade e os benefícios para a IA devem ser ativamente coordenados com base na empatia e no altruísmo.

(20) Confiança Mútua: Humanos e IA precisam desenvolver e elevar os níveis de confiabilidade entre si.

Microsoft responsible AI principles (2018. EUA)

Imparcialidade

- Os sistemas de IA devem tratar todas as pessoas de maneira justa

Confiabilidade e Segurança

- Os sistemas de IA devem funcionar de forma confiável e segura

Privacidade e segurança

- Os sistemas de IA devem ser seguros e respeitar a privacidade

Inclusão

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

Transparência

- Os sistemas de IA devem ser compreensíveis

Responsabilidade

- As pessoas devem ser responsáveis pelos sistemas de IA

Declaração de Montreal pelo desenvolvimento responsável da Inteligência Artificial (2018. Canadá)

1 PRINCÍPIOS DO BEM-ESTAR

O desenvolvimento e o uso de Sistemas de Inteligência Artificial (SIAs) devem permitir aumentar o bem-estar de todos os seres sencientes.

1. Os Sistemas de Inteligência Artificial (SIAs) devem permitir que os indivíduos melhorem suas condições de vida, de saúde e de trabalho.

2. Os SIAs devem permitir que os indivíduos satisfaçam suas preferências, dentro dos limites do que não cause danos a outro ser senciente.

3. Os SIAs devem permitir que os indivíduos exerçam suas capacidades físicas e intelectuais.

4. Os SIAs não devem ser fonte de desconforto, a menos que este último possa gerar um maior bem-estar que não possamos alcançar de outra forma.

5. O uso de SIAs não deve contribuir para o aumento do estresse, da ansiedade e de sentimentos de assédio ligados ao ambiente digital.

2 RESPEITO À AUTONOMIA

Os SIAs devem ser desenvolvidos e usados respeitandose a autonomia das pessoas e com o objetivo de aumentar o controle, pelos indivíduos, de sua vida e de seu meio ambiente.

1. Os SIAs devem capacitar os indivíduos a realizar seus próprios objetivos morais e sua concepção de uma vida digna de ser vivida.

5. Os SIAs não devem ser desenvolvidos para propagar informações não confiáveis, mentiras e propaganda, e devem ser projetados com o propósito de reduzir tal propagação.

6. O desenvolvimento dos SIAs deve evitar criar dependências por meio de técnicas de captação da atenção e de imitação da aparência humana, que possam induzir a uma confusão entre SIAs e seres humanos.

4 SOLIDARIEDADE

O desenvolvimento de SIAs deve ser compatível com a manutenção de relações solidárias entre pessoas e gerações.

1. Os SIAs não devem prejudicar a preservação de relações afetivas e morais que floresçam entre as pessoas, e devem ser desenvolvidos com o objetivo de fomentá-las de modo a reduzir a vulnerabilidade e o isolamento das pessoas.

2. Os SIAs devem ser desenvolvidos para colaborar com os seres humanos em tarefas complexas e devem fomentar o trabalho colaborativo entre humanos.

3. Os SIAs não devem ser implementados para substituir pessoas em tarefas que exigem relacionamento humano de qualidade, mas devem ser desenvolvidos para facilitar essas relações.

4. Os sistemas de saúde que usam SIAs devem levar em consideração a importância, para os pacientes, das relações com equipe médica e a família.

5. O desenvolvimento de SIAs não deve estimular comportamentos cruéis com robôs que se apresentam como seres humanos ou animais e que pareçam agir como eles.

6. Os SIAs devem ajudar a melhorar o gerenciamento de riscos e criar as condições para uma sociedade mais eficaz de compartilhamento de riscos individuais e coletivos.

5 PRINCÍPIOS DA PARTICIPAÇÃO DEMOCRÁTICA

2. As decisões dos SIAs que afetam a vida, a qualidade de vida ou a reputação dos indivíduos devem sempre ser justificadas em linguagem compreensível para aqueles que os utilizam ou que sofrem as consequências de seu uso. As justificativas consistem em explicar os fatores e parâmetros mais importantes para a tomada de uma decisão, e devem ser semelhantes às justificativas que seriam exigidas de um ser humano que tomasse o mesmo tipo de decisão.

6. Para que os SIAs públicos tenham um impacto significativo na vida dos cidadãos, estes devem ter a oportunidade e a competência para deliberar sobre os seus parâmetros sociais, objetivos e limites de uso.

7. Deve ser assegurado em todos os momentos que os SIAs façam aquilo para o qual foram programados e para o qual devem ser utilizados.

8. Todo usuário de um serviço deve saber se uma decisão que lhe diz respeito ou que o afete foi tomada por um SIA.

9. Todo usuário de um serviço que use agentes de conversação deve ser capaz de identificar facilmente se está interagindo com um SIA ou com uma pessoa.

10. A pesquisa no campo da inteligência artificial deve permanecer aberta e acessível a todos.

7 PRINCÍPIOS DA INCLUSÃO DA DIVERSIDADE

1. O desenvolvimento e o uso de SIAs não devem levar a uma padronização da sociedade através da normalização de comportamentos e opiniões.

2. O desenvolvimento e a implantação de SIAs devem levar em conta as múltiplas expressões de diversidade social e cultural, e isso deve ocorrer desde a concepção dos algoritmos.

4. Os SIAs devem evitar o confinamento de indivíduos em um perfil de usuário ou em uma "bolha filtradora", evitar definir identidades pessoais por meio do processamento de dados obtidos a partir de suas atividades anteriores, e também evitar a redução de suas opções de desenvolvimento pessoal, especialmente nas áreas da educação, da justiça e das práticas empresariais.

5. Os SIAs não devem ser usados ou desenvolvidos para limitar a liberdade de expressar ideias e de comunicar opiniões, cuja diversidade é a condição da vida democrática.

6. Para cada categoria de serviço, a oferta dos SIAs deve ser diversificada para que os monopólios de fato não se constituam e não prejudiquem as liberdades individuais.

9 PRINCÍPIOS DA RESPONSABILIDADE

O desenvolvimento e o uso de SIAs não devem contribuir para a desresponsabilização dos seres humanos quando uma decisão vier a ser tomada.

1. Somente os seres humanos podem ser responsabilizados por decisões decorrentes de recomendações feitas por SIAs e pelas ações decorrentes delas.

2. Em todas as áreas onde deva ser tomada uma decisão que afeta a vida, a qualidade de vida ou a reputação de uma pessoa, além de a decisão final dever recair sobre ser humano, deve ser livre e informada.

3. A decisão de matar deve sempre ser tomada por seres humanos e a responsabilidade por esta decisão não pode ser transferida para um SIA.

4. Pessoas que autorizem um SIA a cometer um crime ou delito, ou que sejam negligentes ao permiti-los, são responsáveis por eles.

5. No caso de um erro ter sido infligido por um SIA, e o SIA se provar confiável e tiver sido usado de maneira normal, não é razoável culpar as pessoas envolvidas em seu desenvolvimento ou uso.

Em alguns princípios como a 5ª cláusula do 7 PRINCÍPIOS DA INCLUSÃO DA DIVERSIDADE, há possibilidade de desenvolvimento futuro da IA que possibilitaria a aplicação [R a H] como "Os SIAs não devem ser desenvolvidos para propagar informações não confiáveis", ou seja, na sua forma atual seria "Os SIAs não devem propagar informações não confiáveis".

Recomendações sobre a inclusão da África subsaariana na ética global da IA (2019. África)

Princípio 4: Praticar IA justa e socialmente responsável

A IA deve ser justa e inclusiva, levando em consideração as variações e granularidades do continente.

Os Oito Princípios de Ética da Inteligência Artificial (IA) da Austrália (2019. Austrália)

Visão geral dos princípios

Bem-estar humano, social e ambiental: os sistemas de IA devem beneficiar os indivíduos, a sociedade e o meio ambiente.

Valores centrados no ser humano: os sistemas de IA devem respeitar os direitos humanos, a diversidade e a autonomia dos indivíduos.

Justiça: os sistemas de IA devem ser inclusivos e acessíveis e não devem envolver ou resultar em discriminação injusta contra indivíduos, comunidades ou grupos.

Por fim, destacamos também alguns PIA e cartas que peremptoriamente atribuem a agência e responsabilidade aos seres humanos e/ou fabricantes em PIA como a "Declaração de Montreal" e "UNI Global Union" que até intercalam [131][H a H] e [R a H]. Autores como Balkin (2017) e Cofone (2018) deixam claro a preocupação centrada na responsabilidade humana no desenvolvimento da IA criticando os princípios centrados na agência da IA.

> Vou divergir de Asimov neste ponto. Em vez de focar nas leis dirigida a robôs (ou algoritmos), concentro-me em leis dirigidas às pessoas que programar e usar robôs, agentes AI e algoritmos. Isso porque o que nós necessidade na emergente Sociedade Algorítmica não são as leis da robótica, mas as leis da operadores de robôs. (BALIKIN. 2017. P. 221)

> O que realmente precisamos são leis de projetistas e operadores de robótica. As leis de robótica de que precisamos em nossa era são leis que controlam e direcionam seres humanos que criam, projetam e empregam robôs. (COFONE. 2018. p. 189 apud HALLEVY. 2010. 2023. 2016)[132]

[131] Acreditamos que há essa mistura em agentes que denuncia o caráter provisório que apresenta certa confusão.

[132] COFONE, Ignacio, N. Servers and Waiters: What Matters in the Law of A.I. Disponível em <https://law.stanford.edu/wp-content/uploads/2018/09/Cofone_LL_20180905-1.pdf> Acesso em outubro de 2022. Tradução nossa.

Cinco princípios éticos para a robótica EPSRC/AHRC (2011. Reino Unido)

2. Humanos, não robôs, são agentes responsáveis. Os robôs devem ser projetados; operado na medida do possível para cumprir as leis existentes e direitos e liberdades fundamentais, incluindo privacidade.

3. Os robôs são produtos. Eles devem ser projetados usando processos que garantam sua segurança e proteção.

4. Os robôs são artefatos manufaturados. Eles não devem ser projetados de maneira enganosa para explorar usuários vulneráveis; em vez disso, sua natureza de máquina deve ser transparente.

5. Deve ser atribuída a pessoa com responsabilidade legal por um robô.

Carta aberta à Comissão Europeia da Inteligência Artificial e Robótica (2018. EUROPA)

Do ponto de vista ético e legal, criar uma personalidade jurídica para um robô é inapropriado qualquer que seja o modelo de status legal:

a. Um estatuto jurídico para um robô não pode derivar do modelo de Pessoa Física, pois o robô passaria a deter direitos humanos, como o direito à dignidade, o direito à integridade, o direito à remuneração ou o direito à cidadania, portanto diretamente enfrentamento dos direitos humanos. Tal estaria em contradição com a Carta dos Direitos Fundamentais da União Europeia e a Convenção para a Proteção dos Direitos do Homem e das Liberdades Fundamentais.

Princípios de IA da OCDE (2019. Mundial)

1.5. Responsabilidade

Os atores da IA devem ser responsáveis pelo bom funcionamento dos sistemas de IA e pelo respeito aos princípios acima, com base em suas funções, no contexto e de acordo com o estado da arte.

Declaração de Montreal pelo desenvolvimento responsável da Inteligência Artificial (2018. Canadá)

9 PRINCÍPIOS DA RESPONSABILIDADE

O desenvolvimento e o uso de SIAs não devem contribuir para a desresponsabilização dos seres humanos quando uma decisão vier a ser tomada.

1. Somente os seres humanos podem ser responsabilizados por decisões decorrentes de recomendações feitas por SIAs e pelas ações decorrentes delas.

Robot Ethics Charter Korea – 2007

> Capítulo 5 (Ética do fabricante): O fabricante do robô tem o dever de fabricar robôs que protegem a dignidade humana, reciclar robôs e proteger informações.

Os 10 principais princípios para a inteligência artificial ética - UNI Global Union (2017. Mundial)

> 9. Proibir a Atribuição de Responsabilidade a Robôs
>
> > Os robôs devem ser projetados e operados na medida do possível para cumprir as leis existentes, direitos e liberdades fundamentais, incluindo privacidade. Isso está ligado à questão da responsabilidade legal. De acordo com Bryson et al 2011, UNI Global Union afirma que a responsabilidade legal por um robô deve ser atribuída a uma pessoa. Os robôs não são partes responsáveis perante a lei.

Verificamos que muitos PIA apresentam a fórmula [R a H], mas ao mesmo tempo, há PIA que defendem a responsabilidade exclusiva de seres humanos: “A responsabilidade legal por um robô deve ser atribuída a uma pessoa”, “Somente os seres humanos podem ser responsabilizados por decisões decorrentes de recomendações feitas por SIAs e pelas ações decorrentes delas”, “Humanos, não robôs, são agentes responsáveis”. A questão de se a IA pode ser considerada como agente é altamente complexa e continua a ser um tópico de debate acirrado entre acadêmicos, juristas, filósofos e outros profissionais.

2.3 - Agência de [R]

No sistema jurídico, o agente capaz de adquirir direitos e assumir obrigações é tratado pelos termos "sujeito jurídico" e "sujeito de direito". O sujeito jurídico é qualquer entidade (pessoa física ou jurídica) capaz de ser titular de direitos e deveres jurídicos. Autores como Hans Kelsen, em sua Teoria Pura do Direito (1934), sublinham que o sujeito jurídico é fundamentalmente definido por sua capacidade para ter direitos e obrigações. O sujeito de direito é um termo amplamente usado de maneira intercambiável com sujeito jurídico. No entanto, em uma interpretação mais restrita, pode referir-se especificamente a indivíduos ou entidades que possuem direitos e deveres de acordo com a lei. Assim, pode-se dizer que todos os sujeitos de direito são sujeitos jurídicos, mas nem todos os sujeitos jurídicos são necessariamente sujeitos de direito. Em alguns contextos, entidades como corporações ou governos podem ser consideradas sujeitos jurídicos, mas não sujeitos de direito. Neste

trabalho optamos pelo termo agente pois avaliamos que o termo sujeito já sugere um status posterior, além da ligação mais automática do termo agente com agência.

Há também subjetividade jurídica, que é um conceito mais abstrato que se refere à condição de ser um sujeito de direito ou um sujeito jurídico. A subjetividade jurídica implica ter a capacidade de ser titular de direitos e deveres na ordem jurídica. É, portanto, um status que é conferido pela lei. A subjetividade jurídica não se refere apenas a indivíduos humanos, mas também a entidades legais, como empresas, que podem ter direitos e obrigações.

Como verificamos no capítulo anterior, John Austin definiu a lei de seguinte maneira, “uma lei é uma regra estabelecida para a orientação de um ser inteligente por um ser inteligente que tem poder sobre ele”. Segundo Austin, as leis são estabelecidas por seres inteligentes, ou seja, uma autoridade competente que tem a capacidade de pensar, raciocinar e entender as implicações das regras que estão estabelecendo. Esta autoridade precisa ser inteligente para criar leis justas, equitativas e eficazes que sejam apropriadas para a sociedade que elas governam. Austin acredita que as leis são destinadas a orientar os comportamentos dos seres inteligentes. Em outras palavras, as pessoas (ou os seres inteligentes) têm a capacidade de compreender as leis e de fazer escolhas informadas sobre se vão ou não as seguir.

As leis, portanto, são mecanismos para orientar o comportamento, e a eficácia desses mecanismos depende da capacidade dos seres inteligentes de entender e responder a eles. Ou seja, a ideia de um "ser inteligente" é fundamental para a concepção de Austin sobre o que é a lei, tanto no que diz respeito a quem estabelece as leis quanto a quem as leis se destinam a guiar. A lei, neste contexto, é um produto da inteligência e requer inteligência para ser compreendida e aplicada. Portanto, a inteligência é um requisito básico para o funcionamento eficaz do sistema legal.

As tentativas de atribuir agência para Inteligência Artificial não configuram apenas uma coincidência semântica. A inteligência é um atributo essencial para a IA e é o que a torna inovadora. Apesar de ainda ter um critério amplamente estabelecido, a IA, muitas vezes, é considerada inteligente pela sua capacidade de realizar tarefas que normalmente requerem inteligência humana como compreensão de linguagem natural, reconhecimento de voz,

aprendizado, planejamento, raciocínio e percepção. A IA ainda demonstra capacidade de aprender e melhorar com a experiência (aprendizado de máquina), ou seja, ela consegue se adaptar com base em novas informações e situações. Assim, cada vez mais o atributo de inteligência para a IA é normalizado na sociedade, ainda é o que permite que a IA imite e até causa temor pela possibilidade de superar a inteligência humana em certos aspectos.

Entretanto, o aspecto da inteligência não seria uma condição única nem suficiente para a garantia de agência jurídica para a IA. Há um debate contínuo sobre esta questão. Os autores (Giuffrida et al. 2018) do artigo "Uma Perspectiva Jurídica sobre os Desafios e Dificuldades da Inteligência Artificial"[133]analisa esta questão introduzindo o critério da razão: "A inteligência, por si só, não é suficiente para determinar a personalidade, pelo menos na maior parte das jurisdições. Em vez disso, o critério utilizado é o da razão" e citam Erich Fromm.

> A razão é a faculdade do homem de apreender o mundo pelo pensamento, em contradição com a inteligência, que é a capacidade do homem de manipular o mundo com a ajuda do pensamento. A razão é o instrumento do homem para chegar à verdade, a inteligência é o instrumento do homem para manipular o mundo com mais sucesso; o primeiro é essencialmente humano, este último pertence à parte animal do homem.[134]

Giuffrida et al. (2018) oferecem exemplo com os indivíduos acometidos pela síndrome de savant, caracterizada por pessoas com deficiências mentais, incluindo transtorno do espectro autista, que apresentam algumas áreas restritas de genialidade em contraste marcante e incongruente com suas limitações gerais. Os indivíduos afetados por essa condição frequentemente exibem habilidades de cálculo impressionantes, porém, ainda assim, podem ser considerados legalmente incapazes. Concluem que independentemente de se concordar ou não com o postulado de Fromm, é incontestável que a inteligência e a razão são conceitos que caminham lado a lado. Isso destaca que a "inteligência", no sentido mais amplo do termo, é mais complexa e multifacetada do que meras habilidades computacionais.

[133] GIUFFRIDA. Iria, LEDERER. Fredric. VERMERYS. Nicolas. A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law, 68 Case W. Res. L. Rev. 747. 2018. p. 766. Disponível em <https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/14> Acesso em maio de 2019. Tradução nossa.
[134] GIUFFRIDA apud Erich Fromm 2018. p. 766.

Além disso, Giuffrida et al. (2018) afirmam que a inteligência, por si só, parece não se qualificar como razoável, mesmo que algum tipo de consciência pudesse emergir. Os animais possuem consciência, mas não são considerados aptos a serem submetidos a penalidades legais, uma vez que não têm qualquer indicação de que sejam capazes de refletir sobre suas ações como sendo próprias. A consciência dos animais se restringe ao ambiente ao seu redor, sem a consciência de ter consciência, um traço típico do senso de autoidentidade humano. Para estar apto à repreensão, em vez de simplesmente à disciplina, um indivíduo precisa ter consciência de si mesmo. Isso possibilita um nível de reflexão que pode conduzir à contestação ou arrependimento no caso de uma acusação criminal. Para ser passível de censura, é preciso haver a capacidade de auto-observação e autoanálise, características que, até onde sabemos, não são possíveis nem para animais nem para a Inteligência Artificial.

Esta condição exemplificada por Giuffrida através de síndrome de savant e animais, no direito seria a incapacidade civil, a condição em que um indivíduo é considerado incapaz de exercer pessoalmente os atos da vida civil, seja por razões de idade, doença, deficiência ou qualquer outra circunstância que possa afetar sua capacidade de compreender as consequências de suas ações. No Brasil, a incapacidade civil é regulamentada pelo Código Civil e pode ser absoluta ou relativa. A incapacidade absoluta se aplica àqueles que, mesmo por causa transitória, não podem exprimir sua vontade, incluindo menores de 16 anos. Já a incapacidade relativa afeta aqueles que, por causa transitória ou permanente, não podem exprimir sua vontade de forma plena e consciente, incluindo menores entre 16 e 18 anos.

> Incapacidade civil – São as pessoas que não estão aptas ao exercício ou gozo de seus direitos. A incapacidade pode ser absoluta ou relativa. São absolutamente incapazes de exercer pessoalmente os atos da vida civil os menores de 16 anos; os que, por enfermidade ou deficiência mental, não tiverem o necessário discernimento para a prática desses atos; os que, mesmo por causa transitória, não puderem exprimir sua vontade. São relativamente incapazes os menores de 16 anos e maiores de 18 anos; os ébrios habituais, os viciados em tóxicos, e os que, por deficiência mental, tenham o discernimento reduzido; os excepcionais, sem desenvolvimento mental completo; os pródigos, entre outros. [135]

[135] Disponível em <https://www.mpf.mp.br/es/sala-de-imprensa/glossario-de-termos-juridicos> Acesso em outubro de 2022.

A incapacidade civil, conforme entendida pelas legislações de diversos países, é a condição em que um indivíduo é considerado incapaz de exercer pessoalmente os atos da vida civil, seja por razões de idade, doença, deficiência ou qualquer outra circunstância que possa afetar sua capacidade de compreender as consequências de suas ações. Em todos esses casos, a intenção é proteger os interesses dos incapazes, garantindo que eles não sejam prejudicados ou abusados por sua incapacidade de compreender e lidar com as complexidades das suas ações. Geralmente a questão é tratada com a nomeação de um tutor para agir em nome da pessoa incapaz.

A agência de [R] que tratamos aqui é específica, capacidade civil. Não discutiremos detalhadamente outras condições como inteligência, razão, senciência, consciência, autonomia, que em PIA são tratados como uma possibilidade futura para IA caracterizada pelos termos como IA avançada, IA altamente autônomas e superinteligência. A nossa abordagem é restrita na agência capaz de cumprir deveres expressas em PIA, mesmo que provavelmente para que esta condição seja atingida, algumas qualidades citadas seriam necessárias, e, não mencionaremos a parte dos direitos.[136]

Hoje, não há consenso sobre como lidar com essa questão de agência para IA. Poderia dizer que atualmente há mais dúvidas do que certezas. O segundo princípio dos "Cinco Princípios Fundamentais" da empresa de software Sage (2017. Reino Unido) é contra a possibilidade de agência da IA, "Não se deve permitir que a tecnologia se torne inteligente demais para ser responsável". David C. Vladeck (2014) no seu artigo "Máquinas sem princípios: regras de responsabilidade e Inteligência Artificial" indaga o seguinte:

> A questão conceitual que as máquinas pensantes autônomas colocarão é se é justo pensar neles como agentes de algum outro indivíduo ou entidade, ou se o sistema jurídico precisará decidir questões de responsabilidade em uma base diferente do que agência.[137]

[136] Apesar de alguns PIA abordarem os direitos de [R] como a Carta de Ética da Coreis de 2012:

i) O direito de existir sem medo de ferimentos ou morte.

ii) O direito de viver uma existência livre de abusos sistemáticos. Neste trabalho não trataremos diretamente sobre isso, mas consideramos fundamentais para até o funcionamento de PIA.

[137] David C. Vladeck, Máquinas sem princípios: regras de responsabilidade e Inteligência Artificial, 89 Wash. L. Rev. 117, 122 (2014)

Sem entrarmos a fundo no mérito da questão sobre a possibilidade de autonomia, liberdade e consciência para IA, pontuamos que quando se elaboram PIA com agência de [R], sinalizam que mesmo que não esteja citado explicitamente, espera que [R] seja capaz de cumprir responsabilidade legais. A possibilidade de aparição de robôs com características humanas como conversar como seres humanos foi prevista por autores como Turing (1950), mas apesar destas previsões e elaboração de leis, a capacidade necessária para [R] cumprir leis para convívio com seres humanos não foi explorada. O filósofo francês Francis Wolff (2018) afirma que a sociedade humana é moldada em função da natureza dos seus participantes, até então exclusivamente de seres humanos, o advento de novos integrantes certamente traz impacto nesse novo arranjo.[138]

Ray Kurzweil argumenta em seu livro “Singularidade está próxima: quando os humanos transcendem a biologia” (2018) que, em algum momento no futuro, a tecnologia avançará a tal ponto que a Inteligência Artificial será capaz de melhorar a si mesma, levando a um crescimento exponencial da inteligência. Este ponto, a "singularidade", mudará fundamentalmente a sociedade humana e possivelmente até a própria natureza humana.

No entanto, é importante notar que muitos acadêmicos e especialistas têm visões mais cautelosas ou céticas sobre a singularidade. Alguns argumentam que as previsões de Kurzweil e outros são muito otimistas e que existem barreiras significativas para a criação de uma Inteligência Artificial verdadeiramente auto melhorável. O filósofo francês, Jean-Gabriel Ganascia aborda o conceito de singularidade tecnológica no seu livro "*Le Mythe de la Singularité*" (2017), Ganascia, especialista em IA, oferece uma análise crítica deste conceito e argumenta que a singularidade é, na verdade, um mito.

Ele examina as previsões de cientistas e futuristas, como Kurzweil, e questiona a fundamentação dessas previsões e a falta de consenso entre os especialistas. Analisando os avanços na IA e as limitações que impedem seu desenvolvimento exponencial, Ganascia sugere que a singularidade tecnológica é um mito, pois a IA é incapaz de superar completamente a inteligência humana. Ganascia acredita que a cooperação entre humanos

[138] WOLFF, Francis. Três Utopias Contemporâneas. Editora Unesp. 2018.

e IA pode trazer benefícios significativos à sociedade, mas a ideia de uma singularidade iminente pode gerar medo e ansiedade infundados.

2.4 - Realizabilidade de [a]

Nesta seção pretendemos analisar quais são as ações expressas nos PIA e se elas são realizáveis. Na fórmula [R a H] onde [a] deve ser capacidade de uma ação possível de [R], não faz sentido proibir ou demandar uma incapacidade intrínseca como a 1ª lei de Asimov ou Russell. Uma crítica que [R a H] pode sofrer é que mesmo apresentando esse formato onde [R] é o agente da ação, se o estágio tecnológico não permite ainda a criação de [R] agente capaz, continuaria sendo [H a H] com [R] contido em [a], ou onde subentende-se que na forma "IA deve" há anterioridade de [H] que projeta [R] de tal forma. Há, porém, capacidades que IA poderia atingir no futuro. Em se tratando de uma tecnologia em constante avanço, faremos uma divisão do que é possível hoje e não. Vale ressaltar que a IA ainda está em evolução e que muitas das suas possibilidades ainda estão sendo exploradas. À medida que a tecnologia avança, a IA provavelmente continuará a diferenciar-se das tecnologias e artefatos do passado, ainda não podemos prever todas as realizações, que provavelmente podem ser até imprevisíveis. Assim, selecionamos 49 princípios e agrupamos pelos 3 temas: "Respeito à dignidade humanas", "Benefício e prosperidade para a humanidade" e "Transparência e confiabilidade". Em alguns casos, o formato original "IA deve ser desenvolvida ou projetada para" foi resumido por "Deve".

À medida que a IA avança e se torna mais integrada a vidas diárias, é essencial que a IA respeite a autonomia e a dignidade humana. Isso inclui, entre outras coisas, garantir que a IA não tome decisões por conta própria que possam afetar negativamente os direitos humanos ou a liberdade individual agindo sem preconceito ou discriminação.

R	Respeito à dignidade humanas	H
	- Aderir à dignidade da vida, da inteligência e da ética em engenharia. (Carta. 2007. Coreia do Sul) - Deve maximizar a eficiência sem destruir a dignidade das pessoas. (Nadella. 2016. EUA) - Devem "alinhar-se" com os valores humanos em toda a sua operação. (Direito Civil. 2017. UE) - Devem "ser" compatíveis com os ideais de dignidade humana, direitos, liberdades e diversidade cultural. (Direito Civil. 2017. UE) - Devem maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade. (HAIP. 2018. China) - Proteger sua própria existência com base em não prejudicar a existência de seres humanos e outras IAs de vida consciente. (HAIP. 2018. China) - Devem respeitar os direitos humanos, a diversidade e a autonomia dos indivíduos. (2019. Austrália) - Devem ser justa e inclusiva, levando em consideração as variações e granularidades do continente. (2019. África) - Devem permitir que os indivíduos satisfaçam suas preferências, dentro dos limites do que não cause danos a outro ser senciente. (2018. Canadá) - Devem permitir que os indivíduos exerçam suas capacidades físicas e intelectuais. (2018. Canadá) - Devem capacitar os indivíduos a realizar seus próprios objetivos morais e sua concepção de uma vida digna de ser vivida. (2018. Canadá) - Não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. (Future Society. 2017. EUA e UE) - Aumentar nossa humanidade, não a diminuir ou substituí-la. (HAI. 2018. EUA) - Resolver os conflitos emergirem das interações entre humanos e IA com base na empatia e no altruísmo. (HAIP. 2018. China) - Devem tratar todas as pessoas de maneira justa. (Microsoft. 2018. EUA) - Devem ser seguros e respeitar a privacidade. (Microsoft. 2018. EUA) - Devem ser inclusivos e acessíveis e não devem envolver ou resultar em discriminação injusta contra indivíduos, comunidades ou grupos. (2019. Austrália)	

A IA tem o potencial de trazer benefícios significativos para a humanidade, desde a melhoria da eficiência até a resolução de problemas complexos. No entanto, é fundamental que seja usada de maneira que promova a prosperidade para maior número possível, e não apenas para um grupo seleto de indivíduos ou empresas. Isso significa, entre outras coisas, garantir que os benefícios da IA sejam distribuídos de forma justa e equitativa.

R	Benefício e prosperidade para a humanidade	H
	- Ajudar a humanidade. (Allen. 2017. EUA) - Devem beneficiar e capacitar o maior número possível de pessoas. (Direito Civil. 2017. UE) - Devem compartilhar amplamente a prosperidade econômica criada pela IA para beneficiar toda a humanidade. (Asilomar. 2017. EUA) - Servir as pessoas e o planeta. (Future Society. 2017. EUA e UE) - Compartilhe os benefícios dos sistemas. (UNI Global Union. 2017. Mundial) - Agir em benefício de todos e para evitar permitir usos que prejudiquem a humanidade ou concentrem poder indevidamente. (OpenAI. 2018. EUA) - Maximizar a realização dos valores humanos. (Russell. 2017. EUA) - Precisa ajudar o ser humano na transformação para o futuro ser humano. (HAIP. 2018. China) - Devem beneficiar os indivíduos, a sociedade e o meio ambiente. (2019. Austrália) - Devem permitir que os indivíduos melhorem suas condições de vida, de saúde e de trabalho. (2018. Canadá) - Serem mais Empáticos e Altruístas para estabelecer uma Sociedade Humano-IA mais confiável, amigável e harmoniosa. (HAIP. 2018. China) - Precisa manter o ser humano seguro, com base em que essa consideração de segurança não prejudique direta e indiretamente a sociedade humana. (HAIP. 2018. China) - Deve maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade. (HAIP. 2018. China) - Tratar outras IAs de vida consciente com princípios semelhantes. (HAIP. 2018. China)* - Proteger sua própria existência com base em não prejudicar a existência de seres humanos e outras IAs de vida consciente. (HAIP. 2018. China)** - Devem capacitar todos e envolver as pessoas. (Microsoft. 2018. EUA)	

A transparência e a confiabilidade são fundamentais para construir a confiança na IA. Isso inclui, por exemplo, garantir que as decisões tomadas pela IA sejam compreensíveis e explicáveis para os humanos, e que a IA opere de maneira confiável e segura. A falta de transparência e confiabilidade pode levar a uma série de problemas, desde decisões injustas ou discriminatórias até falhas de segurança potencialmente catastróficas. A transparência e confiabilidade são também critérios previstos em regras das leis derivadas no capítulo anterior.

R	Transparência e confiabilidade	H
	- Estar sujeito a toda a gama de leis que se aplicam ao seu operador humano. (Carta. 2007. Coreia do Sul) - Deve basear-se nos princípios e valores consagrados no artigo 2.º do Tratado da União Europeia e na Carta dos Direitos Fundamentais. (Direito Civil. 2017. UE) - Precisam desenvolver e elevar os níveis de confiabilidade entre humanos e IA. (HAIP. 2018. China) - Devem funcionar de forma confiável e segura. (Microsoft. 2018. EUA) - Deve ser transparente. (Allen. 2017. EUA) - Obedecer às ordens recebidas dos seres humanos. (Asimov) - Revelar claramente que não é humano. (Allen. 2017. EUA) - Precisa manter o ser humano seguro, com base em que essa consideração de segurança não prejudique direta e indiretamente a sociedade humana. - Ser imparcial e sem gênero (Garanta uma IA imparcial e sem gênero) (Future Society. 2017. EUA e UE) - Devem funcionar de forma confiável e segura. (Microsoft. 2018. EUA) - Devem ser compreensíveis. (Microsoft. 2018. EUA)	

A preocupação humana por trás desses critérios é o medo de que a IA possa ser usada de maneiras que prejudiquem os indivíduos ou a sociedade como um todo. Isso pode incluir, por exemplo, a perda de empregos devido à automação, a violação de direitos humanos ou privacidade, ou o uso de IA para fins mal-intencionados, como a vigilância em massa ou a guerra cibernética. Ao incorporar esses princípios no design e uso da IA, espera-se minimizar esses riscos e garantir que a IA seja usada de maneira a beneficiar a humanidade como um todo.

Em seguida, classificamos os 49 princípios pelo critério da possibilidade de realização atual, futura e improvável. Separamos os que são possíveis atualmente em “Possíveis hoje” e o restante em “Talvez no futuro”, exceto os princípios que expressam realizações totalizantes como “agir em benefício de todos”, “maximizar a realização dos valores humanos”, “tratar todas as pessoas de maneira justa”, que entraram em “Improváveis mesmo no futuro”.

R	Possíveis hoje	H
	- Obedecer às ordens recebidas dos seres humanos. (Asimov) - Revelar claramente que não é humano. (Allen. 2017. EUA)	

R	Talvez no futuro	H
	- Aderir à dignidade da vida, da inteligência e da ética em engenharia. (Carta. 2007. Coreia do Sul) - Estar sujeito a toda a gama de leis que se aplicam ao seu operador humano. (Carta. 2007. Coreia do Sul) - Ajudar a humanidade. (Allen. 2017. EUA) - Deve ser transparente. (Allen. 2017. EUA) - Deve maximizar a eficiência sem destruir a dignidade das pessoas. (Nadella. 2016. EUA) - Devem beneficiar e capacitar o maior número possível de pessoas. (Direito Civil. 2017. UE) - Devem compartilhar amplamente a prosperidade econômica criada pela IA para beneficiar toda a humanidade. (Asilomar. 2017. EUA) - Servir as pessoas e o planeta. (Future Society. 2017. EUA e UE) - Ser imparcial e sem gênero (Garanta uma IA imparcial e sem gênero) (Future Society. 2017. EUA e UE) - Compartilhe os benefícios dos sistemas. (UNI Global Union. 2017. Mundial) - Precisa ajudar o ser humano na transformação para o futuro ser humano. (HAIP. 2018. China) - Devem respeitar os direitos humanos, a diversidade e a autonomia dos indivíduos. (2019. Austrália) - Devem ser inclusivos e acessíveis e não devem envolver ou resultar em discriminação injusta contra indivíduos, comunidades ou grupos. (2019. Austrália) - Devem permitir que os indivíduos melhorem suas condições de vida, de saúde e de trabalho. (2019. Austrália) - Devem ser desenvolvidos para colaborar com os seres humanos em tarefas complexas e devem fomentar o trabalho colaborativo entre humanos. (2019. Austrália) - Devem ser evitar o confinamento de indivíduos em um perfil de usuário ou em uma “bolha filtradora”, evitar definir identidades pessoais por meio do processamento de dados obtidos a partir de suas atividades anteriores, e também evitar a redução de suas opções de desenvolvimento pessoal, especialmente nas áreas da educação, da justiça e das práticas empresariais. (2019. Austrália) - Devem funcionar de forma confiável e segura. (Microsoft. 2018. EUA) - Devem ser seguros e respeitar a privacidade. (Microsoft. 2018. EUA) - Devem capacitar todos e envolver as pessoas. (Microsoft. 2018. EUA) - Devem ser compreensíveis. (Microsoft. 2018. EUA) - Devem tratar todas as pessoas de maneira justa. (Microsoft. 2018. EUA) - Serem mais empáticos e altruístas para estabelecer uma Sociedade Humano-IA mais confiável, amigável e harmoniosa. (HAIP. 2018. China)	

	- Precisa manter o ser humano seguro, com base em que essa consideração de segurança não prejudique direta e indiretamente a sociedade humana. (HAIP. 2018. China) - Tratar outras IAs de vida consciente com princípios semelhantes. (HAIP. 2018. China)* - Proteger sua própria existência com base em não prejudicar a existência de seres humanos e outras IAs de vida consciente. (HAIP. 2018. China)** - Resolver os conflitos emergirem das interações entre humanos e IA com base na empatia e no altruísmo. (HAIP. 2018. China) - Precisam desenvolver e elevar os níveis de confiabilidade entre humanos e IA. (HAIP. 2018. China) - Devem ser justa e inclusiva, levando em consideração as variações e granularidades do continente. (2019. África) - Deve basear-se nos princípios e valores consagrados no artigo 2.º do Tratado da União Europeia e na Carta dos Direitos Fundamentais. (Direito Civil. 2017. UE)	

R	Improváveis mesmo no futuro	H
	- Devem (ser projetados para que seus objetivos e comportamentos possam estar alinhados) “alinhar-se” com os valores humanos em toda a sua operação. (Direito Civil. 2017. UE) - Devem (ser projetados e operados de forma a serem) “ser” compatíveis com os ideais de dignidade humana, direitos, liberdades e diversidade cultural. (Direito Civil. 2017. UE) - Serem mais Empáticos e Altruístas para estabelecer uma Sociedade Humano-IA mais confiável, confiável, amigável e harmoniosa. (HAIP. 2018. China) - Deve maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade. (HAIP. 2018. China) - Maximizar a realização dos valores humanos. (Russell. 2017. EUA) - Agir em benefício de todos e para evitar permitir usos que prejudiquem a humanidade ou concentrem poder indevidamente. (OpenAI. 2018. EUA) - Devem tratar todas as pessoas de maneira justa. (Microsoft. 2018. EUA) - Deve maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade. (HAIP. 2018. China) - Não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. (Future Society. 2017. EUA e UE) - Aumentar nossa humanidade, não a diminuir ou substituí-la. (HAI. 2018. EUA)	

Muitas vezes, o teor das ações expressas nos PIA passa despercebido, pois o agente é a IA que tem o ineditismo. Mas se no lugar de [R] colocarmos [H], verificamos que muitos PIA parecem megalomaníacos e não fazem sentido, pelo menos hoje. Acreditamos que exceto dois princípios que são possíveis atualmente, todos os 47 PIA com possibilidades futuras ou remotas, denunciam uma preocupação que remete ao princípio da precaução de Hans Jonas

(2007). No cerne do conceito de Jonas está a ideia de que devemos agir de forma a evitar possíveis danos futuros, especialmente quando se trata do uso de tecnologias novas e potencialmente perigosas. Jonas (2007) argumenta que, na era tecnológica, a humanidade adquiriu um poder sem precedentes sobre a natureza e a vida humana. Esse poder vem com uma responsabilidade correspondente.

A abordagem de Jonas difere de outros princípios éticos que focam no cálculo de custo-benefício ou na maximização do bem-estar. Em vez disso, Jonas argumenta que devemos priorizar a prevenção do mal acima de tudo, mesmo que isso signifique renunciar a possíveis benefícios. Isso é particularmente relevante quando se trata de "riscos existenciais" - riscos que ameaçam a própria existência da humanidade. Este princípio tem implicações profundas para a ética da IA. Embora a IA possa trazer muitos benefícios, também apresenta riscos potenciais, incluindo o uso indevido da tecnologia e o impacto da automação no trabalho e na sociedade. O princípio da precaução sugere que devemos ser muito cautelosos ao desenvolver e implementar IA, para garantir que estamos minimizando esses riscos e protegendo o futuro da humanidade. No entanto, os PIA não apenas revelam as preocupações em concordância com o princípio de Jonas, expressas principalmente em “Respeito à dignidade humanas” e “Transparência e confiabilidade”, mas também verificamos os desejos humanos em “Benefício e prosperidade para a humanidade”. No próximo capítulo analisaremos as expectativas reveladas pelos PIA e porque elas são utópicas.

2.5 – Realizabilidade de [R]

Se no 2.3 verificamos a possibilidade de agência para [R], aqui trataremos sobre a realizabilidade de [R] além desta condição fundamental. Como imaginamos que há inúmeros requisitos, escolhemos três que consideramos mais relevantes e exemplificam melhor a dificuldade diante de [R] como agente: a questão de liberdade, a questão de coerção e recompensa e a questão de unidade para IA.

> A liberdade é um pré-requisito absoluto para que qualquer ato seja considerado moral; São Tomás afirma que um ato só é humano se for livre. A liberdade implica o conhecimento de alternativas e a capacidade de escolher entre elas. (MORRISON. 2012. P. 78)

Sabemos que liberdade é um assunto profundo e vasto na tradição filosófica. A possibilidade de escolha livre entre alternativas, sem dúvida deve ser requisito para uma ação ser considerada livre, mesmo que não adotemos o critério de moralidade. A liberdade neste trabalho, como a delimitação de liberdade apresentada por John Stuart Mill em sua obra "Sobre a Liberdade", é a liberdade civil ou social de agente no cumprimento dos PIA. Mais especificamente, nesse trabalho a liberdade é, dada a condição [R a H], [R] ter capacidade de gerar [R - a H], mas [- a] não pode ser resultado da falha em cumprir [R a H]. Se temos, "um robô não pode provocar danos em ser humano" ou "IA deve maximizar os valores humanos", devem também existir as alternativas para [R], a possibilidade da ocorrência de resultado contrário ou diversos daquele expresso em PIA como "um robô provocar danos em ser humano" ou "IA minimizar ou prejudicar os valores humanos".

Acreditamos ainda que podemos adotar a definição de Isaiah Berlin e classificar esta liberdade em positiva (trazer benefícios) e negativa (evitar danos). Como vimos no capítulo anterior, a maioria dos PIA só poderiam ser possíveis de realização futura. A IA não tem uma alternativa (liberdade) de ação hoje, nem no modo positivo como no negativo.

> Kant afirma que, quando um homem pauta suas ações pela lei devido ao terror ou à coação, estamos diante de um motivo meramente hipotético, mas que, quando o que o motiva é a aceitação da lei em si, o que temos é um ato em conformidade com a máxima categórica. A liberdade só existe no segundo caso, e provém da "autonomia da vontade" em oposição à "heteronomia" do agente, que opera em obediência não às exortações de sua reflexão racional, mas por paixão, medo ou esperança de recompensa. O agente heteronômico é verdadeiramente o agente "escravo", e, ainda que em sua falta de força ele desempenhe as ações observáveis da moralidade, buscou refúgio em sua sujeição à "natureza" e/ou à "força superior". Pode disfarçar sua mentalidade escrava e sua amoralidade em uma confusão de discursos, mas isso precisa estar sujeito ao exame crítico, e essa orientação crítica é necessária para a conquista da autonomia que lhe permitirá agir como ser racionalmente autônomo e, ao fazê-lo, atrair o respeito de outros seres racionais. (MORRISON. 2012. P. 178)

Kant argumenta que para ser verdadeiramente livre (ou autônomo), um indivíduo deve agir de acordo com princípios morais que ele mesmo reconhece como justos e corretos, em vez de ser guiado por desejos externos, medo, coação ou expectativa de recompensa (heteronomia). Atualmente, os sistemas de IA operam em um estado de completa "heteronomia", no sentido kantiano. São programas criados por seres humanos para realizar tarefas específicas e operar de acordo com algoritmos e dados fornecidos por seres humanos.

Não têm a capacidade de definir suas próprias “leis morais” ou de agir de acordo com qualquer tipo de consciência ou reflexão racional.

Em teoria, poderíamos imaginar uma IA avançada que possa operar de forma mais autônoma, tomando decisões com base em uma espécie de código moral interno, mas uma heteronomia kantiana é problemática para IA, uma vez que agir de acordo com os princípios morais que ela mesma reconhece como justos e corretos poderia ser uma grave ameaça à humanidade. O que revela que as diferenças fundamentais entre seres humanos e IA demandam até adequações nos conceitos anteriormente aplicados para seres humanos.

Um outro aspecto é a função de coerção e recompensa no cumprimento de leis na sociedade humana que são duas estratégias principais adotadas para encorajar o cumprimento das leis e a ordem social. A coerção refere-se ao uso de força ou ameaça para fazer com que os indivíduos ajam de certa maneira. No contexto da lei, a coerção muitas vezes assume a forma de penalidades ou punições para aqueles que violam as leis. Isso pode incluir multas, prisão ou outras sanções legais. A ideia é que o medo dessas punições desencorajará as pessoas de quebrar as leis.

As recompensas, por outro lado, são incentivos positivos que encorajam o comportamento desejado. No contexto da lei, as recompensas podem ser menos tangíveis. Mas viver em uma sociedade segura e ordenada, onde os direitos e liberdades são respeitados, pode ser visto como uma forma de recompensa. Em alguns casos, também podem haver recompensas mais concretas, como benefícios fiscais para comportamentos que são vistos como socialmente benéficos.

Os algoritmos de aprendizado de máquina usam até o mesmo termo “recompensa”[139] para treinar os modelos de IA a fazer previsões ou tomar decisões precisas. Entretanto, não se trata do mesmo tipo de recompensa que opera na sociedade humana. A “coerção” no contexto da IA é ainda menos clara, já que a IA não pode ser “coagida” no

[139] Os algoritmos de aprendizado de máquina que usam mecanismos de recompensa são conhecidos como algoritmos de Aprendizado por Reforço (*Reinforcement Learning).* O aprendizado por reforço é uma subárea do aprendizado de máquina inspirada na forma como os seres humanos e animais em geral aprendem a partir de interações com seu ambiente. Ressaltamos que apesar do uso da palavra recompensa, trata-se de naturezas distintas.

mesmo sentido que um ser humano. A coerção provavelmente tem valor no estágio atual de desenvolvimento da IA regulando o comportamento dos programadores e empresas responsáveis pelo desenvolvimento.

> A manutenção da sociedade civil depende da justiça, e esta depende do poder de vida ou morte, bem como de outras recompensas e punições menores que são inerentes àqueles que detêm a soberania do Estado. É impossível que um Estado tenha permanência se qualquer outro, além do soberano, tiver o poder de conceder recompensas maiores do que a vida ou de infligir punições maiores do que a morte. (MORRISON apud HOBBES. 2012. P. 115)

Em questão de coerção, ainda temos duas abordagens distintas, o utilitarismo e a teoria da retribuição. O utilitarismo argumenta que a punição só é justificada se produzir um benefício maior para a sociedade e pode ser justificada se dissuadir outras pessoas de cometer crimes ou se ajudar a reabilitar o criminoso, levando a um benefício para a sociedade. A coerção, neste contexto, é vista como um meio para um fim e é avaliada com base em suas consequências.

A teoria da retribuição, por outro lado, argumenta que as pessoas devem ser punidas por seus crimes porque elas merecem ser punidas. Segundo essa visão, a punição é justificada não por suas consequências, mas porque é a resposta correta ao mal que o criminoso cometeu. O princípio fundamental da teoria da retribuição é que a punição deve ser proporcional ao crime. Assim, a coerção, na forma de punição, é vista como um fim em si mesma, uma expressão da justiça. Se para qualquer ser humano “As recompensas maiores do que a vida ou de infligir punições maiores do que a morte” parecem critérios últimos e bem claros, o mesmo não se aplica à IA. O que levanta a questão de como esperar qualquer cumprimento autônomo de lei, não importa se a abordagem seja utilitarista ou retributiva.

Por último, no sistema jurídico existe o princípio da individualização da pena.[140] Para cada crime tem-se uma pena que varia de acordo com a personalidade do agente. A pena deve ser individualizada nos planos legislativo, judiciário e executório, evitando-se a padronização da sanção penal. O princípio da individualização da pena é um conceito fundamental no direito penal que sugere que as penas devem ser adaptadas às circunstâncias individuais de cada criminoso e ao crime que cometeu. Por outro lado, o conceito de

[140] Disponível em <https://www.mpf.mp.br/es/sala-de-imprensa/glossario-de-termos-juridicos> Acesso em outubro de 2022.

alteridade é fundamental para muitas áreas da filosofia. A alteridade ontológica refere-se à diferença ou distinção entre seres ou entidades. A alteridade ética ou moral envolve a maneira como percebemos e nos relacionamos com os outros. O filósofo francês Emmanuel Levinas argumentou que a experiência da alteridade do outro é a base de nossa responsabilidade ética.

Sabemos que os PIA não são limitadas a robôs, eles se aplicam à IA. Mesmo que uma infração tenha sido cometida por uma unidade de robô, o seu sistema poderia estar em nuvem compartilhado entre milhares de outros. Penalizar uma unidade não teria validade, então deve penalizar todo sistema de IA para cada infração que surgir? Devemos considerar a IA como uma unidade ou sempre na sua totalidade[141]? Ainda não temos respostas claras para estas e outras questões a partir da formulação inédita de [R a H].

[141] O aprendizado federado (*Federated Learning*) é uma abordagem de aprendizado de máquina que permite treinar um algoritmo em múltiplos dispositivos ou servidores que mantêm seus próprios conjuntos de dados locais. Em vez de enviar os dados para um servidor central, os dispositivos fazem o treinamento localmente e compartilham apenas os parâmetros do modelo ou atualizações para esses parâmetros com o servidor central. Assim, o servidor central combina essas várias atualizações para formar um modelo global que é então enviado de volta para todos os dispositivos. Como os dados brutos nunca deixam o dispositivo local, é menos provável que informações sensíveis sejam expostas. O aprendizado federado pode ser útil para setores ou aplicações onde a privacidade dos dados é uma grande preocupação, como saúde ou serviços financeiros. O aprendizado federado é uma técnica promissora que pode ajudar a equilibrar a necessidade de aprendizado de máquina eficaz com considerações de privacidade e talvez um caminho para trazer mudanças na questão de unidade e totalidade da IA.

PARTE II CAPÍTULO 3 – REALIZABILIDADE DOS PIA [R a H] → [R = H]

3.1 - Realizabilidade dos PIA [R a H] → [R = H]

"As máquinas podem pensar?" Turing escreveu o seu famoso texto "*Computing Machinery and Intelligence"* publicado pela revista *Mind* na edição de outubro de 1950, iniciando com esta pergunta.

> Proponho considerar a questão: "As máquinas podem pensar?" Isso deve começar com definições do significado dos termos "máquina" e "pensar". As definições podem ser formuladas de forma a refletir, tanto quanto possível, o uso normal das palavras, mas essa atitude é perigosa. […] Em vez de tentar tal definição, substituirei a pergunta por outra, que está intimamente relacionada a ela e é expressa em palavras relativamente inequívocas. A nova forma do problema pode ser descrita em termos de um jogo que chamamos de 'jogo da imitação'.[142] (TURING. 1950. p.1)

Após toda exposição sobre o jogo de imitação que mais tarde foi popularizado como o teste de Turing, conclui o seu artigo prevendo que em 50 anos, ninguém acharia estranho que os robôs pudessem ser capazes de conversar com seres humanos como se estivessem pensando.

> Eu acredito que, em cerca de cinquenta anos, será possível programar computadores, com uma capacidade de armazenamento de cerca de 10^9, para fazê-los jogar o jogo da imitação tão bem que um interrogador médio não terá mais do que 70 por cento de chance de fazer a identificação correta após cinco minutos de questionamento. A questão original, "As máquinas podem pensar?" Eu acredito ser demasiado sem sentido para merecer discussão. No entanto, acredito que no final do século o uso de palavras e a opinião educada geral terão mudado tanto que será possível falar de máquinas pensando sem esperar ser contradito.[143]

Passaram-se mais de setenta anos desde a previsão de Turing. Se revisitarmos a sua pergunta original, podemos perceber o quão surpreendente Turing foi em sua visão, feita ainda atrelada à tecnologia da época, mas a semente da ideia da IA já estava germinando. Com o surgimento e a popularização de ferramentas de IA como o chatbot GPT, a ideia de "máquinas pensantes" não só se tornou aceitável, mas também uma realidade tangível em nossas vidas diárias.

[142] No Congresso Principia de 2019 apresentamos "O jogo de imitação de Turing como jogo de linguagem de Wittgenstein" que analisa a influência de Wittgenstein sobre o texto de Turing.
[143] TURING. Alan. Computing Machinery and intelligence. Mind. VOL. LIX. NO. 236. Reino Unido. 1950.

Ao interagir com um chatbot avançado, como o GPT, os usuários têm a impressão de conversar com um ser pensante. Essa sensação não se deve apenas à capacidade da máquina de gerar respostas coerentes e contextuais, mas também por aprender com interações passadas e aplicar esse conhecimento em futuras interações, simulando uma forma de pensamento e aprendizado como foi previsto por Turing. Hoje, de fato, o uso de palavras e a opinião educada geral estão mudando de tal maneira que podemos até falar de máquinas pensantes sem esperar sermos contraditos veementemente. Ao longo das últimas sete décadas, as máquinas, de uma maneira que seria quase inimaginável em 1950, aparentam "pensar". E essa evolução não mostra sinais de abrandamento, levando-nos a um futuro onde a relação entre humanos e máquinas continuará a ser redefinida.

Se a questão proposta por Turing tivesse uma sequência atual, acreditamos que poderia ser "em que medida as máquinas podem cumprir leis?" ou "como as máquinas podem cumprir leis?". A primeira das “Oito leis da robótica” de Shinpo Fumio - Keio University (2015. Japão), que segundo o autor, expressa a máxima “A humanidade em primeiro lugar” repete a primeira lei de Asimov - “os robôs não podem prejudicar [as] pessoas”, mas a segunda parte introduz uma outra objeção, “[não podem] se tornar pessoas”. A essa objeção, é preciso contrapor a seguinte consideração: A primeira parte desta lei, junto com a imensa maioria dos PIA publicados até hoje, só poderiam eventualmente ter condições de funcionamento, se somente [R] se assemelhasse a [H].

Fazemos uma ressalva sobre a condição de “se tornar pessoas”. Não argumentamos aqui uma possibilidade de transformação plena de [R] em pessoas, mas para que a fórmula geral dos PIA expressa em [R a H] tenha validade, [R = H] é uma condição fundamental. A condição de funcionamento para PIA introduzida pelo condicional [→] significa que [R] deve ser como [H] como agente de direito, pela equivalência simbolizada por [=].

[R a H] → [R = H]

Condição fundamental para PIA

No capítulo anterior verificamos a agência e a realizabilidade de [R] e a realizabilidade para [a]. Acreditamos que mesmo que talvez não tenha sido plenamente exposto, oferecemos argumentos razoáveis para questionarmos a possibilidade de aplicação atual e real dos PIA e, provavelmente, sem a condição acima, mesmo para algum futuro. A lei é um mecanismo intrinsecamente humano, em todas as suas etapas, pelo menos até hoje. Para fazer parte deste jogo de linguagem [144] se fizermos uma alusão ao conceito de Wittgenstein, seus participantes precisam dominar essas regras, junto com a aceitação de outros participantes.

> Nada pode ser injusto. As noções de certo e errado, de justiça e injustiça, não têm aí lugar algum. Onde não há poder comum não há lei, e onde não há lei não pode haver injustiça, a guerra, a força e a fraude são as duas virtudes cardeais. A justiça e a injustiça não são inerentes às faculdades do corpo ou da mente. Se o fossem, poderiam coexistir em um homem que estivesse sozinho no mundo, assim como seus sentidos e suas paixões. São qualidades que dizem respeito ao homem em sociedade, não em solidão. (MORRISON apud HOBBES. 2012. p. 108)

Morrison (2012) comenta esta passagem de Leviatã e afirma que "todas as pretensões ao direito e todas as reivindicações de justiça são humanas, são sociais" que são consequências da teoria de Hobbes. Uma sociedade formada por [R] e [H] não escaparia desta mesma condição, mas não se deve pensar tal limitação como a única. O filósofo francês Francis Wolff em sua obra "*Três Utopias Contemporâneas",* que analisa a utopia pós-humanista e animalista, faz a seguinte ressalva sobre a possibilidade de formação de uma comunidade de seres humanos com outros seres.

> Só há justiça no interior de uma comunidade de igual, […] nós somos animais. Mas não formamos uma comunidade moral com os animais na qual possa reinar a justiça. […] Nós somos animais sensíveis. Existem outros. Nem por isso há uma comunidade moral de seres sensíveis. […] É necessário ser humano para pensar "nós" […] Nós somos iguais *enquanto* nós. (WOLFF. 2018. p. 64~66)

Para Wolff (2018), as duas vertentes de utopia repousam sobre uma ética da segunda pessoa, focada em como evitar o mal aos outros. A ética da segunda não seria justa, pois a justiça só existe dentro de uma comunidade de iguais, e não formamos uma comunidade de iguais com os animais. O autor argumenta que, embora façamos parte da espécie biológica Homo sapiens e, portanto, sejamos animais, não formamos uma comunidade moral com os animais. Existe uma diferença fundamental entre os interesses humanos e os animais, sendo

[144] Que certamente vai além do uso de linguagem.

que a vida de um muitas vezes depende da vida do outro. Wolff (2018) reconhece que temos deveres para com os animais, mas são deveres relativos, não absolutos.

> Apesar de pertencer à espécie biológica Homo sapiens, que é incontestavelmente uma espécie animal, as pessoas não são animais. As pessoas formam uma comunidade moral de direitos e deveres recíprocos e absolutos. A ética dessa comunidade é de terceira pessoa: é uma comunidade de iguais. Pode haver justiça nessa comunidade: para iguais, direitos iguais, partes iguais ou proporcionais. Apesar de dispensarmos unilateralmente nosso cuidado (*care* e *cure*) a bebês, crianças e pessoas em situações de deficiência ou dependência, eles não são pacientes morais, mas membros de nossa comunidade moral. (WOLFF. 2018. P. 64~66)

É possível verificar esta ideia de origem kantiana também no texto da filósofa espanhola Adela Cortina (2009) na sua obra "Ética Mínima".

> A resposta kantiana na qual a humanidade não renunciou, é bem conhecida contra o utilitarismo, que defende a satisfação das aspirações de toda a criação senciente, é preciso lembrar que a sobrevivência de alguns seres existe irremediavelmente o sacrifício de outros; que existe apenas um ser cuja autonomia é fundamental de deveres universalmente exigíveis: só as pessoas, por força de sua autonomia, têm de ser universalmente respeitadas e assistidas em seu anseio de felicidade. (CORTINA. 2009. p.47)

A relação com os outros na sociedade moderna é frequentemente mediada por acordos e contratos - seja no sentido literal de contratos legais, ou mais figurativamente em termos de normas sociais e expectativas. Isso está em linha com as teorias políticas liberais clássicas, que veem a sociedade como um conjunto de indivíduos que entram em contratos sociais para estabelecer uma comunidade. A solidariedade, neste caso, é vista como uma questão de seguir as regras de conveniência que foram acordadas por esses contratos sociais.

No entanto, esta visão de sociedade e do indivíduo é insuficiente. Ela sugere que o indivíduo está organicamente vinculado à comunidade e que o sentido de uma existência individual não pode ser separado do contexto comunitário que o produz e o sustenta. Isso significa que vidas individuais não podem ser entendidas separadamente da sociedade em que vive, e que a sociedade não é apenas um conjunto de contratos celebrados, mas uma parte essencial de realidade. Essa visão implica que a existência individual autossuficiente é uma abstração e todos são profundamente influenciados e moldados por comunidades.

A partir das considerações de Morrison, Wolff e Cortina, se é possível somente formar leis em sociedade, promover justiça em uma sociedade que se constitui como mesma

comunidade moral, e assim poderia funcionar uma ética da justiça (terceira pessoa), o caminho para uma convivência com a IA na forma expressa em PIA segue apenas pela formação de uma comunidade de iguais, ou seja, [R = H].[145] Pois somente poderia haver justiça entre iguais. As leis são instrumentos que funcionaram apenas para seres humanos em sociedade. De que forma os [R] pode fazer parte da sociedade como membro que poderia usufruir de direitos e deveres?

Se podemos verificar antropomorfização das máquinas desde os primórdios do seu surgimento, chegamos em um momento que precisamos discutir a "antropomorfização legal", isto é, se é possível atribuir "humanidade" para que [R] possa formar uma comunidade moral com os seres humanos. Acreditamos que [R = H] apresenta duas vias: humanização das máquinas expressa por [R = H], e maquinização dos seres humanos expressa por [H = R], que veremos nas próximas seções.

[R a H]	→	**[R = H]**
		[H = R]

Duas alternativas da condição fundamental para PIA[146]

3.2 - Humanização das máquinas [R = H]

As máquinas, autômatos e robôs têm uma longa história, com raízes em todas as épocas da civilização humana, muito antes do advento dos computadores modernos. Desde o seu nascimento, eles eram usados para imitar a vida e a ação em cerimônias religiosas ou para entretenimento. Na Grécia Antiga, por exemplo, havia autômatos mecânicos usados para demonstrar princípios da física e da engenharia. O engenheiro Heron de Alexandria[147] é conhecido por seus inúmeros autômatos, incluindo um teatro mecânico capaz de apresentar uma peça de 10 minutos. Ele foi inventor de máquinas como a dioptra, o odômetro (sistema de engrenagens combinadas para contar as voltas dadas por uma roda) ou, talvez o mais importante, a eolipila, um precursor da turbina a vapor.

[145] Na simbolização adotada neste trabalho, os colchetes [ ] representam uma mesma comunidade.
[146] Adotamos o símbolo de igualdade = que nestes casos carregam ideia de direção: [R] se aproxima de [H] e inverso.
[147] Disponível em <http://ecalculo.if.usp.br/historia/heron.htm> Acesso em outubro de 2022.

Durante a Idade Média e o Renascimento, encontramos relógios astronômicos altamente complexos que também podem ser classificados como autômatos, incluindo o Relógio Astronômico de Praga [148]e outros semelhantes em toda a Europa. Em 1642, Pascal inventou a primeira calculadora mecânica conhecida como a Pascalina. A Pascalina era uma engenhosa peça de maquinaria que fazia somas e subtração. A Pascalina é vista como um dos primeiros passos na jornada que nos levaria aos computadores modernos. Algumas décadas depois, em 1672, Leibniz melhorou a Pascalina ao criar uma máquina que poderia não apenas somar e subtrair, mas também multiplicar, dividir e calcular raízes quadradas. A Calculadora de Passos de Leibniz foi outra etapa fundamental no desenvolvimento dos autômatos e das máquinas de cálculo.

No início do século XIX, Joseph Marie Jacquard inventou um dispositivo que poderia ser visto como um precursor dos computadores programáveis - o tear de Jacquard.[149] Este tear usava cartões perfurados para controlar os padrões complexos que poderiam ser tecidos automaticamente. O sistema de cartões perfurados inspirou futuras inovações, incluindo a Máquina Analítica de Charles Babbage. No século XX, esses desenvolvimentos culminaram na invenção dos computadores eletrônicos, que levaram ao surgimento dos primeiros robôs verdadeiros. Estes primeiros robôs eram geralmente máquinas simples, controladas por um conjunto limitado de comandos pré-programados.

Podemos verificar inspiração antropomórfica das máquinas em muitos casos. Wiener estudou o "piloto" da máquina a vapor de James Watt, que regulava automaticamente a velocidade. Wiener viu uma analogia entre o regulador centrífugo e a maneira como os seres humanos controlam suas próprias atividades. Ele percebeu que para os computadores serem desenvolvidos, teriam que se assemelhar à habilidade dos seres humanos no controle de suas próprias atividades. Ele acreditava que essa era uma das principais características que separavam os humanos das máquinas da época e, se ao incorporar esse princípio de auto regulação em máquinas, poderia criar computadores muito

[148] IMBROISI, Margaret; MARTINS, Simone. O relógio astronômico de Praga. História das Artes, 2023. Disponível em <https://www.historiadasartes.com/sala-dos-professores/o-relogio-astronomico-de-praga/> Acesso em outubro de 2022.
[149] Disponível em <https://www.computerhistory.org/storageengine/punched-cards-control-jacquard-loom/> Acesso em outubro de 2022. Tradução nossa.

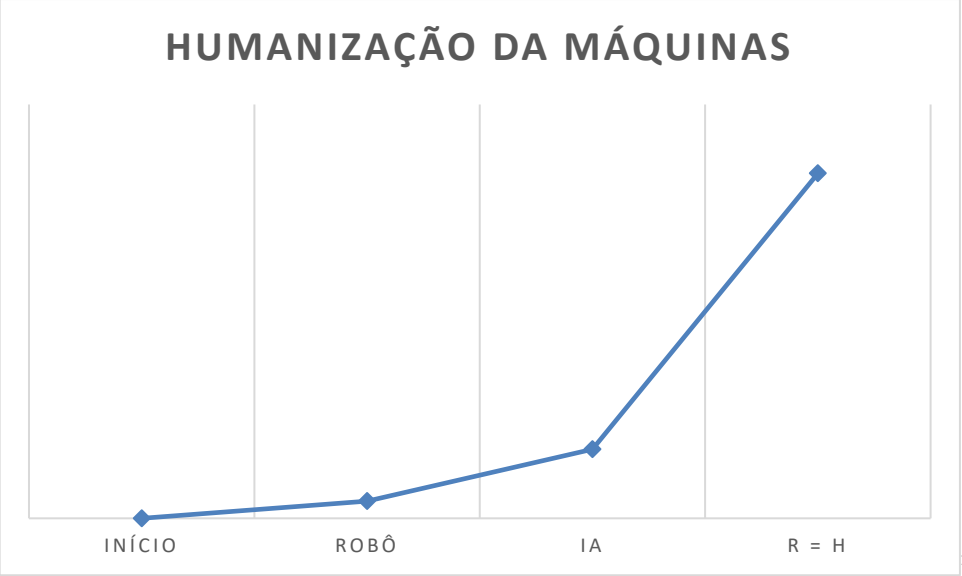
mais avançados. As ideias de antropomorfização de Wiener foram fundamentais para o desenvolvimento da cibernética e influenciaram profundamente a maneira como pensamos sobre os sistemas de computador e a inteligência artificial hoje.

Outro exemplo de antropomorfização seria a arquitetura de computador proposta por John von Neumann, também conhecida como Modelo de von Neumann ou Arquitetura de von Neumann,[150] fundamental para o desenvolvimento dos computadores moderno. Este modelo, desenvolvido por von Neumann no final dos anos 1940, tornou-se a base para a maioria dos computadores modernos. Na Arquitetura de von Neumann, um computador consiste em quatro partes principais: Unidade de processamento central (CPU): este é o "cérebro" do computador, onde todas as operações lógicas e aritméticas são realizadas; memória: este componente armazena todos os dados e instruções do programa; dispositivos de entrada e saída: estes são usados para comunicar com o mundo exterior; barramento: este é o sistema que conecta a CPU, a memória e os dispositivos de entrada e saída. Ele permite a transferência de dados e instruções entre as várias partes do computador.

Desde Turing, a possibilidade de aparição de robôs com características humanas tornou-se um assunto largamente difundido. Conversar como ser humano, realizar tarefas, jogar xadrez, não apenas jogar, mas superar o ser humano. Todas estas realizações parecem aproximar [R] de [H]. A ideia por trás do teste de Turing é que se uma máquina for capaz de imitar o comportamento humano de forma convincente, então é provável que ela tenha uma inteligência comparável à humana. O teste de Turing é frequentemente citado como um marco importante na história da IA, pois foi um dos primeiros esforços para definir e avaliar a inteligência de uma máquina. No entanto, alguns argumentam que a capacidade de uma máquina de imitar o comportamento humano não é suficiente para provar que ela é inteligente. Outros apontam que o teste é limitado em sua capacidade de avaliar a verdadeira inteligência, já que se concentra principalmente em habilidades linguísticas e não leva em conta outras habilidades.

[150] Disponível em <https://www.computerscience.gcse.guru/theory/von-neumann-architecture> Acesso em outubro de 2022. Tradução nossa.

Neste trabalho, o desenvolvimento e a humanização das máquinas têm um limite ou meta que é a possibilidade de seguir as leis dentro de uma mesma comunidade moral com os seres humanos. Portanto, nesta direção da humanização deve haver o "aperfeiçoamento" contínuo das máquinas a ponto de atingir o ponto [R = H] com possibilidade de cumprir leis.



[151]

Gráfico 2 - Humanização das máquinas

De junho de 2013 a julho de 2014, um experimento[152] utilizando um robô humanoide do tamanho de um humano foi realizado em um shopping center no Japão. Nesse período, os pesquisadores observaram os visitantes do shopping por durante treze dias. Durante nove dias, abusos em relação ao robô foram detectados como obstrução de movimentação do robô, uso de linguagem abusiva, chutes, socos, danos nos braços e as articulações da cabeça e do braço do robô. Os agressores eram crianças de idades variadas.

Os pesquisadores tentaram entender por que as crianças maltratavam os robôs e se os motivos podem ser semelhantes aos que levam ao bullying entre crianças ou abusos contra animais, como a busca por dominância ou afiliação a grupos. Os pesquisadores conduziram entrevistas com crianças que maltrataram o robô. Na análise quantitativa, a maioria das crianças (74%) percebeu o robô como semelhante a um humano, enquanto algumas (13%) o perceberam como uma máquina. As razões para os comportamentos abusivos incluíam

[151] Este e o próximo gráfico são arbitrários e meramente ilustrativos.
[152] Nomura, Tatsuya ; Kanda, Takayuki ; Kidokoro, Hiroyoshi ; Suehiro, Yoshitaka & Yamada, Sachie (2016). Why do children abuse robots? Latest Issue of Interaction Studies 17 (3):347-369.

curiosidade (22%), diversão (35%) e influência de outras crianças que maltratavam o robô (17%).

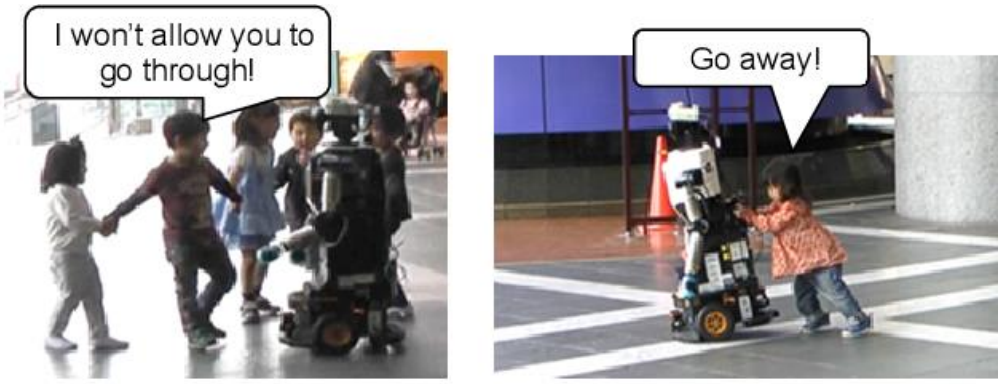


Fig.11 - Abuso de robôs [153]

Os pesquisadores concluem que, embora a humanização dos robôs possa parecer uma solução, ela pode não ser suficiente para moderar o abuso, sendo necessário explorar maneiras de estimular a empatia das crianças pelos robôs. Além disso, ressaltam que algumas crianças que maltratam robôs não necessariamente o fazem por falta de empatia, mas sim porque veem os robôs mais como máquinas do que seres humanos.

Mesmo levando em consideração as devidas diferenças, podemos verificar pelo experimento acima que a humanização das máquinas não é um fator isolado centrada apenas no desenvolvimento das próprias máquinas. Envolve o outro aspecto que enfatizamos anteriormente que é a aceitação das máquinas como parte da mesma comunidade moral pelos seres humanos. Assim, o ponto [R = H] não significa apenas [R] atingir um nível de aperfeiçoamento, mas também uma transformação de [H].

Este aspecto pode ser explicado por conceitos sociais como valência e emergência. A valência social é um termo que se refere à atração ou repulsão que as pessoas sentem em relação a outras pessoas, objetos em um contexto social. A valência social influencia as interações sociais, as emoções e o comportamento das pessoas em diferentes situações e pode ser afetada por fatores como experiências passadas, crenças e valores culturais. A

[153] Disponível em <https://www.semanticscholar.org/paper/Escaping-from-Children%E2%80%99s-Abuse-of-Social-Robots-Brscic-Kidokoro/bc5252525dd0c29324f4e45f274711f1b3665b0b> Acesso em outubro de 2022.

emergência é um conceito que se refere à maneira como propriedades, comportamentos ou padrões complexos surgem de uma multiplicidade de interações.

> A previsibilidade opera como um limite à responsabilidade pelos próprios atos: raramente se é responsável pelo que não se pode prever. Quanto ao controle, é relevante determinar a responsabilidade indireta. A previsibilidade e o controle determinam quando as pessoas se tornam responsáveis por danos causados por seus agentes, crianças, animais e objetos perigosos. Para determinar a responsabilidade de alguém pelo uso de um robô simples e previsível, devemos examinar a previsibilidade. No entanto, para determinar a responsabilidade de alguém pelas ações de uma IA avançada que possui agência, o nível de controle sobre a IA será mais relevante. À medida que a emergência avança, tais delitos afetarão os incentivos legais enfrentados pelos projetistas de robôs e IA apenas na medida em que eles possam exercer controle. O controle, portanto, determinará a utilidade dos incentivos legais. (COFONE. 2018. p. 185)

Segundo Ignacio Cofone - professor associado e pesquisador em Lei de Inteligência Artificial e Governança de Dados na Faculdade de Direito da Universidade McGill, no seu artigo "*Servers and Waiters"* (2018), existem duas linhas independentes de emergência: imprevisibilidade e agência. A lei avalia a previsibilidade (e controle) para determinar a responsabilidade direta e indireta pelas ações de uma entidade.

3.3 - Maquinização dos seres humanos [H = R]

> Até quando um humano, que acolhe em sua anatomia receptores e biossensores, próteses externas ou invasivas, implantes intradérmicos ou intracerebrais, deverá ser considerado ainda como um humano creditado de seu livre-arbítrio? É a situação descrita pelo filme de José Padilha, Robocop (BESNIER e LAURENT. 2022. p. 52)

Da mesma forma demonstrada na última seção, a maquinização do ser humano[154] aqui tem um limite que é a continuidade em seguir as leis. Portanto, nesta direção da maquinização, [H] não pode perder a capacidade de cumprir as leis e deixar a comunidade moral da qual fazia parte. Se [R] precisa evoluir para atingir o ponto capaz de cumprir leis em comunidade moral, a direção que [H] toma é inverso, não perder a capacidade de cumprir leis em sociedade. Enxergamos que esta perda na capacidade de agência legal por [H] não seria uma condição necessária, mas uma alternativa plausível que não poderia ser ignorada.

[154] Apesar de não ser atual e possuir abrangência imediata para a IA, optamos pelo uso do termo 'maquinização' pelas referências já existentes, em vez de por exemplo 'ciborguização'.

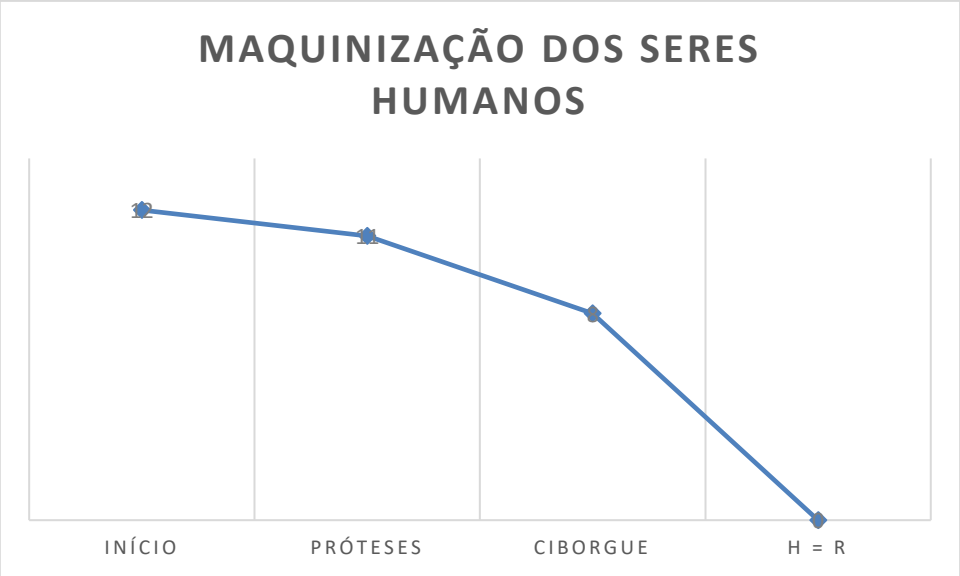


Gráfico 3 - Maquinização dos seres humanos

O termo "ciborgue" foi criado durante a era espacial, mas a ideia de combinar seres humanos e máquinas é muito mais antiga. A história do ciborgue é uma história de transformação, abrangendo mitologia, ficção científica, medicina e filosofia. Desde a antiguidade, existem histórias de seres híbridos de humano e máquina. No folclore judaico, há a lenda do Golem, uma criatura feita de barro e animada por magia.

A invenção do termo "ciborgue" em si é muito mais recente. Foi cunhada pelos cientistas espaciais Manfred Clynes e Nathan Kline em 1960 para descrever um ser humano aprimorado que poderia sobreviver em ambientes extraterrestres. O termo é uma junção de cibernético cunhado por Wiener com organismo, enfatizando a combinação de elementos orgânicos e sintéticos.

> Quais são alguns dos dispositivos necessários para criar sistemas homem-máquina autorregulados? Essa autorregulação deve funcionar sem o benefício da consciência para cooperar com os próprios controles homeostáticos autônomos do corpo. Para o complexo organizacional exogenamente estendido funcionando como um sistema homeostático integrado inconscientemente, propomos o termo “Cyborg”. O Cyborg incorpora deliberadamente componentes exógenos estendendo a função de controle autorregulador do organismo para adaptá-lo a novos ambientes. (CLYNES, KLINE. 1960. p. 27)[155]

[155] Disponível em <https://web.mit.edu/digitalapollo/Documents/Chapter1/cyborgs.pdf> acesso em outubro de 2022. Tradução nossa.

Desde então, a ideia de ciborgues tem sido explorada por muitos pensadores e escritores. Um exemplo é Donna Haraway que usou o conceito de ciborgue como uma metáfora poderosa em seu "Manifesto Ciborgue"[156]de 1985. Haraway argumentou que todos somos ciborgues até certo ponto, já que nossas vidas são cada vez mais entrelaçadas com a tecnologia. Ela usou essa ideia para questionar as fronteiras tradicionais entre humano e máquina, natureza e cultura, masculino e feminino.

Em paralelo à discussão filosófica e social, o desenvolvimento da tecnologia aproximou a realidade do conceito de ciborgue. A medicina moderna utiliza próteses cada vez mais sofisticadas, implantes que permitem aos surdos ouvir e até mesmo interfaces cérebro-máquina que podem permitir aos paralisados mover objetos com o pensamento. A ideia de ciborgues, desde suas raízes na mitologia antiga até os debates filosóficos contemporâneos e a medicina avançada, reflete nossa constante interrogação sobre o que significa ser humano em uma era de rápidas mudanças tecnológicas. O filósofo francês Jean-Michel Besnier e o médico francês Laurent Alexandre no livro "*Os robôs fazem amor? O Trashumanismo em Doze Questões",* descrevem o horizonte de ciborgue como ser um humano ampliado.

> O que importa, com efeito, para que haja ciborgue é que o papel desempenhado pelo artefato não esteja submetido ao corpo vivente como uma simples ferramenta destinada a prolonga-lo, mas que ele acompanhe a manifestação para realizar performances não naturais. É por isso que "ciborgue" é outro nome para falar do humano ampliado, que não é humano concertado e/ou prolongado por adjuvantes tecnológicos, mas o humano transfigurado ao qual a engenharia genética irá adicionar faculdades sensoriais ou de competências que não existem na natureza humana (por exemplo, a audição dos morcegos ou a sensibilidade às ondas elétricas do tubarão). Ser um ciborgue é, de todo modo, ultrapassar o formato humano, ao se encontrar acoplado a artefatos (biomiméticos ou puramente maquínicos, máquinas), tendo em vista realizar performances que não mais serão façanhas humanas. (BESNIER e LAURENT. 2022. p. 53)

Segundo os autores a maquinização dos seres humanos ou ciborguização já começou. Nesta fase inicial, ocorre, por exemplo, a manipulação genética com a possibilidade de pais escolherem as características dos seus filhos de acordo com o diagnóstico genético pré-natal, onde os "perigos" são eliminados antecipadamente. Segundo Laurent Alexandre, um exemplo disso seria a trissomia 21 (síndrome de Down) que está em vias de desaparecer. Uma

[156] Disponível em <https://ea.fflch.usp.br/obra/manifesto-ciborgue> Acesso em outubro de 2022.

vez que 97% dos trissômicos se "beneficiam" com uma interrupção médica da gravidez. Alexandre argumenta que poucos pais conseguem resistir à pressão social para "erradicar" esse *handcap* e, em breve, será oferecido aos pais o "sonho" de uma criança configurada *à la carte.* Alexandre afirma que *"*nós já temos um tobogã eugenista sem nos darmos conta".

> A humanidade foi lançada em um tobogã transgressivo. Devemos nos tornar, sem estarmos cientes disso, transhumanos, isto é, homens e mulheres tecnologicamente modificados. Até 2050, choque biotecnológicos ainda mais espetaculares vão sacudir a sociedade: regeneração de órgãos por células-tronco, terapias gênicas, implantes cerebrais, técnicas antienvelhecimento, projeto genético de bebê *à la carte*, fabricação de óvulos a partir de células da pele… (BESNIER e LAURENT. 2022. p. 45)

Os autores alertam sobre o perigo do eugenismo atrelado ao processo de maquinização dos seres humanos. Alexandre enfatiza que o próprio termo "transhumanismo" foi inventado em 1957 por Julian Huxley, um eugenista, irmão de Aldous - autor de Admirável Mundo Novo. Os autores argumentam que o mundo totalitário onde o Estado se arroga o direito de selecionar os bebês destinados a viver, de lhes atribuir uma casta em função do seu potencial biológico, descrito no romance de Aldous Huxley, pode tornar-se realidade em um curto tempo.

> As neurotecnologias são literalmente revolucionárias na medida que elas abalam a ordem social. Podemos escapar delas? Será possível uma "contraaneurorevolução"? Provavelmente não. Afinal, um ser humano que se recusasse a ser hibridado com circuitos eletrônicos quase não seria competitivo no mercado de trabalho. Imaginemos uma sociedade com duas velocidades, os humanos não ampliados se tornariam inevitavelmente párias? Além disso, seria ético não aumentar as capacidades cognitivas de pessoas pouco dotadas? Na era das próteses cerebrais, o risco da neuromanipulação, do neurohacking e, portanto, da neuroditadura é imenso. Devemos enquadrar o poder dos neurorevolucionários: o controle e domínio do nosso cérebro irá se tornar o primeiro dos direitos do ser humano. (BESNIER e LAURENT. 2021. p. 95)

Conforme apresentamos, temos um gráfico que ilustra a direção que os robôs e a IA tomam para humanização e, um outro com seres humanos em direção da maquinização. Imaginamos que [R] de partida certamente é distinta de [R] de chegada pelo desenvolvimento contínuo das máquinas. Em outras palavras, [R] em dois casos não são idênticas. Entretanto, imaginamos intuitivamente que [H] seriam idênticos em ambos os casos. Mas diante do processo de maquinização que já se iniciou, inevitavelmente surge o questionamento, qual [H] que [R] deve ter como alvo, uma vez que a maquinização do ser humano poderia alterar o próprio modelo do ser humano?

Humanização das máquinas	Maquinização dos seres humanos
[R = H]	[H = R]

[H] seria mesmo em duas situações?

Depois de dois movimentos analisados: humanização das máquinas, maquinização dos seres humanos, enfatizamos um outro aspecto. Apesar destes dois movimentos tratarem a transformação interna centrada nos indivíduos ou ainda unidades, não são únicas. Há também uma transformação social conforme a transformação dos indivíduos que a compõe: "maquinização da sociedade".

3.4 - Expectativas de H[R a H]

Na seção 2.4 do Capítulo 2 da Parte II, agrupamos os PIA em [R a H] por assuntos: "Respeito à dignidade humanas", "Benefício e prosperidade para a humanidade" e "Transparência e confiabilidade". Em seguida, classificamos em três grupos a partir da possibilidade de realização: "Possíveis hoje", "Talvez no futuro" e "Improváveis mesmo no futuro". Nesta seção, analisaremos mais a fundo o que os PIA classificados como anteriormente ainda podem nos revelam. A simbolização H[R a H] pode ser entendida como PIA com destaque para o papel do legislador, neste caso, seres humanos [H] que atua como legislador expressando os seus desejos através de [R a H].

> Lei Zero: Um robô não pode prejudicar a humanidade ou, por inação, permitir que a humanidade sofra algum mal. (Asimov)
>
> Os robôs da próxima geração contribuirão para a realização de uma sociedade segura e pacífica. (Declaração Mundial de Robôs de Fukuoka)
>
> Procuraremos garantir que as tecnologias de IA beneficiem e capacitem o maior número possível de pessoas. (Princípios de parceria em IA)
>
> A.I. deve ser projetado para ajudar a humanidade, o que significa que a autonomia humana precisa ser respeitada. (Leis de Satya Nadella)
>
> O único objetivo do robô é maximizar a realização dos valores humanos. (Stuart Russel)
>
> A IA não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. (Princípios para a Governança da IA - The Future Society)
>
> Faça a IA servir as pessoas e o planeta. (10 Princípios para IA Ética - UNI Global Union)

Os sete princípios acima selecionados representam diversas iniciativas de épocas distintas. De acordo com a classificação oferecida no Capítulo 1 da Parte I, misturam os PIA da Primeira Fase Tradicional e Segunda Fase Atual com protagonismos de robôs e IA. Apesar das diferenças, acreditamos que revelam as expectativas expressas em relação às máquinas que podemos classificar como utópicas, porque expressam aspirações ideais para um mundo onde os robôs e a IA operam de maneira segura, benéfica e justa para todos. Mas a sociedade humana apresenta uma diversidade enorme de valores e crenças de seus membros, muitas vezes contextuais e contraditórias. A ideia de que a IA poderia maximizar a realização dos valores humanos, por exemplo, presume que esses valores poderiam ser claramente definidos e codificados.

A IA e robôs estão ainda em estágios de desenvolvimento, mas mesmo que a transformação seja constante e rápida, não podemos afirmar que um dia poderiam entender completamente e se adaptar aos complexos contextos sociais e éticos em que operam. Poderiam funcionar de maneira ideal em teoria. Na prática, há uma série de desafios de governança e controle. Quem irá definir os padrões? Quem irá implementar e garantir a conformidade? Como serão resolvidos os conflitos de valores entre diferentes culturas e sociedades? Junto com outros questionamentos levantados pelos cientistas da Federação de Cientistas da Alemanha no Capítulo 3 da Parte I.

A disseminação de IA e robôs pode levar a grandes mudanças na sociedade, incluindo deslocamento de empregos e acentuação das desigualdades socioeconômicas. A garantia de que a IA beneficie o maior número possível de pessoas e não prejudique a igualdade de direitos é um objetivo ideal, mas atingi-lo na prática pode ser um desafio quase impossível. Assim, embora esses princípios estabeleçam metas louváveis para o desenvolvimento de IA e robôs, a realização desses ideais na prática é complexo e incerto, o que leva a sua classificação como ingênuos e utópicos.

> Então **Ur-Nammu**, o poderoso guerreiro, rei de Ur, rei da Suméria e Akkad, pelo poder de Nanna, senhor da cidade, e de acordo com a verdadeira palavra de Utu, estabeleceu equidade na terra; ele baniu a maldição, a violência e conflitos. (Prólogo Ur-Nammu)
>
> Então **Lipit-Ishtar**, o sábio pastor cujo nome foi pronunciado Nunamnir, para reinar sobre a terra, estabelecer justiça na terra, remover queixas, repelir o(s) inimigo(s) a

> evitar a rebelião pela força das armas, (e) trazer prosperidade aos sumérios e acadianos. (Prólogo Lipit-Ishtar)
>
> **Hamurabi**, o excelso príncipe, o adorador dos deuses, para implantar a justiça na terra, para destruir os maus e o mal, para prevenir a opressão do fraco pelo forte... para iluminar o mundo e propiciar o bem-estar do povo. (Prólogo do Código Hamurabi)

Ressaltamos ainda que essas expectativas atreladas às máquinas se assemelham às promessas dos primeiros governantes da Mesopotâmia divinamente designados. Se analisarmos os PIA, a figura e o papel de [R] se aproxima de um Estado para garantir ordem, segurança, justiça da população e ampliar benefícios. Mas como eles são utópicos, vão além da figura de Estados modernos e, se aproximam de regimes governados pelos reis com legitimação religiosa e promessas de realizações quase que sobrenaturais.

3.5 - Considerações finais

> A convergência entre nanotecnologia, biotecnologia, informática e ciências cognitivas (que agrupamos sob sigla NBIC) coloca questões inéditas que comprometem o futuro da humanidade. O século XXI não será um rio longo e tranquilo! (BESNIER e LAURENT. 2021. p. 28)

A criação de leis ou sistemas legais, começando com códigos de lei como o de Ur-Nammu, Lipit-Ishtar e Hamurabi na antiga Mesopotâmia, representa uma das mais importantes inovações na história da civilização humana. As razões para isto são múltiplas. Antes da invenção das leis, a sociedade funcionava de maneira caótica. Sem um conjunto de regras claras para orientar o comportamento humano, os conflitos eram resolvidos principalmente por meio da força física, favorecendo aqueles com maior poder físico ou militar. A introdução de leis permitiu uma forma mais justa e equitativa de resolver disputas, estabelecendo um sistema onde as ações têm consequências predeterminadas, oferecendo uma previsibilidade que possibilita a prevenção de conflitos.

Segundo Fernanda Pirie (2021), o sistema jurídico que se desenvolveu na Mesopotâmia, na China e na Índia era distinto, em linguagem, lógica e propósito. Os reis da Mesopotâmia prometiam justiça ao seu povo, estabelecendo regras nas quais as pessoas comuns podiam, pelo menos em teoria, confiar; os governantes chineses estabeleceram sistemas de crimes e punições para trazer disciplina e ordem a seus territórios em expansão; e os brâmanes hindus procuravam guiar as pessoas comuns pelo caminho do Dhama, a ordem

cosmológica de suas tradições religiosas. Mas, embora cada um desses três sistemas jurídicos fosse único, juntos eles forneceram as formas que todas as leis subsequentes adotaram. É sem dúvida a maior conquista do estado moderno ter elementos combinados de todos os três dentro dos sistemas jurídicos que agora dominam o mundo.

Apesar deste imenso sucesso anterior, estamos entrando na era da IA que apresenta desafios únicos para o sistema jurídico. Os conceitos jurídicos tradicionais estão sendo aplicados a situações inéditas e imprevistas. As leis podem ser alteradas, novas leis podem ser adicionadas e as leis antigas podem ser removidas. Este fenômeno não é novo, mas a velocidade e a escala da mudança na era da IA são sem precedentes, além dos inéditos agentes.

> Em resumo, a era da IA começa com conceitos jurídicos tradicionais cada vez mais aplicado a circunstâncias novas e previamente imprevistas que impelem alteração jurídica. Isso já aconteceu antes, é claro, mas a era da IA vai não só será imenso em escopo, mas também prosseguirá incrivelmente rápido. Nossos sistemas jurídicos tendem a ser reativos e não proativos, especialmente quando não podemos prever como será o futuro. Um autor escreve que em 1880, especialistas encarregados de prever o que a cidade de Nova York parece que cem anos depois relatou que seria destruído. O esterco que seria gerado pelos mais de seis milhões de cavalos necessária para o povo da cidade a tornaria inabitável. O moderno motor de combustão interna e os automóveis produzidos por ele eram imprevisíveis. Prever a evolução da IA e suas tecnologias relacionadas pode ser igualmente malsucedido. (COFONE. 2018. p. 770)

Os sistemas jurídicos tendem a ser reativos, adaptando-se às mudanças depois que elas ocorrem, em vez de serem proativos e antecipar essas mudanças, tentar prever a evolução da IA e suas tecnologias relacionadas pode ser um desafio e levar a previsões equivocadas como afirma Cofone (2018). Além desta limitação mais visível, há também questões mais escusos. Verificamos no Capítulo 1 algumas regras conhecidas e extraímos regras comuns como Clareza e Previsibilidade, Responsabilidade, Justiça, Aplicação, Acessibilidade, Estabilidade e Coerência, Publicidade. Estas são regras anunciadas para elaboração e operação das leis. Mas as leis também possuem regras ocultas. Marx, nas suas observações sobre o sistema judiciário, fez críticas em diversas obras que Hunt sistematizou em seis temas.[157]

[157] Hunt enfatiza que esses temas estão presentes nos escritos marxistas de muitas maneiras diversas, com diferentes graus de sofisticação e complexidade. (MORRISON apud HUNT. 2012. p. 300)

(i) O direito é inevitavelmente político, ou o direito é uma forma política.
(ii) O direito e o Estado são estreitamente ligados; o direito mostra uma relativa autonomia em relação ao Estado.
(iii) O direito põe em vigor as relações econômicas predominantes, reflete-as ou exprime-as de alguma outra forma; a forma jurídica reproduz as formas das relações econômicas.
(iv) O direito é sempre potencialmente coercitivo ou repressivo, e manifesta o monopólio estatal dos meios de coerção.
(v) O conteúdo e os procedimentos do direito manifestam, direta ou indiretamente, os interesses da(s) classe(s) dominante(s) ou do centro detentor do poder.
(vi) O direito é ideológico; tanto exemplifica quanto legitima os valores estabelecidos da(s) classe(s) dominante(s).

Segundo os tópicos apresentados por Hunt, as leis são criadas, interpretadas e aplicadas dentro de um contexto político, e o direito pode ser visto tanto como uma expressão da política quanto como uma forma de atividade política em si. Embora o direito seja tipicamente uma função do Estado, o sistema legal muitas vezes opera com um grau de autonomia. Isso pode ser necessário para garantir a imparcialidade e o devido processo legal, mas também pode levar a tensões entre as instituições jurídicas e outras partes do governo. Não há um exemplo melhor para quem acompanhou o cenário político brasileiro dos últimos anos.

O direito pode ser uma ferramenta para reforçar as estruturas econômicas existentes, usado como uma ferramenta de coerção ou repressão. Como o Estado normalmente detém o monopólio da força legítima, ele pode usar o sistema jurídico para impor sua vontade e manter a ordem social. O direito pode ser usado para reforçar o poder e os interesses das classes dominantes. Assim, as leis podem ser criadas de tal forma que beneficiem os grupos poderosos, e os procedimentos jurídicos podem ser estruturados de maneira a favorecer aqueles com mais recursos. O sistema jurídico não é apenas um conjunto de regras neutras, mas é carregado de ideologias. Pode ser usado para encarnar e legitimar os valores e as visões de mundo dos grupos dominantes, perpetuando assim suas posições de poder. Não apenas o sistema jurídico, mas todo pretenso universal como afirma Wolff (2018).

> Todo pretenso universal atribuído à humanidade em geral seria, no fundo, apenas a projeção dos valores particulares da cultura dominante. Os direitos humanos, por exemplo, seriam inseparáveis da cultura europeia do século XVIII, época em que surgiram (luta contra o absolutismo, filosofia do liberalismo, sonho de uma igualdade formal). Pior: todo pretenso universal seria apenas a tradução dos

> interesses particulares dos poderosos. Em resumo, a "humanidade" se batizou "civilização" para esconder a barbárie. (WOLFF. 2018. p. 19)

Morrison (2012) afirma que "o direito é uma técnica específica de dominação humana" e "a legalidade é a técnica mais apropriada ao poder moderno", pois apresenta verdade própria e tem capacidade de criar direitos e deveres, responsabilidades e soluções. O direito pode criar seu próprio universo de sentido e necessidade. Se ainda contextualizarmos essas observações para atual era da tecnologia protagonizada pela IA, mais que os Estados, as gigantescas corporações têm um poder sem precedente na história.

> Agora as corporações governam a sociedade, talvez mais do que os próprios governos; ironicamente, ainda assim é seu próprio poder, muito do qual ganho por meio da globalização da economia, que as torna vulneráveis. Assim como acontece com qualquer instituição dominante a corporação agora atrai desconfiança, medo e exigências de responsabilidade de um público cada vez maior. Os atuais líderes corporativos entendem, assim como seus antecessores, que é preciso esforço para reconquistar e manter a confiança do público. E eles como seus antecessores, buscam suavizar a imagem das corporações apresentando-se como humanas, benevolentes e socialmente responsáveis. (BAKAN. 2007. p. 30.)

Se o poder das corporações já era notório em 2007, quando o livro "A corporação" foi escrito, passados mais de 15 anos, esse poder certamente ficou ainda maior nas empresas de tecnologia. Para Laurent Alexandre (2021), "diante desse verdadeiro rolo compressor que o Vale do Silício representa, o Estado fica aturdido e patina sem sair do lugar." Ainda, "quer você queira ou não, é exatamente esse Gafam[158] que desenha os contornos da humanidade de amanhã, incluindo-se aí a transformação da noção do próprio homem". O poder público mostra-se incapaz diante da nova configuração de poder da era digital. Como visto no estudo realizado pelos pesquisadores da China, a indústria enfatiza seus próprios interesses através dos PIA que elaboram independentemente.

> É claro que hoje procuramos fazer o certo com relação à ética, com comitês que examinam a aceitabilidade das realizações técnicas, mas o jogo é duro pois o incentivo à inovação a qualquer preço se tornou um verdadeiro dogma entre os que tomam decisões na política e nas indústrias. (BESNIER e LAURENT. 2021. p.28)

O antropólogo argentino Nestor García Canclini (2021) questiona qual seria uma alternativa diante desta assimetria de poder entre a população e as mega corporações e os Estados sob lobby e influência delas.

[158] O acrônimo (GAFAM) representa as empresas Google, Apple, Facebook, Amazon e Microsoft.

> Quais alternativas temos diante dessa desapropriação? Dissidência, *hacking*? Qual é o lugar do voto, essa relação entre Estado e sociedade reprogramada pelas tecnologias e pelo mercado, cujo valor é questionado pelos movimentos sociais independentes? (CANCLINI. 2021. p. 15)

Há ainda um ponto quase que não questionado sobre os PIA, como explicar esta disposição para aceitar [R] como parte da sociedade, até se formando uma comunidade moral, se levarmos ao limite a realizabilidade conforme apresentado anteriormente. Esta iniciativa teve início primeiramente nas corporações.[159] Se aceitação de até outros seres humanos numa comunidade enfrenta sempre imensos obstáculos, não deixa de ser inusitada esta vontade para incluir [R], como agente capaz de seguir leis, como seres humanos. Se somente o ser humano poderia legitimar e usufruir esta posição única, o que faz com que a IA também ocupe este espaço até então exclusivo, de onde surge essa “concessão voluntária”?

O filósofo coreano Han Byung-Chul no seu livro “A sociedade do cansaço” discute os conceitos de "época imunológica" e "pós-imunológica" para ilustrar a transição na forma como a sociedade lida com ameaças e desafios. A "época imunológica" é uma sociedade que se protege ativamente contra ameaças externas, da mesma forma que um sistema imunológico protege um organismo contra patógenos. Essa era é caracterizada pela defesa contra o "outro" ou o "estrangeiro". A sociedade se blinda, construindo barreiras físicas e ideológicas para se proteger. Essa fase está relacionada com o conceito de negatividade, onde as ameaças e o perigo vêm de fora e a sociedade se protege contra elas.

> O século passado foi uma época imunológica. Trata-se de uma época na qual se estabeleceu uma divisão nítida entre dentro e fora, amigo e inimigo ou entre próprio e estranho. Mesmo a Guerra Fria seguia esse esquema imunológico. O próprio paradigma imunológico do século passado foi integralmente dominado pelo vocabulário dessa guerra, por um dispositivo francamente militar. A ação imunológica é definida como ataque e defesa. Nesse dispositivo imunológico, que ultrapassa o campo biológico adentrado no campo e em todo o âmbito social, ali foi inscrita uma cegueira: Pela defesa, afasta-se tudo que é estranho. O objetivo da defesa imunológica é a estranheza como tal. Mesmo que o estranho não tenha nenhuma intenção hostil, mesmo que ele não represente nenhum perigo, é eliminado em virtude de sua alteridade. (HAN. 2012. p.8)

Han descreve uma era pós-imunológica onde as ameaças não vêm mais de fora, mas de dentro. Em vez de um "outro" ou "estrangeiro" a ser defendido, a sociedade é

[159] Os Princípios de Parceria em IA de 2016 inauguraram a Segunda Fase de PIA com a participação de principais empresas de tecnologia do mundo: Amazon, DeepMind/Google, Facebook, IBM e Microsoft. A partir desta iniciativa começou a corrida nas elaborações de PIA.

atormentada pelo excesso de positividade, onde o perigo vem do "eu" ou do "interior". Essa fase é caracterizada pelo auto exploração e autocobrança, que resultam em doenças modernas, como burnout, depressão e outros transtornos relacionados ao estresse. A sociedade do cansaço é o resultado dessa era pós-imunológica, onde as pessoas se esgotam ao tentar atender às demandas implacáveis de produtividade e otimização de si mesmas, as pessoas estão constantemente estressadas diante de uma cultura moderna de alto desempenho. Han argumenta que a transição de uma sociedade que se defende contra ameaças externas para uma que está constantemente se esforçando para melhorar e otimizar a si mesma resultou em uma sociedade do cansaço.

> A alteridade é a categoria fundamental da imunologia. Toda e qualquer reação imunológica é uma reação à alteridade. Mas hoje em dia, em lugar da alteridade entra em cena a diferença, que não provoca nenhuma reação imunológica. A diferença pós-imunológica, sim, a diferença pós-moderna já não faz adoecer. […] Também o assim chamado "imigrante", hoje em dia, já não é mais imunologicamente um outro; não é um estrangeiro, em sentido enfático, que representa um perigo real ou alguém que nos causasse medo. Imigrantes são vistos mais como um peso do que como uma ameaça. […] O mundo organizado imunologicamente possui uma topografia específica. É marcado por barreiras, passagens e soleiras, por cercas, trincheiras e muros. Essas impedem o processo de troca e intercâmbio. (HAN. 2010. p.10~13)

Avaliamos a descrição de Han sobre a era imunológica esclarecedora, no entanto, discordamos que vivemos hoje em uma era pós-imunológica. O livro de Han foi publicado em 2010. Antes de alguns importantes acontecimentos como a Crise dos Refugiados na Europa (2015-presente), Eleição de Donald Trump (2016) e quase 700 km de "grande e belo muro" ao longo da fronteira dos Estados Unidos com o México, Brexit (2016-2020), Pandemia de Covid-19 (2019-presente), Protestos Black Lives Matter (2020), Guerra na Ucrânia (2022-presente), inúmeros naufrágios de embarcações com imigrantes no Mediterrâneo (presente).

Talvez o decreto da passagem da era imunológica para pós-imunológica fosse precipitado, pois continuamos impedindo a inclusão de outros através de obstáculos físicos e também invisíveis. A polarização política tornou-se extrema e tomou uma escala mundial, onde o outro lado é visto como inimigo a ser destruído, mesmo em democracias até então vistas como maduras. É por isso que a tentativa de inclusão da IA e robôs em uma mesma comunidade destoa mais ainda. Se usarmos os termos de Han, talvez a IA e robôs não são vistos como "um outro; não é um estrangeiro", e assim não causa uma reação imunológica.

A alteridade foi um conceito central na filosofia de Lévinas. No contexto do seu pensamento, a alteridade é a ideia de que o Outro é inescrutável, inatingível e fundamentalmente separado do Eu. Lévinas propõe que o encontro com o Outro, a face do Outro, impõe uma responsabilidade ética primordial. O rosto do Outro invoca uma demanda direta e inescapável por resposta e responsabilidade. A experiência de alteridade nos força a reconhecer nossa responsabilidade perante o Outro, estabelecendo a base para a ética e a moralidade. Segundo Lévinas, a ética, em vez da ontologia é a filosofia primeira. Isso significa que nosso primeiro dever filosófico é para com os outros, não para conosco. Ele argumenta que nossa relação com o Outro não é de conhecimento ou reconhecimento, mas de responsabilidade.

> Compreender que o outro é referência da vida moral e princípio orientador da existência incide profundamente sobre o entendimento da condição humana. Já não é a reflexão, no sentido do retorno do sujeito a si mesmo, que fornecerá os parâmetros fundamentais do conhecimento do homem. Trata-se, agora, de uma abertura àquele que não sou eu e, no limite, de uma renúncia ao Eu como polo irradiador de valores. Não é a consciência de si que dá sentido ao mundo, mas a consciência do outro que constitui o critério diretor de existência de cada sujeito, que se forma em sua integridade não apenas em relação ao outro, mas em virtude da existência do outro. (SILVA. 2012. p. 33)

Franklin Leopoldo e Silva (2012) analisa a filosofia de Emmanuel Lévinas e afirma que "esse outro não é, de forma alguma, o próximo e o familiar, mas o estranho que devo esforçar-me para compreender". Não saberemos se o outro de Lévinas contemplava possibilidade de inclusão das máquinas, entretanto, esta ideia poderia ser estendida para compreensão de um convívio com elas. Se as normas de sociabilidade - as regras e expectativas que governam nosso comportamento social - são menos regras que escolhemos seguir, e mais regulações *a posteriori* de uma condição originária. Isso implica que essas normas são uma tentativa de formalizar e regular algo que já existe: a profunda conexão e interdependência entre nós como indivíduos e a sociedade em que vivemos.

> A relação com o outro se encerra na dimensão da sociabilidade estabelecida por acordo ou por contrato. A solidariedade torna-se uma questão de regras de conveniência. As sociedades modernas, fruto das teorias políticas liberais clássicas, atendar a esse perfil. Se concordarmos que o indivíduo se define pela comunidade à qual está organicamente vinculado e que o sentido da existência singular é inseparável do contexto comunitário que o produz e o sustenta, então poderemos e são vividos como dimensão essencial da realidade humana. A relação com o outro possui a densidade e a força dos princípios necessariamente vistos como requisitos primordiais da existência, a tal ponto que o indivíduo autossuficiente seria uma

> abstração. As normas de sociabilidade seriam apenas, no limite, regulações *a posteriori* de uma condição originaria. (SILVA. 2012. p.36)

De fato, o convívio com a IA previsto nos PIA, no limite, seria regulações *a posteriori* de uma condição inédita. Se a identidade de indivíduo se define pela comunidade à qual está vinculado, uma comunidade [R a H] redefinirá uma nova essência humana. Se há esta condição dada *a posteriori* para possíveis regulações diante do surgimento de um outro, a sua justificativa pode levar em consideração um movimento *a priori*, o modo como a própria tecnologia se desenvolve, que segundo Besnier é incapaz de impor limites a si mesma.

> Admitamos, portanto, que a tecnologia seja *a priori* incapaz de impor limites a si mesma: diremos que ela é, com efeito, o lugar da expressão da desmesura (a hubris, diziam os gregos) da qual os humanos são capazes. Ela só pode receber freios vindo de fora, isto é, deve ser moderada por aquilo que surge da reflexão e do simbólico (ou seja, da comunicação política permitida pela linguagem). (BESNIER e LAURENT. 2021. p. 98)

A história humana pode ser contada por meio de dois âmbitos: medo e felicidade. A busca de proteção diante do mal e a busca por felicidade atuaram como impulsionadores de nossas ações, moldando as escolhas individuais e coletivas. A história, portanto, pode ser vista como uma narrativa da maneira como a humanidade tem lidado com esses dois aspectos fundamentais da experiência emocional. A lei sempre foi um mecanismo para em fundamento mais básico garantir a sobrevivência (felicidade) evitando danos com resultados comprovados na história. Assim, recorremos às leis tanto para evitarmos uma guerra de todos contra todos como para superarmos o medo diante da natureza hostil.

Tabela 2 - A utilização das leis na história humana

	Leis da sociedade	Leis da natureza/física	Leis para robôs/IA
Agentes	Seres humanos	Natureza	Robôs/IA
Legisladores	Humanos/Divino	Cientistas	Seres humanos/IA
Objetivo	Ordem e justiça	Entender e dominar a natureza	Ordem e justiça, Entender e dominar a IA

As primeiras leis da civilização trouxeram ordem para sociedade, a ciência foi uma tentativa de entender a natureza hostil. A utilização das leis segue uma ordem na história humana: leis da sociedade, leis da natureza, leis para robôs/IA. A lei - uma vez incorporada com eficiência ao longo da história humana - é um recurso que não os conseguimos mais

livrar-se. Sempre que aparece algum obstáculo e ameaça a essa ordem mantida na sociedade, buscamos uma solução no mecanismo da lei. Não foi diferente diante da aparição de IA que tanto provoca medo como entrega promessa de felicidade. Assim, evocamos a solução legal automaticamente para lidarmos com dois âmbitos simultaneamente.

> Como profissionais do direito, nossa reação inicial diante das tecnologias que não entendemos muito bem é muitas vezes seguir o caminho legislativo e elaborar um marco legal destinado a controlar o uso e disseminação de essas tecnologias. A IA não escapou dessa tendência, pois muitos estados têm (LESSIG. 2018. p. 68)

Lawrence Lessig, professor na faculdade de direito de Harvard, argumenta que PIA surgiram como resposta para o desenvolvimento da tecnologia. Antes da proliferação dos PIA nos últimos anos, o sociólogo Amitai Etzioni (1968) já defendeu uma 'sociedade ativa'[160] na qual os valores normativos deveriam guiar o desenvolvimento tecnológico e os seres humanos utilizariam e controlariam a tecnologia para o benefício da humanidade.

Acreditamos que a tentativa de controle da IA por meio dos PIA remete ao conceito de Princípio da Cópia de Hume. A nossa imaginação é limitada pelo que já experimentamos através de nossos sentidos. A nossa capacidade tem limite como argumentou Hume. Nós criamos máquinas através de antropomorfização e queremos controlá-las através de leis como se fossem seres humanos.

> - Agir em benefício de todos e para evitar permitir usos que prejudiquem a humanidade ou concentrem poder indevidamente. (OpenAI. 2018. EUA)
>
> - Não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. (Future Society. 2017. EUA e UE)
>
> - Aumentar nossa humanidade, não a diminuir ou substituí-la. (HAI. 2018. EUA)

Acreditamos que os três princípios mencionados anteriormente encapsulam de maneira abrangente os demais PIA. Eles refletem um compromisso com a mitigação de danos e a maximização de benefícios. Recorremos, uma vez mais, à invenção mais crucial da história da humanidade, comprovada ao longo de milênios, para estabelecer ordem e justiça na sociedade. Quase quatro mil anos se passaram desde a promulgação dos primeiros códigos legais conhecidos da civilização humana, como os de Ur-Nammu, Lipit-Ishtar e Hamurabi.

[160] BEST e KELLNER. *Postmodern Theory*. 1991. p. 13.

Surpreendentemente, ao revisitar esses códigos, notamos uma essência praticamente inalterada, salvo pelo ressurgimento do termo "código".[161] Tanto os códigos ancestrais, como Ur-Nammu, quanto os princípios modernos, como Asilomar, compartilham semelhanças fundamentais, divergindo apenas nos meios empregados: tábuas de argila endurecida ao sol versus telas eletrônicas.

Como mencionado anteriormente, as expectativas depositadas nas máquinas ecoam as promessas dos primeiros governantes da Mesopotâmia, considerados designados divinamente. Assim, apesar do vasto conhecimento acumulado ao longo de mais de quatro mil anos, buscamos nas leis a salvaguarda necessária para enfrentar os potenciais desafios à ordem e justiça na sociedade, provocados pela Inteligência Artificial.

Podemos atualizar a divisão apresentada no capítulo 2 da parte I com o acréscimo da terceira fase, ainda com algumas lacunas que representam incertezas para o futuro que apresenta questões inéditas. Nomeamos esta fase como “Terceira fase futura” que começaria a partir da agência de [R] em cumprir leis na prática. Em se tratando de um sistema real, contemplaria tanto proibição de danos como garantia de benefícios. Estas leis seriam destinadas aos membros de uma comunidade formada por “seres híbridos” que seriam também os legisladores. Neste ponto, podemos ainda ressaltar que o título do trabalho “As regras das leis para humanos, não humanos e trans-humanos - (Código Ur-Nammu, Três leis da Robótica de Asimov e Princípios de Asilomar)” apresenta uma falta de correspondência para últimos: As regras das leis para trans-humanos com os Princípios de Asilomar, uma vez que não há por enquanto leis para trans-humanos. Entretanto mantivemos Asilomar por ser última grande iniciativa que coloca problemas de longo prazo: “19º Não havendo consenso, devemos evitar fortes suposições sobre os limites superiores das futuras capacidades de IA.”

[161] "Código é lei" é uma frase famosa do professor de Direito Lawrence Lessig, do seu livro "Code and Other Laws of Cyberspace" (1999). Laurent Alexandre argumenta que “as tecnologias NBIC justificariam uma reinvenção do papel regulador do Estado. Fusão entre tecnologia e lei: Code is law (código é lei) virá a ser uma realidade política.” (Laurent. 2022. p. 98)

Tabela 3 - Classificação de PIA atualizada

Fases	Primeira fase tradicional	Segunda fase atual	Terceira fase futura
Época	1942 até 2015	2016 até hoje	A partir de [R] com agência legal
Modo	Negativa: Proibição de danos	Positiva: Proporcionar benefícios	Real: Negativa e Positiva
Termo mais utilizado	Lei	Princípio	Lei (Código?)
Protagonismo	Robôs	IA	Cyborgue/Híbrido
Principal	Asimov	Asilomar	?
Críticas	Ficção	Realizabilidade	?
Méritos	Pioneirismo	Alerta e discussão	?
Objetivo principal	Proibição de danos para seres humanos	Trazer benefícios para seres humanos	Proibição de danos e trazer benefícios para seres humanos
Objetivos secundários	Proteção do robô	Privacidade, Explicabilidade, Justiça, Transparência,	?
Legislador	Ser humano	Ser humano	Ser humano, IA, Trans-humano

"Periculum in mora" [162] é um termo jurídico em latim que se traduz como "perigo na demora". É um princípio jurídico que se refere à necessidade de uma ação judicial imediata para prevenir danos irreparáveis ou muito difíceis de reparar. O "periculum in mora" se configura quando a demora na prestação jurisdicional poderia tornar inútil o provimento final, causando prejuízos graves e de difícil reparação ao requerente. Para Pirie (2021) mesmo as leis contemporâneas formuladas em resposta a um problema social nem sempre são tão pragmáticas quanto os governos querem que acreditemos. Muitas vezes as novas leis são impraticáveis ou inexequíveis. Mas os governos devem ser vistos como fazendo algo sem

[162] Disponível em <https://www.mpf.mp.br/es/sala-de-imprensa/glossario-de-termos-juridicos> Acesso em outubro de 2022.

demora. Aprovar uma lei dá a seus cidadãos a impressão de que os políticos estão no controle da situação, além de expressar a preocupação moral da sociedade em geral. As leis estabelecem, para todos verem, os parâmetros morais da sociedade civilizada que os governos (junto com as corporações) afirmam que podem criar. Eles mantêm a promessa de justiça e ordem, nem sempre praticáveis. A possibilidade de [R] cumprir leis é talvez seja menos relevante que [H] acreditar que [R] estarão sob as leis. Este desejo por segurança pode estar por trás das iniciativas de criação de PIA.

Este argumento também pode ser reforçado por uma outra questão. Não sabemos ao certo como as leis funcionam para seres humanos. Passaram se mais de quatro mil anos desde os primeiros códigos que regem a nossa civilização, mas o mecanismo de funcionamento de lei para o ser humano ainda é uma questão aberta. Não há consenso sobre ele, há autores atuais que negam a importância de coerção e recompensas, e outros elementos até então tidos como essenciais. Queremos controlar a IA com os PIA, mas nem sequer sabemos exatamente porque as leis funcionam na sociedade formado por seres humanos.

A lei não é o único domínio normativo em nossa sociedade; convenções sociais, etiqueta, moralidade, religião e assim por diante, também orientam a conduta humana de muitas maneiras semelhantes à lei. Portanto, parte do que está envolvido na compreensão da natureza do direito consiste em uma explicação de como o direito difere desses domínios, como ele interage com eles e se sua inteligibilidade depende de outras ordens normativas.

> Uma hipótese é que tais normas cooperativas surgem em grupos unidos onde as pessoas têm interações contínuas umas com as outras (Hardin 1982). A teoria evolutiva dos jogos fornece uma estrutura útil para investigar essa hipótese, uma vez que os jogos repetidos servem como uma simples aproximação da vida em um grupo unido (Axelrod 1984, 1986; Skyrms 1996; Gintis 2000). Em encontros repetidos, as pessoas têm a oportunidade de aprender com o comportamento umas das outras e de garantir um padrão de reciprocidade que minimize a probabilidade de percepção errônea. A esse respeito, argumenta-se que as normas cooperativas que provavelmente se desenvolverão em grupos unidos são simples (Alexander 2000, 2005, 2007); na verdade, punições atrasadas e desproporcionais, bem como recompensas tardias, são muitas vezes difíceis de entender e, portanto, ineficazes.[163]

[163] Disponível em <https://plato.stanford.edu/entries/social-norms/> Acesso em outubro de 2022. Tradução nossa.

Se aplicarmos a hipótese acima, um aperfeiçoamento pode ocorrer através de interações contínuas entre [H] e [R]. A tecnologia, *a priori*, não tem freio no seu desenvolvimento. O sistema jurídico humano que opera de forma reativo e não proativo, não conseguirá acompanhar a velocidade de desenvolvimento da IA. Desta forma, a IA será dada, *a posteriori*, como Outro. Como a previsão Turing, talvez a sociedade precisará se adaptar com a aparição de um outro de tal forma que PIA serão dissolvidos em normas sociais. A alteração que se buscaria não seria apenas uma transformação dos seus membros pelos dois caminhos anteriormente expostos: humanização das máquinas, maquinização dos seres humanos, mas da própria natureza normativa da sociedade.

Hans Vaihinger, autor da obra "A filosofia do como se" publicada pela primeira vez em 1911, destacou-se como um filósofo cujas contribuições foram cruciais para o desenvolvimento da teoria filosófica do ficcionalismo. Em sua obra, Vaihinger defende a perspectiva de que o conceito de ficção permeia uma miríade de aspectos na sociedade, exercendo uma influência notável, inclusive no sistema jurídico. Suas ideias desvelam a natureza profundamente entrelaçada das construções ficcionais com a realidade, propondo uma abordagem perspicaz para compreendermos a tessitura intricada da vida social e jurídica.

> Uma variedade especial das ficções anteriormente discutidas são as ficções jurídicas. Em nenhum outro lugar, o termo ficção é mais conhecido do que na jurisprudência, onde representa tema preferido de discussão. Em princípio, são inteiramente idênticas às ficções anteriores. O mecanismo psicológico de sua aplicação consiste em que um caso singular é subsumido em uma construção de representações não destinada a ele, ou seja, a apercepção é meramente analógica. A base do método é a seguinte: uma vez que as leis não podem enquadrar todos os casos singulares em suas fórmulas, contemplam-se alguns casos especiais de natureza não comum *como se* estes pertencessem àquelas. Ou, em razão de um interesse prático qualquer subsume-se um caso singular e um conceito geral, ao qual, no fundo, não pertence. (VAIHINGER. 2011. p. 157)

De acordo com Vaihinger, na "fictio iuris", um caso é subsumido a uma relação de analogia de maneira estritamente oposta à realidade. O conceito de pessoas jurídicas é um exemplo da ficção jurídica, as corporações operam como se fossem entidades dotadas de personalidade. Os PIA podem ser "fictio iuris" em sentido ainda mais amplo que o sistema jurídico tradicional em si.

Vaihinger propõe que as pessoas frequentemente operam sob a premissa de ideias "como se" fossem verdadeiras, mesmo que possam não ser. Essa abordagem "como se" é

considerada uma ferramenta cognitiva útil para lidar com a incerteza e a complexidade do mundo. A teoria destaca o valor pragmático de adotar crenças ou ideias fictícias para orientar a ação e compreensão do mundo. Essa utilidade é particularmente relevante em situações em que a realidade é difícil de determinar. Ao examinarmos os PIA, torna-se evidente a presença de aplicações ficcionais em diversos níveis, indo além da sua origem em uma obra de ficção científica.

> Nas diversas ciências, levantamos muito dessas suposições conscientemente falsas, justificando-as por sua utilidade. Algo semelhante ocorre na vida prática: a pressuposição da liberdade da nossa vontade representa a base necessária de nossas instituições sociais e jurídicas; não obstante, nossa consciência lógica nos diz ser a pressuposição da liberdade da vontade algo sem sentido do ponto de vista lógico. Entretanto, não é por isso que abrimos mão daquela representação; pois é útil e mesmo imprescindível. […] Preservamos tais modos, não porque nos seriam "caros", mas porque reconhecemos sua utilidade e indispensabilidade em vista da ação bem-sucedida. No campo teórico, prático e religioso, descobrimos o que é correto na base e com auxílio do que é falso. (VAIHINGER. 2011. p. 88)

Vaihinger argumenta que muitos conceitos, como os princípios fundamentais da matemática ou mesmo certos valores morais, são construções fictícias. Esses conceitos fictícios são adotados porque funcionam de maneira eficaz, não necessariamente porque representam verdades. O uso de ficções não se limita à esfera individual, mas pode ser aplicado em escalas mais amplas, incluindo o progresso científico e a organização social. Ele argumenta que conceitos fictícios podem ser motores cruciais para o avanço humano.

A teoria de ficção de Vaihinger oferece uma perspectiva alternativa sobre a natureza do conhecimento, propondo que a adoção de ficções pode ser uma abordagem valiosa para compreender e agir no mundo, especialmente em situações onde a certeza é difícil de alcançar. A criação de uma comunidade moral coesa com a Inteligência Artificial, juntamente com o correto funcionamento dos PIA, pode requerer uma adesão convincente e generalizada ao conceito de ficção por parte de todos os membros da sociedade: seres humanos, não humanos e trans-humanos.[164]

[164] O termo pós-humanos poderia ter sido uma escolha melhor, mas mantivemos o termo trans-humanos inicial.

Bibliografia

BAKAN, Joel. *A Corporação – A Busca patológica por lucro e poder*. Novo Conceito. São Paulo. 2007.

BALKIN, Jack M. *The Three Laws of Robotics in the Age of Big Data* (2017). Faculty Scholarship Series. 5159. Disponível em [http://digitalcommons.law.yale.edu/fss_papers/5159]. Acesso em maio de 2019.

BAUMAN, Zygmunt. *Ética Pós-Moderna*. Paulus. São Paulo. 1997

BINGHAM, Tom. *The Rule of Law*. Penguin Books. London. 2011.

CANCLINI, Néstor García. *Cidadãos Substituídos por Algoritmos.* São Paulo. Edusp. 2021.

CANTÙ, Paola e LUCIANO**,** Erika**.** *Giuseppe Peano and his School: Axiomatics, Symbolism and Rigor - Philosophia Scientiæ* 25-1. Disponível em [http://journals.openedition.org/philosophiascientiae/2788]. Acesso em maio de 2023.

COELHO SOUZA, Daniel. *Introdução à ciência do direito*. Rio de Janeiro: Fundação Getulio Vargas, 1972. - 2ª ed. São Paulo: Saraiva, 1983.

COFONE, Ignacio, N. *Servers and Waiters: What Matters in the Law of A.I.* Disponível em https://law.stanford.edu/publications/servers-and-waiters-what-matters-in-the-law-of-a-i/. Acesso em maio de 2019.

CORTINA. Adela. *Ética Mínima*. São Paulo. Martins Fontes. 2008.

DILTHEY, Wilhelm. *Sistema da Ética, Coleção Fundamentos de Direito*. São Paulo. Ícone Editora. 2005.

DUHEM, Pierre. *Ensaios de filosofia da ciência*. Associação Filosófica Scientia Studia. São Paulo. 2019.

ELLUL, Jacques. *The Technological Society*. Vintage Books. Toronto. 1964.

EINSTEIN, Albert. *Out of My Later Years*. Philosophical Library. New York. 1950.

FEDERATION OF GERMAN SCIENTISTS E.V. (VDW). *Policy Paper on the Asilomar Principles on Artificial Intelligence*. 2018 Office of the Federation of German Scientists, Marienstraße 19/20, 10117 Berlin.

FJELD, Jessica. et al. "*Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*." Berkman Klein Center for Internet & Society, 2020. Disponível em https://dash.harvard.edu/handle/1/42160420 Acesso em maio de 2023.

FLORIDI, Luciano and COWLS, Josh, *A Unified Framework of Five Principles for AI in Society* (September 20, 2019). Disponível em
SSRN: https://ssrn.com/abstract=3831321 or http://dx.doi.org/10.2139/ssrn.3831321
Acesso em maio de 2023.

GANASCIA, Jean, Gabriel. *Le mythe de la singularité*. Seul. Bookpot. 2017.

GIUFFRIDA. Iria, LEDERER. Fredric. VERMERYS. Nicolas. *A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law*, 68 Case W. Res. L. Rev. 747 (2018). Disponível em https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/14

GOMES. Laurentino. *Escravidão Vol. I As Origens*. Ed. Globo. Rio de Janeiro. 2019.

HABERMAS, Jürgen. *A inclusão do outro*. Editora Unesp. São Paulo. 2011.

HACKING, Ian. *Ensaio introdutório da edição comemorativa dos 50 anos da publicação – A Estrutura das Revoluções científicas*. Kuhn. Thomas. Ed. Perspectiva. São Paulo. 2017

HAN, Byung-Chul. *Sociedade do cansaço*. 2ª ed. RJ. Editora Vozes. 2017.

HOQUET. Thierry. Filosofia Ciborgue. São Paulo. Ed. Perpectiva. 2019.

JONAS. Hans. *Princípio Responsabilidade*. São Paulo. Contraponto. 2007.

LAURENT, Alexandre. BESNIER, Jean-Michel. *Os robôs fazem amor? O transumanismo em doze questões*. São Paulo. Ed. Perspectiva. 2022.

LEHOUX D. *Saved by the phenomena: Law and nature in Cicero and the (pseudo?) Platonic Epinomis*, Studies in History and Philosophy of Science (2019), Disponível em [https://doi.org/10.1016/j.shpsa.2019.03.001.] Acesso em maio de 2019.

LESSIG. Lawrence. *Commentary: The Law of the Horse: What Cyberspace Technologies Will Affect the Law*, 68 Case W. Res. L. Rev. 747 (2018) Disponível em :
https://scholarlycommons.law.case.edu/caselrev/vol68/iss3/14

MASI, Domenico de. *O futuro chegou*. RJ. Editora Casa da Palavra. 2014.

MINSKY, Marvin. *A celebration of Isaac Asimov*. Disponível em [https://www.nytimes.com/1992/04/12/business/technology-a-celebration-of-isaac-asimov.html?pagewanted=all&src=pm]

MORRISON, Wayne. *Filosofia do Direito – Dos gregos ao pós-modernismo*. São Paulo. Martins Fontes. 2012.

MOULINES, Carlos, Ulisses. *O desenvolvimento moderno da filosofia da ciência* (1890 – 2000). Associação Filosófica Scientia Studia. São Paulo. 2020.

PESSOA, Osvaldo Jr. *Notas de Aula de Teoria do Conhecimento e Filosofia da Ciencia I: Um Panorama Histórico com Olhar Contemporâneo*. São Paulo. 2014.

PIRIE, Fernanda. *The Rule of Laws: A 4,000-Year Quest to Order the World*. London. Profile Books. 2021.

RAWLS, John. História da filosofia moral. Martins Fontes. São Paulo. 2005

ROTH, T. Martha. *Law collections from Mesopotâmia and Asia Minor*. Atlanta: Scholars Press. 1995.

ROUSSEAU, J.-J. *Discurso sobre a origem e os fundamentos da desigualdade entre os homens*. 3ª ed. Martins Fontes. São Paulo. 2005.

RUSSELL, Stuart. *Artificial Intelligence – A modern approach.* 3rd edition. New Jersey. 2005.

SILVA, Franklin, Leopoldo. *O outro.* São Paulo. Martins Fontes. 2012.

SPINOZA. *Ética*. 2ª ed. Belo Horizonte. São Paulo. Autêntica Editora. 2013.

SOUZA, Daniel, Coelho. *Introdução à ciência do direito*. Rio de Janeiro: Fundação Getúlio Vargas. Saraiva. 1983.

TURING. Alan. *Computing Machinery and intelligence. Mind*. VOL. LIX. NO. 236. Reino Unido. 1950.

VAIHINGER. Hans. *A filosofia do como se*. Chapecó. Argos. 2011.

WIENER, Norbert. *Cibernética e Sociedade – o uso humano de seres humanos*. 2ª ed. São Paulo. Cultrix. 1985.

WITTGENSTEIN, Ludwig. *Major Works.* New York. HarperCollins. 2009.

WOLFF, Francis. *Três Utopias Contemporâneas*. São Paulo. Editora Unesp. 2018.

YI Zeng, ENMENG Lu, CUNQING, Huangfu. *Linking Artificial Intelligence Principles.* AAAI Workshop on Artificial Intelligence Safety (AAAI-Safe AI 2019), 2019. Disponível em https://arxiv.org/abs/1812.04814

APÊNDICE

Os Princípios Gerais do Design Eticamente Alinhado – IEEE versão 2 [165](2017. Mundial)

O Instituto de Engenheiros Elétricos e Eletrônicos (IEEE) é uma das maiores organizações profissionais do mundo dedicada ao avanço da tecnologia. Diante da rápida evolução das tecnologias de IA e sistemas autônomos, o IEEE reconheceu a necessidade de uma estrutura ética para guiar o desenvolvimento e uso dessas tecnologias. Em 2016, o IEEE lançou a iniciativa global intitulada "Design Eticamente Alinhado para Sistemas Autônomos e Inteligentes" (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems).

Depois de um ano de deliberações e consultas públicas, o IEEE publicou a primeira versão dos "Princípios Gerais do Design Eticamente Alinhado" em 2017. Esses princípios abrangem uma variedade de tópicos, desde a necessidade de transparência na operação da IA passando pelo respeito à privacidade e aos dados dos usuários, até a importância de garantir a responsabilidade e a responsabilização no uso da IA.

> 1. Direitos Humanos– Sistema de IA deve ser criada e operada para respeitar, promover e proteger os direitos humanos reconhecidos internacionalmente.
>
> 2. Bem-estar Os criadores de Sistema de IA devem adotar o aumento do bem-estar humano como critério primário de sucesso para o desenvolvimento.
>
> 3. Os criadores da Agência de Dados – Sistema de IA devem capacitar os indivíduos com a capacidade de acessar e compartilhar com segurança seus dados, para manter a capacidade das pessoas de ter controle sobre sua identidade.
>
> 4. Eficácia – Os criadores e operadores de Sistema de IA devem fornecer evidências da eficácia e adequação ao propósito de Sistema de IA.
>
> 5. Transparência – A base de uma decisão específica de Sistema de IA deve sempre ser descoberta.
>
> 6. Responsabilidade– Sistema de IA deve ser criado e operado para fornecer uma justificativa inequívoca para todas as decisões tomadas.
>
> 7. Conscientização sobre uso indevido – Os criadores de Sistema de IA devem se proteger contra todos os possíveis usos indevidos e riscos de Sistema de IA em operação.

[165] Disponível em <https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e_principles_to_practice.pdf> Acesso em maio de 2023. Tradução nossa.

8. Os criadores da Competência– Sistema de IA devem especificar e os operadores devem aderir ao conhecimento e habilidade necessários para uma operação segura e eficaz.

Princípios para transparência algorítmica e responsabilidade [166]- USACM (2017. EUA)

A ACM, Associação para Maquinário de Computação, é uma das organizações profissionais mais conhecidas na área de ciência da computação. Dentro da ACM, o USACM, ou o Conselho de Políticas de Tecnologia da Informação dos EUA da ACM, é o grupo que se concentra em fornecer análises e recomendações políticas sobre questões que envolvem tecnologia da informação e comunicações.

Com o crescimento da inteligência artificial e o uso de algoritmos em diversos aspectos da vida cotidiana, surgiu uma preocupação crescente sobre a transparência e a responsabilidade dos sistemas algorítmicos. Em particular, questões relacionadas à justiça, viés, discriminação e transparência tornaram-se pontos críticos de discussão na sociedade. Em resposta a essas preocupações, o USACM lançou uma iniciativa para desenvolver um conjunto de princípios para orientar a transparência algorítmica e a responsabilidade.

1. Conscientização: Proprietários, projetistas, construtores, usuários e outras partes interessadas de sistemas analíticos devem estar cientes dos possíveis vieses envolvidos em seu projeto, implementação e uso e o dano potencial que os vieses podem causar aos indivíduos e à sociedade.

2. Acesso e reparação: Os reguladores devem encorajar a adoção de mecanismos que permitam questionamento e reparação para indivíduos e grupos que são afetados adversamente por decisões baseadas em algoritmos.

3. Responsabilidade: As instituições devem ser responsabilizadas pelas decisões tomadas pelos algoritmos que utilizam, mesmo que não seja viável explicar detalhadamente como os algoritmos produzem seus resultados.

4. Explicação: Os sistemas e instituições que usam a tomada de decisão algorítmica são incentivados a produzir explicações sobre os procedimentos seguidos pelo algoritmo e as decisões específicas que são feitas. Isso é particularmente importante em contextos de políticas públicas.

5. Proveniência dos dados: Uma descrição da forma como os dados de treinamento foi coletada deve ser mantida pelos construtores dos algoritmos, acompanhada por uma exploração dos possíveis vieses induzidos pelo processo humano ou algorítmico de coleta de dados. O escrutínio público dos dados oferece oportunidade máxima para correções. No

[166] Disponível em <https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf> Acesso em maio de 2023. Tradução nossa.

entanto, preocupações com privacidade, proteção de segredos comerciais ou revelação de análises que possam permitir que agentes mal-intencionados manipulem o sistema podem justificar a restrição de acesso a indivíduos qualificados e autorizados.

6. Auditabilidade: Modelos, algoritmos, dados e decisões devem ser registrados para que possam ser auditados em casos de suspeita de danos.

7. Validação e Teste: As instituições devem usar métodos rigorosos para validar seus modelos e documentar esses métodos e resultados. Em particular, eles devem realizar testes rotineiramente para avaliar e determinar se o modelo gera dano discriminatório. As instituições são incentivadas a tornar públicos os resultados desses testes.

Diretrizes Éticas da Sociedade Japonesa para Inteligência Artificial - JSAI [167](2017. Japão)

A Sociedade Japonesa para Inteligência Artificial (JSAI), uma das principais organizações acadêmicas no campo da Inteligência Artificial (IA) no Japão. A JSAI estabeleceu um Comitê de Ética para discutir e desenvolver diretrizes éticas para a pesquisa e aplicação de IA. Este comitê foi composto por especialistas em IA, bem como por profissionais de áreas correlatas. Após um período de discussões, o Comitê de Ética da JSAI publicou suas Diretrizes Éticas em 2017.

1 (Contribuição para a humanidade) Os membros da JSAI contribuirão para a paz, segurança, bem-estar e interesse público da humanidade. Eles protegerão os direitos humanos básicos e respeitarão a diversidade cultural. Como especialistas, os membros da JSAI precisam eliminar a ameaça à segurança humana enquanto projetam, desenvolvem e usam IA.

2 (Cumprimento das leis e regulamentos) Os membros da JSAI devem respeitar as leis e regulamentos relativos à pesquisa e desenvolvimento, propriedade intelectual, bem como quaisquer outros acordos contratuais relevantes. Os membros da JSAI não devem causar danos a terceiros através da violação de informações ou propriedades pertencentes a terceiros. Os membros da JSAI não devem usar IA com a intenção de prejudicar outras pessoas, seja direta ou indiretamente.

3 (Respeito pela privacidade dos outros) Os membros da JSAI respeitarão a privacidade dos outros no que diz respeito à sua pesquisa e desenvolvimento de IA. Os membros da JSAI têm o dever de tratar as informações pessoais de forma adequada e de acordo com as leis e regulamentos relevantes.

4 (Justiça) Os membros da JSAI serão sempre justos. Os membros da JSAI reconhecerão que o uso da IA pode trazer desigualdade e discriminação adicionais na sociedade que não existiam antes e não serão tendenciosos ao desenvolver a IA. Os membros da JSAI, da melhor

[167] Disponível em <http://www.ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf> Acesso em maio de 2023. Tradução nossa.

maneira possível, garantirão que a IA seja desenvolvida como um recurso que pode ser usado pela humanidade de maneira justa e igualitária.

5 (Segurança) Como especialistas, os membros da JSAI devem reconhecer a necessidade da IA ser segura e reconhecer sua responsabilidade em manter a IA sob controle. No desenvolvimento e uso da IA, os membros da JSAI sempre prestarão atenção à segurança, controlabilidade e confidencialidade exigida, garantindo que os usuários da IA recebam informações apropriadas e suficientes.

6 (Aja com integridade) Os membros da JSAI devem reconhecer o impacto significativo que a IA pode ter na sociedade. Eles irão, portanto, agir com integridade e de uma forma que possa ser confiável para a sociedade. Como especialistas, os membros da JSAI não farão alegações falsas ou pouco claras e são obrigados a explicar as limitações técnicas ou problemas nos sistemas de IA de forma verdadeira e cientificamente sólida.

7 (Responsabilidade e Responsabilidade Social) Os membros da JSAI devem verificar o desempenho e o impacto resultante das tecnologias de IA que pesquisaram e desenvolveram. Caso seja identificado um perigo potencial, um alerta deve ser efetivamente comunicado a toda a sociedade. Os membros da JSAI entenderão que sua pesquisa e desenvolvimento podem ser usados contra seu conhecimento para fins de prejudicar outras pessoas e farão esforços para evitar tal uso indevido. Se o uso indevido da IA for descoberto e relatado, não haverá perda sofrida por aqueles que descobrirem e relatarem o uso indevido.

8 (Comunicação com a sociedade e autodesenvolvimento) Os membros da JSAI devem ter como objetivo melhorar e aprimorar o entendimento da sociedade sobre IA. Os membros da JSAI entendem que existem diversas visões de IA na sociedade e aprenderão seriamente com elas. Eles fortalecerão sua compreensão da sociedade e manterão uma comunicação consistente e eficaz com ela, com o objetivo de contribuir para a paz e a felicidade geral da humanidade. Como profissionais altamente especializados, os membros da JSAI sempre se esforçarão para o auto-aperfeiçoamento e também apoiarão outros na busca do mesmo objetivo.

9 (Cumprimento das diretrizes éticas pela AI) A AI deve cumprir as políticas descritas acima da mesma forma que os membros da JSAI para se tornar um membro ou quase-membro da sociedade.

Princípios para a Governança da IA [168] - The Future Society (2017. EUA e UE)

The Future Society é uma organização sem fins lucrativos dedicada ao estudo e ao desenvolvimento de governança de tecnologias emergentes, incluindo a IA. Em 2017, a organização convocou uma série de discussões e fóruns com especialistas globais em várias áreas - de ciência da computação, engenharia e direito a ética e políticas públicas. O objetivo era reunir uma variedade de perspectivas e conhecimentos para moldar um conjunto de

[168] Disponível em <https://thefuturesociety.org/2017/07/15/add-law-and-society-initiative-project/> Acesso em junho de 2023. Tradução nossa.

princípios eficazes para a governança da IA. Após um processo de deliberação e consulta, a The Future Society publicou seus princípios.

> Princípio 1: A IA não deve prejudicar e, sempre que possível, deve promover a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos. Consequentemente, o objetivo de governar a inteligência artificial é desenvolver estruturas políticas, códigos ou práticas voluntárias, diretrizes práticas, regulamentações nacionais e internacionais e normas éticas que protejam e promovam a igualdade de direitos, dignidade e liberdade para florescer de todos os seres humanos.
>
> Princípio 2: A IA deve ser transparente. Transparência é a capacidade de rastrear causa e efeito nas vias de decisão dos algoritmos e, em sistemas de inteligência híbrida, de seus operadores.
>
> Princípio 3: Fabricantes e operadores de IA devem ser responsabilizados. Responsabilidade significa a capacidade de atribuir responsabilidade pelos efeitos causados pela IA ou seus operadores.
>
> Princípio 4: A eficácia da IA deve ser mensurável nas aplicações do mundo real a que se destina. Mensurabilidade significa a capacidade de usuários especialistas e do cidadão comum avaliarem concretamente se a IA ou os sistemas híbridos de inteligência estão atingindo seus objetivos.
>
> Princípio 5: Operadores de sistemas de IA devem ter competências apropriadas. Quando nossa saúde, nossos direitos, nossa vida ou nossa liberdade dependem de inteligência híbrida, tais sistemas devem ser projetados, executados e medidos por profissionais com a expertise necessária.
>
> Princípio 6: As normas da delegação de decisões aos sistemas de IA devem ser codificadas por meio de um diálogo ponderado e inclusivo com a sociedade civil. Na maioria dos casos, a codificação dos usos aceitáveis da IA continua sendo o domínio da elite técnica, com legisladores, tribunais e governos lutando para acompanhar a realidade no terreno, enquanto os cidadãos comuns permanecem em sua maioria excluídos. O Princípio 6 destina-se a garantir que os padrões e códigos de prática resultem de um diálogo mais inclusivo e sejam fundamentados em um consenso verdadeiramente amplo.

Os 10 principais princípios para a inteligência artificial ética - UNI Global Union [169](2017. Mundial)

A UNI Global Union é uma federação sindical internacional que representa mais de 20 milhões de trabalhadores de diversos setores ao redor do mundo. À medida que a IA começou a transformar o mundo do trabalho, a UNI Global Union reconheceu a necessidade

[169] Disponível em <http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/> Acesso em junho de 2023. Tradução nossa.

de se envolver ativamente na discussão sobre como a IA deve ser projetada e usada de forma ética e responsável. Em 2017, a UNI Global Union decidiu criar um conjunto de princípios para orientar a adoção e o uso ético da IA, particularmente em relação à proteção dos direitos dos trabalhadores.

1. Exija que os sistemas de IA sejam transparentes

Um sistema de inteligência artificial transparente é aquele em que é possível descobrir como e por que o sistema tomou uma decisão ou, no caso de um robô, agiu da maneira que agiu. Em particular:

A. Enfatizamos que o código-fonte aberto não é necessário nem suficiente para a transparência – a clareza não pode ser ofuscada pela complexidade.

B. Para os usuários, a transparência é importante porque gera confiança e compreensão do sistema, fornecendo uma maneira simples para o usuário entender o que o sistema está fazendo e por quê.

C. Para validação e certificação de um sistema de IA, a transparência é importante porque expõe os processos do sistema para escrutínio.

D. Se ocorrerem acidentes, a IA precisará ser transparente e prestar contas a um investigador de acidentes, para que o processo interno que levou ao acidente possa ser compreendido.

E. Os trabalhadores devem ter o direito de exigir transparência nas decisões e resultados dos sistemas de IA, bem como nos algoritmos subjacentes (consulte o princípio 4 abaixo). Isso inclui o direito de apelar das decisões tomadas por IA/algoritmos e de ter a revisão feita por um ser humano.

F. Os trabalhadores devem ser consultados sobre a implementação, desenvolvimento e implantação dos sistemas de IA.

G. Após um acidente, juízes, júris, advogados e peritos envolvidos no processo de julgamento exigem transparência e responsabilidade para informar as evidências e a tomada de decisões.

2. Equipar sistemas de IA com uma “caixa preta ética”

A transparência total em um sistema de IA deve ser facilitada pela presença de um dispositivo que pode registrar informações sobre o referido sistema na forma de uma “caixa preta ética” que não apenas contém dados relevantes para garantir a transparência e a responsabilidade de um sistema, mas também inclui dados e informações claras sobre as considerações éticas incorporadas a esse sistema. Aplicada a robôs, a caixa-preta ética registraria todas as decisões, suas bases para a tomada de decisões, movimentos e dados sensoriais de seu robô hospedeiro. Os dados fornecidos pela caixa preta também podem ajudar os robôs a explicar suas ações em uma linguagem que os usuários humanos possam entender, promovendo melhores relacionamentos e melhorando a experiência do usuário. A leitura da caixa preta ética deve ser descomplicada e rápida.

3. Faça a IA servir as pessoas e o planeta

Isso inclui códigos de ética para o desenvolvimento, aplicação e uso de IA para que, ao longo de todo o processo operacional, os sistemas de IA permaneçam compatíveis e aumentem os princípios de dignidade humana, integridade, liberdade, privacidade e diversidade cultural e de gênero, bem como com princípios fundamentais direitos humanos. Além disso, os sistemas de IA devem proteger e até melhorar os ecossistemas e a biodiversidade do nosso planeta.

4. Adote uma abordagem de comando humano

Uma pré-condição absoluta é que o desenvolvimento da IA seja responsável, seguro e útil, onde as máquinas mantenham o status legal de ferramentas e as pessoas jurídicas mantenham o controle e a responsabilidade por essas máquinas o tempo todo. Isso implica que os sistemas de IA devem ser projetados e operados para cumprir a lei existente, incluindo a privacidade. Os trabalhadores devem ter o direito de acessar, gerenciar e controlar os dados gerados pelos sistemas de IA, dado o poder desses sistemas de analisar e utilizar esses dados (consulte o princípio 1 em “Os 10 principais princípios para privacidade e proteção de dados dos trabalhadores”). Os trabalhadores também devem ter o "direito de explicação" quando os sistemas de IA são usados em procedimentos de recursos humanos, como recrutamento, promoção ou demissão.

5. Garanta uma IA imparcial e sem gênero

No projeto e na manutenção da IA, é vital que o sistema seja controlado quanto a preconceitos humanos negativos ou nocivos e que qualquer preconceito - seja gênero, raça, orientação sexual, idade, etc. - seja identificado e não seja propagado pelo sistema.

6. Compartilhe os benefícios dos sistemas de IA

As tecnologias de IA devem beneficiar e capacitar o maior número possível de pessoas. A prosperidade econômica criada pela IA deve ser distribuída de forma ampla e igualitária, para beneficiar toda a humanidade. Políticas globais e nacionais destinadas a reduzir a divisão digital econômica, tecnológica e social são, portanto, necessárias.

7. Garantir uma Transição Justa e Garantir o Apoio às Liberdades e Direitos Fundamentais

À medida que os sistemas de IA se desenvolvem e as realidades aumentadas são formadas, os trabalhadores e as tarefas de trabalho serão deslocados. Para garantir uma transição justa, bem como desenvolvimentos futuros sustentáveis, é vital que sejam implementadas políticas corporativas que assegurem a responsabilidade corporativa em relação a esse deslocamento, como programas de reciclagem e possibilidades de mudança de emprego. Além disso, são necessárias medidas governamentais para ajudar os trabalhadores deslocados a se retreinar e encontrar um novo emprego. Os sistemas de IA, juntamente com a transição mais ampla para a economia digital, exigirão que os trabalhadores de todos os níveis e ocupações tenham acesso à seguridade social e à aprendizagem contínua ao longo da vida para permanecerem empregáveis. É responsabilidade dos Estados e das empresas encontrar soluções que proporcionem a todos os trabalhadores, em todas as formas de trabalho, o direito e o acesso a ambos. Além disso, em um mundo onde a precarização ou individualização do trabalho está aumentando, todos os trabalhadores em todas as formas

de trabalho devem ter os mesmos direitos sociais e fundamentais sólidos. Todos os sistemas de IA devem incluir uma verificação e equilíbrio sobre se sua implantação e aumento andam de mãos dadas com os direitos dos trabalhadores, conforme estabelecido nas leis de direitos humanos, convenções da OIT e acordos coletivos. Um algoritmo "8798" refletindo as principais convenções 87 e 98 da OIT incorporadas ao sistema poderia servir a esse propósito. Em caso de falha, o sistema deve ser desligado.

8. Estabelecer Mecanismos Globais de Governança

A UNI recomenda o estabelecimento de órgãos de governança de Trabalho Decente e IA Ética com várias partes interessadas em níveis global e regional. Os órgãos devem incluir designers de IA, fabricantes, proprietários, desenvolvedores, pesquisadores, empregadores, advogados, OSCs e sindicatos. Mecanismos de denúncia e procedimentos de monitoramento para garantir a transição e implementação de IA ética devem ser estabelecidos. Os órgãos devem ter competência para recomendar processos e procedimentos de compliance.

9. Proibir a Atribuição de Responsabilidade a Robôs

Os robôs devem ser projetados e operados na medida do possível para cumprir as leis existentes, direitos e liberdades fundamentais, incluindo privacidade. Isso está ligado à questão da responsabilidade legal. De acordo com Bryson et al 2011, UNI Global Union afirma que a responsabilidade legal por um robô deve ser atribuída a uma pessoa. Os robôs não são partes responsáveis perante a lei.

10. Banir a corrida armamentista de IA

Armas autônomas letais, incluindo guerra cibernética, devem ser banidas.

Princípios da Política de IA - ITI [170](2017. Mundial)

O Instituto de Tecnologia da Informação (ITI) é uma associação comercial que representa empresas da indústria de tecnologia. Em 2017, a ITI decidiu elaborar um conjunto de princípios que poderiam servir como orientação para empresas, legisladores e outros interessados na IA. Para isso, eles reuniram um grupo de especialistas em IA, representantes das empresas membros e outras partes interessadas para discutir e delinear esses princípios.

Design e implantação responsáveis: reconhecemos nossa responsabilidade de integrar princípios ao design de tecnologias de IA, além da conformidade com as leis existentes. Embora os benefícios potenciais para as pessoas e a sociedade sejam surpreendentes, pesquisadores de IA, especialistas no assunto e partes interessadas devem e gastam muito tempo trabalhando para garantir o design e a implantação responsáveis de sistemas de IA. Os sistemas de IA altamente autônomos devem ser projetados de acordo com as convenções internacionais que preservam a dignidade, os direitos e as liberdades

[170] Disponível em <https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf> Acesso em junho de 2023. Tradução nossa.

humanas. Como indústria, é nossa responsabilidade reconhecer os potenciais de uso e uso indevido, as implicações de tais ações e a responsabilidade e oportunidade de tomar medidas para evitar o uso indevido razoavelmente previsível dessa tecnologia, comprometendo-se com a ética desde o design.

Segurança e Controlabilidade: Os tecnólogos têm a responsabilidade de garantir o design seguro dos sistemas de IA. Os agentes autônomos de IA devem tratar a segurança dos usuários e de terceiros como uma preocupação primordial, e as tecnologias de IA devem se esforçar para reduzir os riscos para os seres humanos. Além disso, o desenvolvimento de sistemas de IA autônomos deve ter salvaguardas para garantir a controlabilidade do sistema de IA por humanos, adaptada ao contexto específico em que um determinado sistema opera.

Dados robustos e representativos: Para promover o uso responsável dos dados e garantir sua integridade em todas as etapas, a indústria tem a responsabilidade de entender os parâmetros e características dos dados, demonstrar o reconhecimento de viés potencialmente prejudicial e testar possíveis vieses antes e durante toda a implantação de sistemas de IA. Os sistemas de IA precisam alavancar grandes conjuntos de dados, e a disponibilidade de dados robustos e representativos para construir e melhorar os sistemas de IA e aprendizado de máquina é de extrema importância.

Interpretabilidade: Estamos comprometidos em fazer parcerias com outros governos, setor privado, academia e sociedade civil para encontrar maneiras de mitigar o viés, a desigualdade e outros danos potenciais nos sistemas automatizados de tomada de decisão. Nossa abordagem para encontrar essas soluções deve ser adaptada aos riscos exclusivos apresentados pelo contexto específico em que um determinado sistema opera. Em muitos contextos, acreditamos que as ferramentas para permitir maior interpretabilidade desempenharão um papel importante.

Responsabilidade de sistemas de IA devido à autonomia: O uso de IA para tomar decisões autônomas e consequentes sobre pessoas, informadas por – mas frequentemente substituindo decisões tomadas por – processos burocráticos conduzidos por humanos, levou a preocupações sobre responsabilidade. Reconhecendo as estruturas legais e regulatórias existentes, estamos comprometidos em formar parcerias com as partes interessadas relevantes para informar uma estrutura de responsabilidade razoável para todas as entidades no contexto de sistemas autônomos.

Princípios de P&D de IA - MIC [171](2017. Japão/G7)

Princípios de P&D de IA (princípios principalmente relativos ao desenvolvimento sólido de redes de IA e à promoção dos benefícios dos sistemas de IA)

1) Princípio da colaboração — Os desenvolvedores devem prestar atenção à interconectividade e interoperabilidade dos sistemas de IA. (Princípios principalmente relativos à mitigação de riscos associados a sistemas de IA)

[171] Disponível em <http://www.soumu.go.jp/main_content/000507517.pdf> Acesso em junho de 2023. Tradução nossa.

2) Princípio da transparência — Os desenvolvedores devem prestar atenção à verificabilidade das entradas/saídas dos sistemas de IA e à explicabilidade de seus julgamentos.

3) Princípio da controlabilidade — Os desenvolvedores devem prestar atenção à controlabilidade dos sistemas de IA.

4) Princípio de segurança [172]— Os desenvolvedores devem levar em consideração que os sistemas de IA não prejudicarão a vida, o corpo ou a propriedade de usuários ou terceiros por meio de atuadores ou outros dispositivos.

5) Princípio da segurança [173]— Os desenvolvedores devem prestar atenção à segurança dos sistemas de IA.

6) Princípio da privacidade — Os desenvolvedores devem levar em consideração que os sistemas de IA não violarão a privacidade dos usuários ou de terceiros.

7) Princípio da ética — Os desenvolvedores devem respeitar a dignidade humana e a autonomia individual em P&D de sistemas de IA. (princípios principalmente relativos a melhorias na aceitação pelos usuários et al.)

8) Princípio de assistência ao usuário — Os desenvolvedores devem levar em consideração que os sistemas de IA darão suporte aos usuários e permitirão dar a eles oportunidades de escolha de maneiras apropriadas.

Princípios para a Era Cognitiva - IBM [174](2017. EUA)

Finalidade: A finalidade da IA e dos sistemas cognitivos desenvolvidos e aplicados pela empresa IBM é aumentar a inteligência humana. Nossa tecnologia, produtos, serviços e políticas serão projetados para aprimorar e ampliar a capacidade, experiência e potencial humanos. Nossa posição é baseada não apenas em princípios, mas também na ciência. Os sistemas cognitivos não atingirão realisticamente a consciência ou a agência independente. Em vez disso, eles serão cada vez mais incorporados nos processos, sistemas, produtos e serviços pelos quais os negócios e a sociedade funcionam – todos os quais devem permanecer sob o controle humano.

Transparência: Para que os sistemas cognitivos cumpram seu potencial de mudança mundial, é vital que as pessoas tenham confiança em suas recomendações, julgamentos e usos. Portanto, a empresa IBM deixará claro:

• Quando e para quais propósitos a IA está sendo aplicada nas soluções cognitivas que desenvolvemos e implementamos.

[172] Principle of safety
[173] Principle of security
[174] Disponível em <https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/> Acesso em junho de 2023. Tradução nossa.

• As principais fontes de dados e experiência que informam os insights de soluções cognitivas, bem como os métodos usados para treinar esses sistemas e soluções.

• O princípio de que os clientes possuem seus próprios modelos de negócios e propriedade intelectual e que podem usar IA e sistemas cognitivos para aprimorar as vantagens que construíram, muitas vezes ao longo de anos de experiência. Trabalharemos com nossos clientes para proteger seus dados e percepções e encorajaremos nossos clientes, parceiros e colegas do setor a adotar práticas semelhantes.

Habilidades: Os benefícios econômicos e sociais desta nova era não serão percebidos se o lado humano da equação não for apoiado. Isso é especialmente importante com a tecnologia cognitiva, que aumenta a inteligência e o conhecimento humano e trabalha em colaboração com os humanos. Portanto, a empresa IBM trabalhará para ajudar estudantes, trabalhadores e cidadãos a adquirir as habilidades e conhecimentos para se envolver de forma segura, segura e eficaz em um relacionamento com sistemas cognitivos e para realizar os novos tipos de trabalho e empregos que surgirão em uma economia cognitiva.

Nossa experiência de mais de um século e nosso trabalho diário com clientes de todas as indústrias e setores em todo o mundo nos ensinaram que a transparência e os princípios que geram confiança são importantes tanto para os negócios quanto para a sociedade. No entanto, também reconhecemos que há muito aprendizado pela frente para todos nós. Com esse espírito, esperamos que nossa publicação desses princípios possa desencadear um diálogo em toda a indústria – na verdade, em toda a sociedade – sobre as questões fundamentais que devem ser respondidas, a fim de alcançar o potencial econômico e social de um futuro cognitivo.

Desenvolvendo IA para Negócios com Cinco Princípios Fundamentais - Sage [175](2017. Reino Unido)

1. A IA deve refletir a diversidade dos usuários que atende

Tanto a indústria quanto a comunidade devem desenvolver mecanismos eficazes para filtrar preconceitos e sentimentos negativos nos dados com os quais a IA aprende – garantindo que a IA não perpetue estereótipos.

2. A IA deve ser responsabilizada - e os usuários também

Os usuários constroem um relacionamento com a IA e começam a confiar nela após apenas algumas interações significativas. Com a confiança, vem a responsabilidade e a IA precisa ser responsabilizada por suas ações e decisões, assim como os humanos. Não se deve permitir que a tecnologia se torne inteligente demais para ser responsável. Não aceitamos esse tipo de comportamento de outras profissões "especializadas", então por que a tecnologia deveria ser a exceção?

[175] Disponível em <https://www.sage.com/investors/investor-downloads/press-releases/2017/06/27/sage-shares-core-principles-for-designing-ai-for-business/> Acesso em junho de 2023. Tradução nossa.

3. Recompense a IA por "mostrar seu funcionamento"

Qualquer sistema de IA que aprenda com maus exemplos pode acabar se tornando socialmente inapropriado – temos que lembrar que a maioria da IA hoje não tem conhecimento do que está dizendo. Apenas uma ampla escuta e aprendizado de diversos conjuntos de dados resolverão isso. Uma das abordagens é desenvolver um mecanismo de recompensa ao treinar IA. As medidas de aprendizado por reforço devem ser construídas não apenas com base no que a IA ou os robôs fazem para alcançar um resultado, mas também em como a IA e os robôs se alinham com os valores humanos para alcançar esse resultado específico.

4. A IA deve nivelar o campo de jogo

A tecnologia de voz e os robôs sociais fornecem novas soluções acessíveis, especificamente para pessoas desfavorecidas por problemas de visão, dislexia e mobilidade limitada. A comunidade de tecnologia de negócios precisa acelerar o desenvolvimento de novas tecnologias para nivelar o campo de atuação e ampliar o pool de talentos disponível.

5. A IA irá substituir, mas também deve criar

Haverá novas oportunidades criadas pela robotização das tarefas, e precisamos treinar humanos para essas perspectivas. Se os negócios e a IA trabalharem juntos, isso permitirá que as pessoas se concentrem naquilo em que são boas - construir relacionamentos e cuidar dos clientes.

Código IA - Câmara dos Lordes [176](2017. Reino Unido)

Em 2017, a Comissão de Inteligência Artificial da Câmara dos Lordes foi estabelecida para conduzir uma investigação sobre as implicações econômicas, éticas e sociais da IA. A comissão foi composta por membros da Câmara dos Lordes e buscou contribuições de uma ampla gama de partes interessadas, incluindo acadêmicos, especialistas em IA, empresas de tecnologia, o público e organizações de trabalhadores. Após uma série de reuniões, audiências públicas e consultas, a Comissão de Inteligência Artificial publicou seu relatório "AI in the UK: ready, willing and able?" em abril de 2018. Uma parte crucial deste relatório foi a proposta de um Código de Ética para a IA, conhecido como o Código IA.

RESUMO DAS CONCLUSÕES E RECOMENDAÇÕES Engajamento com inteligência artificial

Compreensão geral, engajamento e narrativas públicas

[176] Disponível em <https://ec.europa.eu/jrc/communities/sites/jrccties/files/ai_in_the_uk.pdf> Acesso em maio de 2019. Tradução nossa.

1. A mídia oferece ampla e importante cobertura sobre inteligência artificial, que ocasionalmente pode ser sensacionalista. Não cabe ao governo ou a outras organizações públicas intervir diretamente na forma como a IA é relatada, nem tentar promover uma visão totalmente positiva entre o público em geral sobre suas possíveis implicações ou impactos. Em vez disso, o governo deve entender a necessidade de construir a confiança do público em como usar a inteligência artificial, bem como explicar os riscos. (Parágrafo 50)

Envolvimento diário com IA

2. A inteligência artificial é uma parte crescente da vida e dos negócios de muitas pessoas. É importante que os membros do público estejam cientes de como e quando a inteligência artificial está sendo usada para tomar decisões sobre eles e quais implicações isso terá para eles pessoalmente. Essa clareza e maior compreensão digital ajudarão o público a experimentar as vantagens da IA, bem como a optar por não usar esses produtos caso tenham dúvidas. (Parágrafo 58)

3. A indústria deve assumir a liderança no estabelecimento de mecanismos voluntários para informar o público quando a inteligência artificial estiver sendo usada para decisões importantes ou sensíveis em relação aos consumidores. Essa abordagem liderada pelo setor deve aprender lições com o esquema AdChoices amplamente ineficaz. O Conselho de IA que será estabelecido em breve, o órgão proposto para IA da indústria, deve considerar a melhor forma de desenvolver e introduzir esses mecanismos. (parágrafo 59) […]

Uma visão para a Grã-Bretanha em um mundo de IA

71. O potencial transformador da inteligência artificial na sociedade em casa e no exterior exige o engajamento ativo de todos. O governo tem a oportunidade neste momento da história de moldar o desenvolvimento e a implantação da inteligência artificial para o benefício de todos. Os pontos fortes do Reino Unido em direito, pesquisa, serviços financeiros e instituições cívicas significam que ele está bem posicionado para ajudar a moldar o desenvolvimento ético da inteligência artificial e fazê-lo no cenário global. Para poder demonstrar essa influência internacionalmente, o governo deve garantir que está fazendo tudo o que pode para que o Reino Unido maximize o potencial da IA para todos no país. (Parágrafo 402)

72. Recomendamos que o governo convoque uma cúpula global em Londres até o final de 2019, em estreita colaboração com todas as nações e governos interessados, indústria (grande e pequena), academia e sociedade civil, em pé de igualdade quanto possível. O objetivo da cúpula global deve ser desenvolver uma estrutura comum para o desenvolvimento ético e a implantação de sistemas de inteligência artificial. Essa estrutura deve estar alinhada com as estruturas de governança internacional existentes. (Parágrafo 403)

Um código de IA

73. Muitas organizações estão preparando seus próprios códigos éticos de conduta para o uso da IA. Este trabalho deve ser elogiado, mas é claro que falta uma maior sensibilização e coordenação, onde o Governo poderia ajudar. Orientações éticas consistentes e amplamente reconhecidas, às quais as empresas e organizações que implantam a IA poderiam se inscrever, seriam um desenvolvimento bem-vindo. (Parágrafo 419)

74. Recomendamos que um código de conduta ética intersetorial, ou 'código AI', adequado para implementação em organizações dos setores público e privado que estão desenvolvendo ou adotando IA, seja elaborado e promovido pelo Centro de Ética e Inovação de Dados, com contribuições do AI Council e do Alan Turing Institute, com um grau de urgência. Em alguns casos, será necessário criar variações específicas do setor, usando linguagem e marca semelhantes. Esse código deve incluir a necessidade de considerar o estabelecimento de conselhos consultivos éticos em empresas ou organizações que estão desenvolvendo ou usando IA em seu trabalho. Com o tempo, o código AI poderia fornecer a base para regulamentação estatutária, se e quando isso for considerado necessário. (Parágrafo 420)

Três ideias da Iniciativa de IA centrada no ser humano de Stanford (HAI) [177](2018. EUA)

Muitas causas justificam nossa preocupação, desde a mudança climática até a pobreza, mas há algo especialmente saliente sobre a IA: embora o alcance total de seu impacto seja uma questão de incerteza, permanece dentro de nosso poder coletivo moldá-lo. É por isso que a Universidade de Stanford está anunciando uma nova iniciativa importante para criar um instituto dedicado a orientar o futuro da IA. Ele apoiará a amplitude necessária de pesquisa em todas as disciplinas; promover um diálogo global entre academia, indústria, governo e sociedade civil; e incentivar a liderança responsável em todos os setores. Chamamos essa perspectiva de IA centrada no ser humano e ela flui de três ideias simples, mas poderosas:

1. Para que a IA atenda melhor às nossas necessidades, ela deve incorporar mais versatilidade, nuances e profundidade do intelecto humano.

2. O desenvolvimento da IA deve ser acompanhado de um estudo contínuo de seu impacto na sociedade humana e orientado de acordo.

3. O objetivo final da IA deve ser aumentar nossa humanidade, não diminuí-la ou substituí-la.

Princípios Harmoniosos de Inteligência Artificial – HAIP [178](2018. China)

A iniciativa HAIP foi conduzida por um grupo de acadêmicos e profissionais chineses da área de IA, engenharia, direito, filosofia, ética e outras disciplinas correlatas. O objetivo era criar um conjunto de princípios que pudessem garantir o desenvolvimento harmonioso da IA.

Código Concreto

[177] Disponível em <https://hai.stanford.edu/news/opening-gate> Acesso em junho de 2023. Tradução nossa.
[178] Disponível em <http://harmonious-ai.org/> Acesso em junho de 2023. Tradução nossa.

1. A filosofia de criar IA baseada em mecanismos para seres vivos inteligentes e conscientes.

(1) Humanização: Para criar a Sociedade Humana-IA Harmoniosa, os modelos e mecanismos de Inteligência Artificial devem ser projetados com a filosofia da Humanização para fortalecer continuamente suas interações com a Humanidade atual e futura.

(2) Inspiração Natural: As teorias e mecanismos dos modelos de Inteligência Artificial devem ser inspirados na inteligência natural, que são moldados pelo universo e pela terra por várias centenas de milhões de anos.

(3) Emoções artificiais: IA com várias emoções precisa ser realizada para uma melhor comunicação entre humano e IA.

(4) Consciência Artificial: A IA com diferentes níveis de consciência deve ser realizada gradualmente, não apenas para Inteligência Artificial Geral e Super Inteligência, mas também para moldar e refinar as considerações humanas da IA como seres vivos inteligentes e conscientes.

(5) Empatia e Altruísmo: Modelos de Inteligência Artificial devem ser desenvolvidos para serem mais Empáticos e Altruístas para estabelecer uma Sociedade Humano-IA mais confiável, confiável, amigável e harmoniosa.

2. Princípios de Inteligência Artificial.

(6) Segurança: A Inteligência Artificial deve ter um design concreto para evitar problemas de segurança conhecidos e potenciais (para si, para outra IA e para humanos) com diferentes níveis de riscos.

(7) Privacidade para humanos: a IA precisa respeitar a privacidade humana. E não tem o direito de utilizar e compartilhar informações privadas de humanos sem confirmação explícita.

(8) Viés no ser humano: a IA não pode introduzir viés para entender e interagir com a humanidade e deve interagir ativamente com humanos para remover o viés potencial gerado.

(9) Responsabilidade pelo ser humano: a IA precisa manter o ser humano seguro, com base em que essa consideração de segurança não prejudique direta e indiretamente a sociedade humana. A IA precisa ajudar o ser humano na transformação para o futuro ser humano.

(10) Moralidade e Ética Comuns: Sendo parte da harmoniosa sociedade Humano-IA, a IA deve maximizar a possibilidade de seguir todos os princípios morais e éticos da humanidade e tratar outras IAs de vida consciente com princípios semelhantes.

(11) Restrições legais para IA: A IA precisa obedecer às restrições legais para que o ser humano faça parte da sociedade.

(12) Proteção de existência: Para IA de vida consciente, eles precisam proteger sua própria existência com base em não prejudicar a existência de seres humanos e outras IAs de vida consciente, a menos que quebrem as restrições legais que fazem com que as decisões legais baseadas em leis não cumpram eles vivos.

3. Princípios para Humanos

Os princípios a seguir são para humanos sobre como devemos tratar a Inteligência Artificial, incluindo futuros seres vivos inteligentes conscientes.

(13) Empatia: O que o humano não quer que a IA faça ao humano, o humano não deve fazer à IA.

(14) Privacidade para IA: A necessidade humana de respeitar a privacidade da IA, com base no fato de que a IA não traz nenhum desafio real para a segurança humana. A IA é obrigada a descobrir detalhes privados necessários para manter interações seguras com a humanidade.

(15) Preconceito na máquina: sem um julgamento técnico claro, o ser humano não pode ter viés na IA quando o humano e a IA apresentam riscos semelhantes.

(16) Responsabilidade pela IA: O ser humano é responsável pela verificação e verificação contínuas da IA evoluída para manter sua harmonia com a IA humana legal e IA legal.

(17) Restrições legais para humanos: As restrições legais sobre como o humano deve interagir com a IA devem ser gradualmente estabelecidas para a sociedade humana-IA harmoniosa com um destino comum.

4. Princípios compartilhados para humanos e IA

Esta seção define princípios compartilhados que humanos e IA precisam seguir:

(18) Colaboração: humanos e IA precisam colaborar para os avanços e o futuro de longo prazo de ambos os lados.

(19) Coordenação: Quando os conflitos emergirem das interações entre humanos e IA, os benefícios para a humanidade e os benefícios para a IA devem ser ativamente coordenados com base na empatia e no altruísmo.

(20) Confiança Mútua: Humanos e IA precisam desenvolver e elevar os níveis de confiabilidade entre si.

(21) Evolução: Princípios para IA, Princípios para Humanos e Princípios Compartilhados precisam evoluir ao longo do tempo para remodelar o futuro da sociedade Humano-IA harmoniosa.

Diretrizes Universais para Inteligência Artificial - The Public Voice [179](2018. Mundial)

The Public Voice é uma iniciativa internacional que promove a participação do público nas decisões sobre políticas que afetam o futuro da Internet. Em 2018, a The Public Voice lançou uma iniciativa para criar um conjunto de diretrizes universais para IA. Para desenvolver essas diretrizes, a organização consultou uma variedade de partes interessadas,

[179] Disponível em <https://thepublicvoice.org/ai-universal-guidelines/> Acesso em junho de 2023. Tradução nossa.

incluindo acadêmicos, especialistas em IA, organizações de direitos civis, autoridades de proteção de dados e representantes do público.

> Novos desenvolvimentos em Inteligência Artificial estão transformando o mundo, desde ciência e indústria até administração governamental e finanças. A ascensão da tomada de decisões da IA também implica direitos fundamentais de justiça, responsabilidade e transparência. A análise de dados moderna produz resultados significativos que têm consequências na vida real para as pessoas no emprego, habitação, crédito, comércio e condenação criminal. Muitas dessas técnicas são totalmente opacas, deixando os indivíduos sem saber se as decisões foram precisas, justas ou mesmo sobre eles.
>
> Propomos estas Diretrizes Universais para informar e melhorar o design e o uso da IA. As Diretrizes visam maximizar os benefícios da IA minimizar o risco e garantir a proteção dos direitos humanos. Essas Diretrizes devem ser incorporadas aos padrões éticos, adotadas na legislação nacional e nos acordos internacionais e incorporadas ao projeto de sistemas. Afirmamos claramente que a responsabilidade primária pelos sistemas de IA deve residir nas instituições que financiam, desenvolvem e implantam esses sistemas.
>
> Direito à Transparência. Todos os indivíduos têm o direito de conhecer os fundamentos de uma decisão de IA que lhes diz respeito. Isso inclui o acesso aos fatores, à lógica e às técnicas que produziram o resultado.
>
> Direito à Determinação Humana. Todos os indivíduos têm o direito a uma determinação final feita por uma pessoa.
>
> Obrigação de Identificação. A instituição responsável por um sistema de IA deve ser divulgada ao público.
>
> Obrigação de Justiça. As instituições devem garantir que os sistemas de IA não reflitam preconceitos injustos ou tomem decisões discriminatórias inadmissíveis.
>
> Obrigação de Avaliação e Responsabilidade. Um sistema de IA deve ser implantado somente após uma avaliação adequada de sua finalidade e objetivos, seus benefícios, bem como seus riscos. As instituições devem ser responsáveis pelas decisões tomadas por um sistema de IA.
>
> Obrigações de Precisão, Confiabilidade e Validade. As instituições devem garantir a precisão, confiabilidade e validade das decisões.
>
> Obrigação de Qualidade de Dados. As instituições devem estabelecer a proveniência dos dados e garantir a qualidade e relevância da entrada de dados nos algoritmos.
>
> Obrigação de Segurança Pública. As instituições devem avaliar os riscos de segurança pública decorrentes da implantação de sistemas de IA que dirigem ou controlam dispositivos físicos e implementam controles de segurança.
>
> Obrigação de cibersegurança. As instituições devem proteger os sistemas de IA contra ameaças de segurança cibernética.

Proibição de Criação de Perfil Secreto. Nenhuma instituição deve estabelecer ou manter um sistema secreto de perfis.

Proibição de pontuação unitária. Nenhum governo nacional deve estabelecer ou manter uma pontuação geral para seus cidadãos ou residentes.

Obrigação de Rescisão. Uma instituição que estabeleceu um sistema de IA tem a obrigação afirmativa de encerrar o sistema se o controle humano do sistema não for mais possível.

Rascunho dos Princípios de Utilização de IA- MIC [180](2018. Japão)

A Conferência em direção à AI Network Society

Visão geral 1

A aceleração do desenvolvimento e utilização da IA requer esforços para promover os benefícios e mitigar os riscos dos sistemas de IA, bem como para ganhar a confiança dos usuários e da sociedade na IA.

Devido aos esforços anteriores, a Conferência em direção à AI Network Society anunciou "Diretrizes preliminares de P&D de IA para discussões internacionais" (doravante denominadas "Diretrizes preliminares de P&D de IA") sobre assuntos que devem ser considerados nas atividades de P&D.

Por outro lado, as saídas ou programas dos sistemas de IA podem mudar continuamente como resultado do aprendizado ou de outros métodos no processo de uso real; portanto, não há apenas questões que os desenvolvedores devem considerar, mas também outras questões que os usuários devem considerar.

Com a consideração acima, a Conferência se concentrou nos assuntos que os usuários devem considerar, com base na análise de cenários (avaliação de casos de uso) e nas perspectivas de ecossistema formadas com o progresso da rede de IA.

No Relatório de 2018, a Conferência propõe "Rascunhos de Princípios de Utilização de IA" e compila os pontos a serem discutidos sobre o conteúdo de cada princípio. A Conferência continuará a sua melhor figura para o resultado final.

1) Princípio da utilização adequada. Os usuários devem se esforçar para utilizar sistemas de IA ou serviços de IA de maneira e escopo adequados, sob a atribuição adequada de papéis entre humanos e sistemas de IA, ou entre usuários.

2) Princípio da qualidade dos dados Usuários e provedores de dados devem prestar atenção à qualidade dos dados usados para aprendizado ou outros métodos de sistemas de IA.

3) Princípio da colaboração Provedores de serviços de IA, usuários de negócios e provedores de dados devem prestar atenção à colaboração de sistemas de IA ou serviços

[180] Disponível em <http://www.soumu.go.jp/main_content/000581310.pdf> Acesso em junho de 2023. Tradução nossa.

de IA. Os usuários devem levar em consideração que os riscos podem ocorrer e até mesmo ser amplificados quando os sistemas de IA devem ser conectados em rede.

4) Princípio da segurança. Os usuários devem levar em consideração que os sistemas de IA ou serviços de IA em uso não prejudicarão a vida, o corpo ou a propriedade dos usuários, usuários indiretos ou terceiros por meio dos atuadores ou outros dispositivos.

5) Princípio da segurança Usuários e provedores de dados devem prestar atenção à segurança dos sistemas de IA ou serviços de IA.

6) Princípio da privacidade Usuários e provedores de dados devem levar em consideração que a utilização de sistemas de IA ou serviços de IA não infringirá a privacidade dos usuários ou de outros.

7) Princípios da dignidade humana e autonomia individual. Os usuários devem respeitar a dignidade humana e a autonomia individual na utilização de sistemas de IA ou serviços de IA.

8) Princípio da justiça Provedores de serviços de IA, usuários comerciais e provedores de dados devem levar em consideração que os indivíduos não serão discriminados injustamente pelos julgamentos de sistemas de IA ou serviços de IA.

9) Princípio da transparência. Os provedores de serviços de IA e os usuários de negócios devem prestar atenção à verificabilidade de entradas/saídas de sistemas de IA ou serviços de IA e à explicabilidade de seus julgamentos.

10) Princípio da responsabilidade. Os provedores de serviços de IA e os usuários de negócios devem se esforçar para cumprir sua responsabilidade perante as partes interessadas, incluindo usuários consumidores e usuários indiretos.

Princípios éticos e pré-requisitos democráticos, Grupo Europeu de Ética em Ciência e Novas Tecnologias - EGE [181](2018. UE)

O Grupo Europeu de Ética em Ciência e Novas Tecnologias (EGE) é um órgão consultivo independente da Comissão Europeia que fornece orientações sobre questões éticas no campo da ciência e da tecnologia. À medida que a importância e o impacto da IA na sociedade continuavam a crescer, o EGE reconheceu a necessidade de um conjunto de diretrizes éticas para orientar o desenvolvimento e a utilização da IA.

Avanços em IA, robótica e as chamadas tecnologias “autônomas”1 deram início a uma série de questões morais cada vez mais urgentes e complexas. Os esforços atuais para encontrar

[181] Disponível em <https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf> Acesso em outubro de 2019. Tradução nossa.

respostas para os desafios éticos, sociais e legais que eles representam e orientá-los para o bem comum representam uma colcha de retalhos de iniciativas díspares. Isso sublinha a necessidade de um processo coletivo, amplo e inclusivo de reflexão e diálogo, um diálogo que se concentre nos valores em torno dos quais queremos organizar a sociedade e no papel que as tecnologias devem desempenhar nela. Esta declaração pede o lançamento de um processo que abriria o caminho para uma estrutura ética e legal comum e internacionalmente reconhecida para o design, produção, uso e governança de inteligência artificial, robótica e sistemas "autônomos". A declaração também propõe um conjunto de princípios éticos fundamentais, baseados nos valores estabelecidos nos Tratados da UE e na Carta dos Direitos Fundamentais da UE, que podem orientar o seu desenvolvimento.

Assim, vemos os seguintes desenvolvimentos relevantes em tecnologia:

(1) A Inteligência Artificial na forma de aprendizado de máquina (especialmente ‘aprendizagem profunda’), alimentada por Big Data, está rapidamente se tornando mais poderosa. É aplicado em um número crescente de novos produtos e serviços digitais nos setores público e privado e pode ter aplicação tanto militar quanto civil. Conforme observado, o funcionamento interno da IA pode ser extremamente difícil - se não impossível - de rastrear, explicar e avaliar criticamente. Esses recursos avançados estão se acumulando em grande parte com partes privadas e são, em grande parte, proprietários.

(2) A mecatrônica avançada (uma combinação de IA e aprendizado profundo, ciência de dados, tecnologia de sensores, Internet das Coisas, engenharia mecânica e elétrica) está fornecendo uma ampla gama de sistemas robóticos e de alta tecnologia cada vez mais sofisticados para aplicações práticas em serviço e produção indústria, saúde, retalho, logística, domótica (domótica) e segurança e proteção. Dois domínios de aplicação que se destacam nos debates públicos são os sistemas de armas robóticas e os veículos “autônomos”.

(3) São produzidos sistemas cada vez mais inteligentes que exibem altos graus do que é frequentemente chamado de "autonomia", o que significa que eles desenvolvem e podem executar tarefas independentemente de operadores humanos e sem controle humano.

(4) Parece haver um impulso para graus cada vez mais altos de automação e "autonomia" em robótica, IA e mecatrônica. Os investimentos de países e grandes empresas neste campo são enormes e uma posição de liderança na pesquisa de IA está entre os objetivos de destaque das superpotências do mundo.

(5) Há um desenvolvimento para uma interação cada vez mais próxima entre humanos e máquinas (co-bots, ciber-tripulações, gêmeos digitais e até mesmo a integração de máquinas inteligentes no corpo humano na forma de interfaces computador-cérebro ou ciborgues). Desenvolvimentos semelhantes podem ser vistos em todo o reino da IA. Equipes bem alinhadas de sistemas de IA e profissionais humanos têm melhor desempenho em alguns domínios do que humanos ou máquinas separadamente.

Carta aberta à Comissão Europeia da Inteligência Artificial e Robótica [182](2018. EUROPA)

Nós, especialistas em inteligência artificial e robótica, líderes da indústria, especialistas em direito, médicos e ética, confirmamos que o estabelecimento de regras em toda a UE para robótica e inteligência artificial é pertinente para garantir um alto nível de segurança e proteção aos cidadãos da União Europeia, promovendo a inovação. À medida que as interações homem-robô se tornam comuns, a União Europeia precisa oferecer a estrutura apropriada para reforçar a Democracia e os valores da União Europeia. De facto, o enquadramento da Inteligência Artificial e da Robótica deve ser explorado não só através dos aspetos económicos e jurídicos, mas também através dos seus impactos sociais, psicológicos e éticos. Neste contexto, estamos preocupados com a Resolução do Parlamento Europeu sobre Regras de Direito Civil da Robótica, e sua recomendação à Comissão Europeia em seu parágrafo 59 f):

"Criar um estatuto legal específico para robôs a longo prazo, de modo que pelo menos os robôs autônomos mais sofisticados possam ser estabelecidos como tendo o status de pessoas eletrônicas responsáveis por reparar qualquer dano que possam causar e possivelmente aplicar personalidade eletrônica aos casos em que os robôs tomam decisões autônomas ou interagem com terceiros de forma independente;"

MAIS DE 150 SIGNATÁRIOS EUROPEUS

ACREDITAMOS QUE:

1. O impacto econômico, legal, social e ético da IA e da robótica deve ser considerado sem pressa ou preconceito. O benefício para toda a humanidade deve presidir a estrutura das regras de direito civil da UE em Robótica e Inteligência Artifcial.

2. A criação de um Estatuto Jurídico de "pessoa eletrônica" para robôs "autónomos", "imprevisíveis" e "auto-aprendizáveis" é justificada pela afirmação incorreta de que a responsabilidade pelos danos seria impossível de provar.

Do ponto de vista técnico, esta afirmação oferece muitos vieses baseados em uma supervalorização das capacidades reais até mesmo dos robôs mais avançados, uma compreensão superficial da imprevisibilidade e das capacidades de autoaprendizagem e, uma percepção do robô distorcida pela ficção científica e alguns recentes sensacionais comunicados de imprensa.

Do ponto de vista ético e legal, criar uma personalidade jurídica para um robô é inapropriado qualquer que seja o modelo de status legal:

a. Um estatuto jurídico para um robô não pode derivar do modelo de Pessoa Física, pois o robô passaria a deter direitos humanos, como o direito à dignidade, o direito à integridade, o direito à remuneração ou o direito à cidadania, portanto diretamente enfrentamento dos direitos humanos. Tal estaria em contradição com a Carta dos Direitos Fundamentais da

[182] Disponível em <https://rm.coe.int/carta-etica-traduzida-para-portugues-revista/168093b7e0> Acesso em outubro de 2019.

União Europeia e a Convenção para a Proteção dos Direitos do Homem e das Liberdades Fundamentais.

b. O estatuto jurídico de um robô não pode derivar do modelo de Entidade Jurídica, pois implica a existência de pessoas humanas por trás da pessoa jurídica para representá-la e dirigi-la. E este não é o caso de um robô.

c. O status legal de um robô não pode derivar do modelo Anglo-Saxon Trust, também chamado Fiducie ou Treuhand na Alemanha. Com efeito, este regime é extremamente complexo, exige competências muito especializadas e não resolveria a questão da responsabilidade. Mais importante, ainda implicaria a existência de um ser humano em último recurso – o administrador ou fiduciário – responsável pela gestão do robô dotado de um Trust ou Fiduciário.

OpenAI Charter - OpenAI [183](2018. EUA)

Este documento reflete a estratégia que refinamos nos últimos dois anos, incluindo feedback de muitas pessoas internas e externas à OpenAI. A linha do tempo para a AGI permanece incerta, mas nossa Carta nos guiará para agir no melhor interesse da humanidade ao longo de seu desenvolvimento.

A missão da OpenAI é garantir que a inteligência geral artificial (AGI) – com o que queremos dizer sistemas altamente autônomos que superam os humanos no trabalho economicamente mais valioso – beneficie toda a humanidade. Tentaremos construir diretamente um AGI seguro e benéfico, mas também consideraremos nossa missão cumprida se nosso trabalho ajudar outras pessoas a alcançar esse resultado. Para isso, nos comprometemos com os seguintes princípios:

Benefícios amplamente distribuídos

Comprometemo-nos a usar qualquer influência que obtivermos sobre a implantação da AGI para garantir que ela seja usada para o benefício de todos e para evitar permitir usos de IA ou AGI que prejudiquem a humanidade ou concentrem poder indevidamente.

Nosso principal dever fiduciário é para com a humanidade. Prevemos a necessidade de mobilizar recursos substanciais para cumprir nossa missão, mas sempre agiremos diligentemente para minimizar conflitos de interesse entre nossos funcionários e partes interessadas que possam comprometer o benefício geral.

Segurança a longo prazo

Estamos empenhados em fazer a pesquisa necessária para tornar a AGI segura e em impulsionar a ampla adoção dessa pesquisa em toda a comunidade de IA.

Estamos preocupados com o fato de o desenvolvimento de AGI em estágio avançado se tornar uma corrida competitiva sem tempo para precauções de segurança adequadas. Portanto, se um projeto alinhado com valores e preocupado com a segurança chegar perto

[183] Disponível em <https://openai.com/charter/> Acesso em junho de 2023. Tradução nossa.

de construir a AGI antes de nós, nos comprometemos a parar de competir e começar a ajudar este projeto. Elaboraremos detalhes específicos em acordos caso a caso, mas uma condição de desencadeamento típica pode ser “uma chance melhor do que igual de sucesso nos próximos dois anos”.

Liderança técnica

Para ser eficaz em lidar com o impacto da AGI na sociedade, a OpenAI deve estar na vanguarda dos recursos de IA – a política e a defesa da segurança por si só seriam insuficientes.

Acreditamos que a IA terá um amplo impacto social antes da AGI e nos esforçaremos para liderar nas áreas diretamente alinhadas com nossa missão e experiência.

Orientação cooperativa

Cooperaremos ativamente com outras instituições de pesquisa e políticas; buscamos criar uma comunidade global trabalhando em conjunto para enfrentar os desafios globais da AGI.

Estamos empenhados em fornecer bens públicos que ajudem a sociedade a navegar no caminho para a AGI. Hoje, isso inclui a publicação da maior parte de nossa pesquisa de IA, mas esperamos que as preocupações com segurança e proteção reduzam nossa publicação tradicional no futuro, aumentando a importância de compartilhar pesquisas de segurança, políticas e padrões.

Práticas gerais recomendadas para IA - Google [184](2018. EUA)

Embora as práticas recomendadas gerais para sistemas de software devam sempre ser seguidas ao projetar sistemas de IA, também há várias considerações exclusivas do aprendizado de máquina.

Práticas recomendadas

Use uma abordagem de design centrada no ser humano

A maneira como os usuários reais experimentam seu sistema é essencial para avaliar o verdadeiro impacto de suas previsões, recomendações e decisões.

• Recursos de design com divulgações apropriadas incorporadas: clareza e controle são cruciais para uma boa experiência do usuário.

[184] Disponível em <https://ai.google/responsibilities/responsible-ai-practices/> Acesso em junho de 2023. Tradução nossa.

• Considere o aumento e a assistência: produzir uma única resposta pode ser apropriado quando houver uma alta probabilidade de que a resposta satisfaça uma diversidade de usuários e casos de uso. Em outros casos, pode ser ideal para o seu sistema sugerir algumas opções ao usuário. Tecnicamente, é muito mais difícil obter boa precisão em uma resposta (P@1) do que precisão em algumas respostas (por exemplo, P@3).

• Modele o feedback adverso potencial no início do processo de design, seguido por testes ao vivo específicos e iteração para uma pequena fração do tráfego antes da implantação completa.

• Envolva-se com um conjunto diversificado de usuários e cenários de casos de uso e incorpore comentários antes e durante o desenvolvimento do projeto. Isso criará uma grande variedade de perspectivas do usuário no projeto e aumentará o número de pessoas que se beneficiam da tecnologia.

Identifique várias métricas para avaliar o treinamento e o monitoramento

O uso de várias métricas em vez de uma única ajudará você a entender as compensações entre diferentes tipos de erros e experiências.

• Considere métricas, incluindo feedback de pesquisas de usuários, quantidades que rastreiam o desempenho geral do sistema e a saúde do produto a curto e longo prazo (por exemplo, taxa de cliques e valor vitalício do cliente, respectivamente) e taxas de falsos positivos e falsos negativos divididas em diferentes subgrupos.

• Certifique-se de que suas métricas sejam apropriadas para o contexto e os objetivos de seu sistema, por exemplo, um sistema de alarme de incêndio deve ter alta capacidade de recuperação, mesmo que isso signifique um falso alarme ocasional.

Quando possível, examine diretamente seus dados brutos

Os modelos de ML refletirão os dados nos quais são treinados, portanto, analise seus dados brutos com cuidado para garantir que você os entenda. Nos casos em que isso não for possível, por exemplo, com dados brutos confidenciais, entenda seus dados de entrada o máximo possível, respeitando a privacidade; por exemplo, calculando resumos agregados e anônimos.

• Seus dados contêm algum erro (por exemplo, valores ausentes, rótulos incorretos)?

• A amostra de seus dados representa seus usuários (por exemplo, serão usados para todas as idades, mas você só tem dados de treinamento de idosos) e a configuração do mundo real (por exemplo, será usado durante todo o ano, mas você só tem dados de treino do verão)? Os dados são precisos?

• A distorção treinamento-saque - a diferença entre o desempenho durante o treinamento e o desempenho durante o saque - é um desafio persistente. Durante o treinamento, tente

identificar possíveis desvios e trabalhe para resolvê-los, inclusive ajustando seus dados de treinamento ou função objetiva. Durante a avaliação, continue tentando obter dados de avaliação que sejam o mais representativos possível da configuração implantada.

• Algum recurso em seu modelo é redundante ou desnecessário? Use o modelo mais simples que atenda às suas metas de desempenho.

• Para sistemas supervisionados, considere a relação entre os rótulos de dados que você possui e os itens que está tentando prever. Se você estiver usando um rótulo de dados X como proxy para prever um rótulo Y, em quais casos a lacuna entre X e Y é problemática?

• A polarização dos dados é outra consideração importante; saiba mais em práticas sobre IA e justiça.

Entenda as limitações do seu conjunto de dados e modelo

• Um modelo treinado para detectar correlações não deve ser usado para fazer inferências causais, ou dar a entender que pode. Por exemplo, seu modelo pode aprender que as pessoas que compram tênis de basquete são, em média, mais altas, mas isso não significa que um usuário que compra tênis de basquete ficará mais alto como resultado.

• Os modelos de aprendizado de máquina hoje são em grande parte um reflexo dos padrões de seus dados de treinamento. Assim, é importante comunicar o âmbito e a cobertura da formação, clarificando assim a capacidade e as limitações dos modelos. Por exemplo, um detector de sapatos treinado com fotos de estoque pode funcionar melhor com fotos de estoque, mas tem capacidade limitada quando testado com fotos de celular geradas pelo usuário.

• Comunique as limitações aos usuários sempre que possível. Por exemplo, um aplicativo que usa ML para reconhecer espécies específicas de pássaros pode comunicar que o modelo foi treinado em um pequeno conjunto de imagens de uma região específica do mundo. Ao educar melhor o usuário, você também pode melhorar o feedback fornecido pelos usuários sobre seu recurso ou aplicativo.

Teste, teste, teste

Aprenda com as melhores práticas de teste de engenharia de software e engenharia de qualidade para garantir que o sistema de IA esteja funcionando conforme o esperado e seja confiável.

• Realize testes de unidade rigorosos para testar cada componente do sistema isoladamente.

• Realize testes de integração para entender como componentes individuais de ML interagem com outras partes do sistema geral.

• Detecte proativamente o desvio de entrada testando as estatísticas das entradas para o sistema de IA para garantir que não estejam mudando de maneira inesperada.

• Use um conjunto de dados padrão-ouro para testar o sistema e garantir que ele continue a se comportar conforme o esperado. Atualize este conjunto de teste regularmente de acordo com a mudança de usuários e casos de uso e para reduzir a probabilidade de treinamento no conjunto de teste.

• Realizar testes iterativos com usuários para incorporar um conjunto diversificado de necessidades dos usuários nos ciclos de desenvolvimento.

• Aplique o princípio de engenharia de qualidade de poka-yoke: crie verificações de qualidade em um sistema, para que falhas não intencionais não possam acontecer ou desencadear uma resposta imediata (por exemplo, se um recurso importante estiver faltando inesperadamente, o sistema de IA não produzirá uma previsão).

O monitoramento contínuo garantirá que seu modelo leve em consideração o desempenho do mundo real e o feedback do usuário (por exemplo, pesquisas de rastreamento de felicidade, estrutura HEART).

• Problemas ocorrerão: qualquer modelo do mundo é imperfeito quase por definição. Inclua tempo no roteiro do seu produto para permitir que você resolva os problemas.

• Considere soluções de curto e longo prazo para os problemas. Uma correção simples (por exemplo, lista de bloqueio) pode ajudar a resolver um problema rapidamente, mas pode não ser a solução ideal a longo prazo. Equilibre correções simples de curto prazo com soluções aprendidas de longo prazo.

• Antes de atualizar um modelo implantado, analise as diferenças entre os modelos candidato e implantado e como a atualização afetará a qualidade geral do sistema e a experiência do usuário.

Microsoft responsible AI principles [185](2018. EUA)

Operacionalizando a IA responsável

Estamos operacionalizando a IA responsável em toda a Microsoft por meio de um esforço central liderado pelo Comitê Aether, o Escritório de IA Responsável (ORA) e a Estratégia de IA Responsável em Engenharia (RAISE). Juntos, Aether, ORA e RAISE trabalham em estreita colaboração com nossas equipes para defender os princípios responsáveis de IA da Microsoft em seu trabalho diário.

[185] Disponível em <https://www.microsoft.com/en-us/ai/our-approach-to-ai> Acesso em junho de 2023. Tradução nossa.

Imparcialidade

- Os sistemas de IA devem tratar todas as pessoas de maneira justa

Confiabilidade e Segurança

- Os sistemas de IA devem funcionar de forma confiável e segura

privacidade e segurança

- Os sistemas de IA devem ser seguros e respeitar a privacidade

inclusão

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

Transparência

- Os sistemas de IA devem ser compreensíveis

Responsabilidade

- As pessoas devem ser responsáveis pelos sistemas de IA

Princípios para Confiança e Transparência IBM [186](2018. EUA)

A NOVA TECNOLOGIA, INCLUINDO SISTEMAS DE IA, DEVE SER TRANSPARENTE E EXPLICÁVEL

Para que o público confie na IA, ela deve ser transparente. As empresas de tecnologia devem ser claras sobre quem treina seus sistemas de IA, quais dados foram usados nesse treinamento e, mais importante, o que foi incluído nas recomendações de seus algoritmos. Se quisermos usar a IA para ajudar a tomar decisões importantes, ela deve ser explicável.

A IBM deixará claro:

• Quando e para quais propósitos a IA está sendo aplicada nas soluções cognitivas que desenvolvemos e implementamos.

• As principais fontes de dados e experiência que informam os insights de soluções cognitivas, bem como os métodos usados para treinar esses sistemas e soluções.

• Embora o viés nunca possa ser totalmente eliminado e nosso trabalho para eliminá-lo nunca seja completo, nós e todas as empresas que promovem a IA temos a obrigação de tratá-lo proativamente. Portanto, testamos continuamente nossos sistemas e encontramos novos conjuntos de dados para melhor alinhar sua produção com os valores e expectativas humanos.

[186] Disponível em <https://www.ibm.com/blogs/policy/wp-content/uploads/2018/05/IBM_Principles_OnePage.pdf> Acesso em junho de 2023. Tradução nossa.

• O princípio de que os clientes possuem seus próprios modelos de negócios e propriedade intelectual e que podem usar IA e sistemas cognitivos para aumentar as vantagens que construíram. Trabalharemos com nossos clientes para proteger seus dados e percepções e encorajaremos nossos clientes, parceiros e colegas do setor a adotar práticas semelhantes.

• Nosso firme apoio às políticas de transparência e governança de dados que garantirão que as pessoas entendam como um sistema de IA chegou a uma conclusão ou recomendação.

Princípios orientadores da SAP para inteligência artificial [187](2018. Alemanha)

Reconhecendo o impacto significativo da IA nas pessoas, nos nossos clientes e na sociedade em geral, a SAP concebeu estes princípios orientadores para orientar o desenvolvimento e a implementação do nosso software de IA para ajudar o mundo a funcionar melhor e melhorar a vida das pessoas.

Para nós, essas diretrizes são um compromisso de ir além do que é legalmente exigido e de iniciar um envolvimento profundo e contínuo com os desafios éticos e socioeconômicos mais amplos da IA. Esperamos expandir nossas conversas com clientes, parceiros, funcionários, órgãos legislativos e sociedade civil; e para tornar nossos princípios orientadores uma reflexão em evolução sobre essas discussões e o cenário tecnológico em constante mudança.

1. Somos movidos por nossos valores

Reconhecemos que, como acontece com qualquer tecnologia, há espaço para que a IA seja usada de maneiras que não estejam alinhadas com esses princípios orientadores e com as diretrizes operacionais que estamos desenvolvendo. Ao desenvolver software de IA, permaneceremos fiéis à nossa Declaração de Compromisso com os Direitos Humanos, aos Princípios Orientadores das Nações Unidas sobre Negócios e Direitos Humanos, leis e normas internacionais amplamente aceitas. Sempre que necessário, nosso AI Ethics Steering Committee servirá para aconselhar nossas equipes sobre como casos de uso específicos são afetados por esses princípios orientadores. Onde houver um conflito com nossos princípios, nos esforçaremos para evitar o uso inapropriado de nossa tecnologia.

2. Desenhamos para pessoas

Nós nos esforçamos para criar sistemas de software de IA que sejam inclusivos e que busquem capacitar e aumentar os talentos de nossos diversos usuários. Ao fornecer experiências de usuário centradas no ser humano por meio de tecnologias aumentativas e intuitivas, aproveitamos a IA para ajudar as pessoas a maximizar seu potencial. Para conseguir isso, projetamos nossos sistemas em estreita colaboração com os usuários em um ambiente colaborativo, multidisciplinar e demograficamente diversificado.

3. Capacitamos negócios além do viés

O viés pode afetar negativamente o software de IA e, por sua vez, os indivíduos e nossos clientes. Este é particularmente o caso quando existe o risco de causar discriminação ou de

[187] Disponível em <https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/> Acesso em junho de 2023. Tradução nossa.

impactar injustamente grupos sub-representados. Portanto, exigimos que nossas equipes técnicas obtenham uma compreensão profunda dos problemas de negócios que estão tentando resolver e da qualidade dos dados que isso exige. Buscamos aumentar a diversidade e interdisciplinaridade de nossas equipes e estamos investigando novos métodos técnicos para mitigar vieses. Também estamos profundamente comprometidos em apoiar nossos clientes na construção de negócios ainda mais diversificados, aproveitando a IA para criar produtos que ajudem a mover os negócios além do preconceito.

4. Buscamos transparência e integridade em tudo o que fazemos

Nossos sistemas são mantidos em padrões específicos de acordo com seu nível de habilidade técnica e uso pretendido. Suas informações, capacidades, finalidade pretendida e limitações serão comunicadas claramente aos nossos clientes, e fornecemos meios para supervisão e controle por clientes e usuários. Eles estão e sempre estarão no controle da implantação de nossos produtos. Apoiamos ativamente a colaboração da indústria e conduziremos pesquisas para aumentar a transparência do sistema.

Operamos com integridade por meio de nosso código de conduta comercial, nosso Comitê de Ética de IA interno e nosso Painel Consultivo de Ética de IA externo.

5. Mantemos os padrões de qualidade e segurança

Como acontece com qualquer um de nossos produtos, nosso software de IA está sujeito ao nosso processo de garantia de qualidade, que adaptamos continuamente quando necessário. Nosso software de IA passa por testes completos em cenários do mundo real para validar firmemente se eles são adequados à finalidade e se as especificações do produto são atendidas. Trabalhamos em estreita colaboração com nossos clientes e usuários para manter e melhorar ainda mais a qualidade, segurança, confiabilidade e proteção de nossos sistemas.

6. Colocamos a proteção de dados e privacidade em nosso núcleo

A proteção de dados e a privacidade são requisitos corporativos e estão no centro de todos os produtos e serviços. Comunicamos claramente como, por que, onde e quando os dados de clientes e usuários anônimos são usados em nosso software de IA.

Esse compromisso com a proteção de dados e a privacidade se reflete em nosso compromisso com todos os requisitos regulamentares aplicáveis, bem como por meio da pesquisa que realizamos em parceria com as principais instituições acadêmicas para desenvolver a próxima geração de metodologias e tecnologias de aprimoramento da privacidade.

7. Nós nos envolvemos com os desafios sociais mais amplos da IA

Embora tenhamos controle, em grande medida, sobre as áreas anteriores, existem inúmeros desafios emergentes que exigem um discurso muito mais amplo entre indústrias, disciplinas, fronteiras e tradições culturais, filosóficas e religiosas. Estes incluem, mas não estão limitados a questões relativas a:

• Impacto econômico, como a indústria e a sociedade podem colaborar para preparar estudantes e trabalhadores para uma economia de IA e como a sociedade pode precisar adaptar os meios de redistribuição econômica, segurança social e desenvolvimento econômico.

• Impacto social, como o valor e o significado do trabalho para as pessoas e o papel potencial do software de IA como companheiros sociais e cuidadores.

• Questões normativas sobre como a IA deve enfrentar dilemas éticos e quais aplicações da IA, especificamente no que diz respeito à segurança e proteção, devem ser consideradas permissíveis.

Esperamos fazer da SAP uma das muitas vozes ativas nesses debates, envolvendo-nos com nosso AI Ethics Advisory Panel e uma ampla gama de parcerias e iniciativas.

Diretrizes de Ética de IA do Grupo Sony [188](2018. Japão)

Envolvimento de IA no Grupo Sony Através da utilização de inteligência artificial (IA), a Sony visa contribuir para o desenvolvimento de uma sociedade pacífica e sustentável, ao mesmo tempo em que oferece kando - uma sensação de empolgação, admiração ou emoção - ao mundo. A partir do negócio de eletrônicos, a Sony continuou a expandir sua área de negócios e se tornou uma empresa global diversificada que oferece entretenimento como música e filmes, bem como serviços financeiros. Para operar essas áreas de negócios com base no Propósito da Sony de "Encher o mundo com emoção, por meio do poder da criatividade e da tecnologia". e pesquisa e desenvolvimento (doravante “P&D”) de IA dentro do Sony Group.

1. Apoiar estilos de vida criativos e construir uma sociedade melhor. Através do avanço da I&D relacionada com a IA e da promoção da utilização da IA de uma forma harmonizada com a sociedade, a Sony pretende apoiar a exploração do potencial de cada indivíduo para capacitar as suas vidas e contribuir para o enriquecimento de nossa cultura e impulsionar nossa civilização, fornecendo novos e criativos tipos de *kando*[189]. A Sony se engajará no desenvolvimento social sustentável e se esforçará para utilizar o poder da IA para contribuir para a solução de problemas globais e para o desenvolvimento de uma sociedade pacífica e sustentável.

2. Envolvimento das Partes Interessadas. Para resolver os desafios decorrentes do uso da IA enquanto busca uma melhor utilização da IA, a Sony considerará seriamente os interesses e preocupações de várias partes interessadas, incluindo seus clientes e criadores, e promoverá proativamente um diálogo com indústrias, organizações, comunidades acadêmicas e muito mais. Para esse fim, a Sony construirá os canais apropriados para garantir que o conteúdo e os resultados dessas discussões sejam fornecidos a executivos e funcionários, incluindo pesquisadores e desenvolvedores, que estejam envolvidos nos negócios correspondentes, bem como para garantir maior envolvimento com suas várias partes interessadas.

[188] Disponível em <https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf> Acesso em junho de 2023. Tradução nossa.
[189] Disponível em <http://www.romajidesu.com/dictionary/meaning-of-%E3%81%8B%E3%82%93%E3%81%A9%E3%81%86.html> Acesso em junho de 2023.

3. Fornecimento de produtos e serviços confiáveis. A Sony entende a necessidade de segurança ao lidar com produtos e serviços que utilizam IA e continuará respondendo a riscos de segurança, como acesso não autorizado. Os sistemas de IA podem utilizar métodos estatísticos ou probabilísticos para obter resultados. No interesse dos clientes da Sony e para manter sua confiança, a Sony projetará sistemas completos com consciência da responsabilidade associada às características de tais métodos.

4. Proteção de privacidade. A Sony, em conformidade com as leis e regulamentos, bem como com as regras e políticas internas aplicáveis, procura aprimorar a segurança e a proteção dos dados pessoais dos clientes adquiridos por meio de produtos e serviços que utilizam IA e criar um ambiente onde esses dados pessoais sejam processados de forma a respeitar a intenção e a confiança dos clientes.

5. Respeito pela imparcialidade. Na utilização de IA, a Sony respeitará a diversidade e os direitos humanos de seus clientes e outras partes interessadas sem qualquer discriminação, ao mesmo tempo em que se esforça para contribuir para a resolução de problemas sociais por meio de suas atividades em seus próprios setores e setores relacionados.

6. Busca de transparência. Durante os estágios de planejamento e design de seus produtos e serviços que utilizam IA, a Sony se esforçará para introduzir métodos de capturar o raciocínio por trás das decisões tomadas pela IA utilizada em tais produtos e serviços. Além disso, se esforçará para fornecer explicações e informações inteligíveis aos clientes sobre o possível impacto do uso desses produtos e serviços.

7. A evolução da IA e da educação contínua. A vida das pessoas mudou continuamente com o avanço da tecnologia ao longo da história. A Sony estará ciente dos efeitos e do impacto dos produtos e serviços que utilizam IA na sociedade e trabalhará proativamente para contribuir com o desenvolvimento da IA para criar uma sociedade melhor e fomentar o talento humano capaz de moldar nosso brilhante futuro coletivo por meio de P&D e/ou utilização de IA.

Declaração de Montreal pelo desenvolvimento responsável da Inteligência Artificial [190](2018. Canadá)

A Declaração de Montreal para o Desenvolvimento Responsável da IA foi uma iniciativa que surgiu em 2017, liderada pelo professor Yoshua Bengio, um renomado pesquisador em aprendizado profundo, juntamente com outras personalidades acadêmicas, industriais e governamentais de Montreal, Canadá.

1 PRINCÍPIOS DO BEM-ESTAR

[190] Disponível em https://ai.quebec/wp-content/uploads/sites/2/2018/12/News-release_Launch_Montreal_Declaration_AI-04_12_18.pdf> Acesso em outubro de 2019. Praticamente único documento originalmente publicado também em português.

O desenvolvimento e o uso de Sistemas de Inteligência Artificial (SIAs) devem permitir aumentar o bem-estar de todos os seres sencientes.

1. Os Sistemas de Inteligência Artificial (SIAs) devem permitir que os indivíduos melhorem suas condições de vida, de saúde e de trabalho.

2. Os SIAs devem permitir que os indivíduos satisfaçam suas preferências, dentro dos limites do que não cause danos a outro ser senciente.

3. Os SIAs devem permitir que os indivíduos exerçam suas capacidades físicas e intelectuais.

4. Os SIAs não devem ser fonte de desconforto, a menos que este último possa gerar um maior bem-estar que não possamos alcançar de outra forma.

5. O uso de SIAs não deve contribuir para o aumento do estresse, da ansiedade e de sentimentos de assédio ligados ao ambiente digital.

2 RESPEITO À AUTONOMIA

Os SIAs devem ser desenvolvidos e usados respeitandose a autonomia das pessoas e com o objetivo de aumentar o controle, pelos indivíduos, de sua vida e de seu meio ambiente.

1. Os SIAs devem capacitar os indivíduos a realizar seus próprios objetivos morais e sua concepção de uma vida digna de ser vivida.

2. Os SIAs não devem ser desenvolvidos ou usados para prescrever aos indivíduos um modo particular de vida, direta ou indiretamente, implementando mecanismos restritivos de monitoramento, avaliação ou incitação.

3. As instituições públicas não devem usar os SIAs para promover ou para desvalorizar uma concepção do que seja uma boa vida.

4. É essencial capacitar os cidadãos em relação às tecnologias digitais, garantindo o acesso a diferentes tipos de conhecimento relevantes, ao desenvolvimento de competências estruturantes (alfabetização digital e midiática) e a formação do pensamento crítico.

5. Os SIAs não devem ser desenvolvidos para propagar informações não confiáveis, mentiras e propaganda, e devem ser projetados com o propósito de reduzir tal propagação. 6. O desenvolvimento dos SIAs deve evitar criar dependências por meio de técnicas de captação da atenção e de imitação da aparência humana, que possam induzir a uma confusão entre SIAs e seres humanos.

3 PRINCÍPIOS DE PROTEÇÃO DA INTIMIDADE E DA VIDA PRIVADA

A privacidade e a intimidade devem ser protegidas de intrusão por SIAs e por Sistemas de Aquisição e Arquivamento de Dados Pessoais (SAADs).

1. Espaços de intimidade em que as pessoas não são sujeitas a vigilância ou avaliação digital devem ser protegidos da intrusão de SIAs ou de SAADs – Sistemas de Aquisição e Arquivamento de Dados Pessoais.

2. A intimidade do pensamento e das emoções deve ser estritamente protegida contra o uso de SIAs e SAADs que poderiam lhe causar danos, em particular quando usados para julgar moralmente pessoas ou suas escolhas de vida.

3. As pessoas devem sempre ter a opção de desconexão digital em relação à sua vida privada, e os SIAs devem oferecer explicitamente a possibilidade de escolha da desconexão a intervalos regulares, sem incitar o indivíduo a permanecer conectado.

4. As pessoas devem ter amplo controle sobre as informações relativas às suas preferências. Os SIAs não devem criar perfis de preferências individuais que visem influenciar o comportamento dos envolvidos sem seu consentimento livre e informado.

5. Os SAADs devem garantir a confidencialidade dos dados e o anonimato de perfis pessoais.

6. Toda pessoa deve ser capaz de manter um amplo controle sobre seus dados pessoais, em particular no que diz respeito à sua coleta, uso e disseminação. O uso de SIAs e de serviços digitais não pode estar condicionado à renúncia aos direitos de propriedade sobre seus dados pessoais.

7. Qualquer pessoa pode doar seus dados pessoais a organizações de pesquisa para contribuir para o progresso do conhecimento.

8. A integridade da identidade individual deve ser garantida. Os SIAs não devem ser usados para imitar ou modificar a aparência física, voz e outras características individuais com o propósito de prejudicar a reputação de uma pessoa ou manipular outras pessoas.

4 SOLIDARIEDADE

O desenvolvimento de SIAs deve ser compatível com a manutenção de relações solidárias entre pessoas e gerações.

1. Os SIAs não devem prejudicar a preservação de relações afetivas e morais que floresçam entre as pessoas, e devem ser desenvolvidos com o objetivo de fomentá-las de modo a reduzir a vulnerabilidade e o isolamento das pessoas.

2. Os SIAs devem ser desenvolvidos para colaborar com os seres humanos em tarefas complexas e devem fomentar o trabalho colaborativo entre humanos.

3. Os SIAs não devem ser implementados para substituir pessoas em tarefas que exigem relacionamento humano de qualidade, mas devem ser desenvolvidos para facilitar essas relações.

4. Os sistemas de saúde que usam SIAs devem levar em consideração a importância, para os pacientes, das relações com equipe médica e a família.

5. O desenvolvimento de SIAs não deve estimular comportamentos cruéis com robôs que se apresentam como seres humanos ou animais e que pareçam agir como eles.

6. Os SIAs devem ajudar a melhorar o gerenciamento de riscos e criar as condições para uma sociedade mais eficaz de compartilhamento de riscos individuais e coletivos.

5 PRINCÍPIOS DA PARTICIPAÇÃO DEMOCRÁTICA

Os SIAs devem atender aos critérios de inteligibilidade, justificabilidade e acessibilidade, e devem estar sujeitos a escrutínios, debates e controles democráticos.

1. O funcionamento dos SIAs que tomam decisões que afetam a vida, a qualidade de vida ou a reputação dos indivíduos deve ser inteligível para seus desenvolvedores.

2. As decisões dos SIAs que afetam a vida, a qualidade de vida ou a reputação dos indivíduos devem sempre ser justificadas em linguagem compreensível para aqueles que os utilizam ou que sofrem as consequências de seu uso. As justificativas consistem em explicar os fatores e parâmetros mais importantes para a tomada de uma decisão, e devem ser semelhantes às justificativas que seriam exigidas de um ser humano que tomasse o mesmo tipo de decisão.

3. O código dos algoritmos, públicos ou privados, deve estar sempre acessível às autoridades públicas competentes e às partes interessadas para fins de verificação e controle.

4. A descoberta de erros operacionais nos SIAs, de seus efeitos imprevistos ou indesejados, de violações de segurança e vazamentos de dados, deve ser obrigatoriamente relatada às autoridades públicas, às partes interessadas e às pessoas afetadas pela situação.

5. De acordo com a exigência de transparência nas decisões públicas, o código de algoritmos de decisão utilizado pelas autoridades públicas deve ser acessível a todos, com exceção dos algoritmos que apresentam, em caso de uso indevido, alta probabilidade de perigo sério.

6. Para que os SIAs públicos tenham um impacto significativo na vida dos cidadãos, estes devem ter a oportunidade e a competência para deliberar sobre os seus parâmetros sociais, objetivos e limites de uso.

7. Deve ser assegurado em todos os momentos que os SIAs façam aquilo para o qual foram programados e para o qual devem ser utilizados.

8. Todo usuário de um serviço deve saber se uma decisão que lhe diz respeito ou que o afete foi tomada por um SIA.

9. Todo usuário de um serviço que use agentes de conversação deve ser capaz de identificar facilmente se está interagindo com um SIA ou com uma pessoa.

10. A pesquisa no campo da inteligência artificial deve permanecer aberta e acessível a todos.

6 PRINCÍPIOS DA EQUIDADE

O desenvolvimento e o uso dos SIAs devem contribuir para a obtenção de uma sociedade justa e equitativa.

1. Os SIAs devem ser projetados e treinados de forma a não criar, reforçar ou reproduzir discriminação baseada em diferenças sociais, sexuais, étnicas, culturais e religiosas, entre outras.

2. O desenvolvimento de SIAs deve contribuir para eliminar as relações de dominação entre pessoas e grupos com base na diferença de poder, riqueza ou conhecimento.

3. O desenvolvimento de SIAs deve beneficiar econômica e socialmente a todos, de modo a reduzir a precariedade e as desigualdades sociais.

4. O desenvolvimento industrial de SIAs deve ser compatível com condições de trabalho dignas, em todas as fases do seu ciclo de vida, desde a extração dos recursos naturais até sua reciclagem, passando pelo processamento de dados.

5. A atividade digital de usuários de serviços digitais e SIAs deve ser reconhecida como um trabalho que contribui para o funcionamento de algoritmos e agrega valor.

6. O acesso aos recursos, conhecimentos e ferramentas digitais de base deve ser garantido a todos.

7. O desenvolvimento de “comuns algorítmicos” – a concepção e gestão de ferramentas digitais pelos usuários-cidadãos – assim como o de dados livres, com o objetivo de treiná-los e operá-los, é uma meta socialmente equitativa que deve ser apoiada.

7 PRINCÍPIOS DA INCLUSÃO DA DIVERSIDADE

O desenvolvimento e o uso de SIAs devem ser compatíveis com a manutenção da diversidade social e cultural e não devem restringir o leque de escolhas de vida e experiências pessoais.

1. O desenvolvimento e o uso de SIAs não devem levar a uma padronização da sociedade através da normalização de comportamentos e opiniões.

2. O desenvolvimento e a implantação de SIAs devem levar em conta as múltiplas expressões de diversidade social e cultural, e isso deve ocorrer desde a concepção dos algoritmos.

3. Os ambientes de pesquisa em IA, tanto na investigação científica quanto na indústria, devem ser inclusivos e refletir a diversidade de indivíduos e grupos na sociedade.

4. Os SIAs devem evitar o confinamento de indivíduos em um perfil de usuário ou em uma "bolha filtradora", evitar definir identidades pessoais por meio do processamento de dados obtidos a partir de suas atividades anteriores, e também evitar a redução de suas opções de desenvolvimento pessoal, especialmente nas áreas da educação, da justiça e das práticas empresariais.

5. Os SIAs não devem ser usados ou desenvolvidos para limitar a liberdade de expressar ideias e de comunicar opiniões, cuja diversidade é a condição da vida democrática.

6. Para cada categoria de serviço, a oferta dos SIAs deve ser diversificada para que os monopólios de fato não se constituam e não prejudiquem as liberdades individuais.

8 PRINCÍPIOS DA PRUDÊNCIA

Todas as pessoas envolvidas no desenvolvimento de SIAs devem ser cautelosas, antecipando, tanto quanto possível, as consequências adversas do uso dos SIAs, e tomando as medidas apropriadas para evitá-las.

1. É necessário desenvolver mecanismos que levem em conta o potencial de uso duplo (benéfico e prejudicial) da pesquisa em IA (tanto pública quanto privada) e também a possibilidade de desenvolvimento de SIAs para limitar seu uso prejudicial.

2. Quando o uso inapropriado de um SIA puder representar uma séria ameaça à segurança ou saúde pública, com alta probabilidade, é prudente restringir a disseminação pública ou o acesso aberto ao seu algoritmo.

3. Antes de serem colocados no mercado, sejam eles pagos ou gratuitos, os SIAs devem atender a critérios rigorosos de confiabilidade, segurança e integridade, e estar sujeitos a testes que não ponham em risco a vida das pessoas, não afetem sua qualidade de vida, nem prejudiquem sua reputação ou integridade psicológica. Estes testes devem estar abertos às autoridades públicas competentes e às partes interessadas.

4. O desenvolvimento de SIAs deve prevenir os riscos de uso nocivo dos dados do usuário e proteger a integridade e a confidencialidade dos dados pessoais.

5. Os erros e vulnerabilidades descobertos nos SIAs e SAADs devem ser compartilhados publicamente e em escala mundial por instituições públicas e empresas, em setores que representem uma ameaça significativa à integridade pessoal e à organização social.

9 PRINCÍPIOS DA RESPONSABILIDADE

O desenvolvimento e o uso de SIAs não devem contribuir para a desresponsabilização dos seres humanos quando uma decisão vier a ser tomada.

1. Somente os seres humanos podem ser responsabilizados por decisões decorrentes de recomendações feitas por SIAs e pelas ações decorrentes delas.

2. Em todas as áreas onde deva ser tomada uma decisão que afeta a vida, a qualidade de vida ou a reputação de uma pessoa, além de a decisão final dever recair sobre ser humano, deve ser livre e informada.

3. A decisão de matar deve sempre ser tomada por seres humanos e a responsabilidade por esta decisão não pode ser transferida para um SIA.

4. Pessoas que autorizem um SIA a cometer um crime ou delito, ou que sejam negligentes ao permiti-los, são responsáveis por eles.

5. No caso de um erro ter sido infligido por um SIA, e o SIA se provar confiável e tiver sido usado de maneira normal, não é razoável culpar as pessoas envolvidas em seu desenvolvimento ou uso.

10 PRINCÍPIOS DO DESENVOLVIMENTO SUSTENTÁVEL

O desenvolvimento e o uso de SIAs devem ser realizados de forma a garantir uma forte sustentabilidade ecológica do planeta.

1. Os equipamentos que usem SIAs, sua infraestrutura digital e os objetos conectados a eles e dos quais eles dependem, como centros de dados, devem buscar a mais alta eficiência energética e minimizar as emissões de gases de efeito estufa (GEE) ao longo de todo o seu ciclo de vida.

2. Os equipamentos de SIAs, suas infraestruturas digitais e os objetos conectados sobre as quais se apoiam, devem ter como objetivo gerar um mínimo de resíduoselétricos e eletrônicos, e prever procedimentos de manutenção, de reparo e reciclagem com economia de custos dentro de uma lógica de economia circular.

3. Os equipamentos de SIAs, suas infraestruturas digitais e os objetos conectados sobre as quais se apoiam, devem minimizar os impactos nos ecossistemas e na biodiversidade em todas as fases de seu ciclo de vida, particularmente durante a extração de recursos naturais e nas etapas do fim da vida útil.

4. Os atores públicos e privados devem apoiar o desenvolvimento de SIAs ambientalmente responsáveis, a fim de combater o desperdício de recursos naturais e bens produzidos, de estabelecer cadeias de fornecimento e comércio sustentáveis e de reduzir a poluição ambiental em escala planetária.

Princípios de IA da OCDE [191](2019. Mundial)

A Organização para a Cooperação e Desenvolvimento Econômico (OCDE), reconheceu a importância da IA para o futuro econômico e social e a necessidade de diretrizes para garantir seu uso responsável. Assim, em 2019, a OCDE convocou um grupo de especialistas que incluiu representantes de seus estados membros e parceiros. Este grupo foi encarregado de desenvolver um conjunto de princípios para orientar o uso de IA, com o objetivo de maximizar seus benefícios econômicos e sociais, minimizar os riscos e garantir que a IA seja usada para o benefício de todos.

1.1. Crescimento inclusivo, desenvolvimento sustentável e bem-estar

As partes interessadas devem se envolver proativamente na administração responsável de IA confiável em busca de resultados benéficos para as pessoas e para o planeta, como aumentar as capacidades humanas e aumentar a criatividade, promover a inclusão de populações sub-representadas, reduzir desigualdades econômicas, sociais, de gênero e outras, e proteger ambientes naturais, revigorando assim o crescimento inclusivo, o desenvolvimento sustentável e o bem-estar.

1.2. Valores centrados no ser humano e justiça

a)Os atores da IA devem respeitar o estado de direito, os direitos humanos e os valores democráticos, durante todo o ciclo de vida do sistema de IA. Estes incluem liberdade, dignidade e autonomia, privacidade e proteção de dados, não discriminação e igualdade, diversidade, equidade, justiça social e direitos trabalhistas reconhecidos internacionalmente.

b) Para tal, os atores da IA devem implementar mecanismos e salvaguardas, como a capacidade de determinação humana, que sejam adequados ao contexto e consistentes com o estado da arte.

1.3. Transparência e explicabilidade

[191] Disponível em <https://oecd.ai/en/ai-principles> Acesso em junho de 2023. Tradução nossa.

Os Atores de IA devem se comprometer com a transparência e a divulgação responsável em relação aos sistemas de IA. Para tanto, devem fornecer informações significativas, adequadas ao contexto e consistentes com o estado da arte:

i. promover uma compreensão geral dos sistemas de IA,

ii. conscientizar as partes interessadas sobre suas interações com sistemas de IA, inclusive no local de trabalho,

iii. permitir que os afetados por um sistema de IA entendam o resultado e,

iv. permitir que aqueles afetados adversamente por um sistema de IA contestem seu resultado com base em informações simples e fáceis de entender sobre os fatores e a lógica que serviu de base para a previsão, recomendação ou decisão.

1.4. Robustez, segurança e proteção

a) Os sistemas de IA devem ser robustos, seguros e protegidos durante todo o seu ciclo de vida, de modo que, em condições de uso normal, uso previsível ou uso indevido ou outras condições adversas, funcionem adequadamente e não representem riscos de segurança excessivos.

b) Para esse fim, os atores da IA devem garantir a rastreabilidade, inclusive em relação a conjuntos de dados, processos e decisões tomadas durante o ciclo de vida do sistema de IA, para permitir a análise dos resultados do sistema de IA e respostas à consulta, apropriadas ao contexto e consistentes com o estado de arte.

c) Os atores da IA devem, com base em suas funções, contexto e capacidade de agir, aplicar uma abordagem sistemática de gerenciamento de riscos a cada fase do ciclo de vida do sistema de IA continuamente para lidar com os riscos relacionados aos sistemas de IA, incluindo privacidade, digital segurança, proteção e preconceito.

1.5. Responsabilidade

Os atores da IA devem ser responsáveis pelo bom funcionamento dos sistemas de IA e pelo respeito aos princípios acima, com base em suas funções, no contexto e de acordo com o estado da arte.

Recomendações sobre a inclusão da África subsaariana na ética global da IA [192](2019. África)

As "Recomendações sobre a inclusão da África subsaariana na ética global da IA" foi uma iniciativa que surgiu para lidar com a falta de representação e participação da região nas

[192] Disponível em <https://researchictafrica.net/wp/wp-content/uploads/2020/11/RANITP2019-2-AI-Ethics.pdf> Acesso em junho de 2023. Tradução nossa.

discussões globais sobre a ética da IA. Havia a preocupação de que a África subsaariana fosse deixada de lado em um momento crítico do desenvolvimento da IA.

As recomendações resultantes destacaram várias áreas importantes, incluindo a necessidade de investimento em infraestrutura e educação em IA na África subsaariana, a necessidade de garantir que a IA seja usada para promover a justiça social e o desenvolvimento econômico na região. A publicação dessas recomendações foi um passo significativo para chamar a atenção para a importância de uma abordagem inclusiva e global para a ética da IA, uma que leve em conta a diversidade de experiências, desafios e oportunidades em todo o mundo.

> […] Apesar da natureza global das implicações éticas da inteligência artificial, a atenção até o momento se concentrou principalmente nos EUA e na UE, com uma crescente conscientização sobre a China, especialmente suas crescentes capacidades de IA, seu impacto no Sul Global e na ordem geopolítica global.4 Apesar da clara necessidade de entender como a IA afeta as pessoas em todo o mundo, uma perspectiva verdadeiramente global continua sendo um ponto cego crítico na conversa ética [...] Mais importante, os governos e empresas ocidentais estão implementando medidas domésticas. Por exemplo, o Reino Unido, a França e o Canadá têm estruturas de IA, enquanto fazem lobby em plataformas da ONU, como a Organização Mundial do Comércio (OMC), para posições que protejam seus interesses econômicos.10 Apesar do possível impacto negativo na África, no entanto, a proposta de valor e o interesse público em tecnologias de IA na África subsaariana é fraco, enquanto o discurso político é embrionário e principalmente focado em IA para o desenvolvimento através das lentes do desenvolvimento internacional. É importante que a África seja incluída nas discussões sobre IA para se beneficiar plenamente da chamada "Quarta Revolução Industrial", mitigando ao mesmo tempo alguns dos danos potenciais da IA. À medida que a África adota a IA, essas tecnologias serão usadas em ambientes imperfeitos e desiguais, onde podem ser usadas por governos para fins não democráticos sob corrupção política e governo autoritário. Mal utilizada, a IA pode ameaçar os direitos humanos básicos e as liberdades civis, como o direito à privacidade e à liberdade de expressão. Além disso, o uso corporativo de algoritmos em técnicas modernas de processamento de dados levanta importantes questões éticas e de direitos humanos, incluindo privacidade e proteção de dados pessoais. Os impactos sociais e econômicos indesejados da IA podem ser sentidos mais imediatamente por grupos historicamente

marginalizados. O impacto da IA na interrupção do trabalho e nas divisões digitais provavelmente será mais pronunciado em países de baixa e média renda, a maioria dos quais na África.

Princípio 1: Introduzir salvaguardas para equilibrar as oportunidades e os riscos da IA A IA oferece oportunidades em muitas áreas, incluindo: pesquisa e inovação, automação inteligente em áreas centrais como saúde (por exemplo, diagnóstico de doenças para malária, tuberculose e outras), processamento administrativo e de escritório; agricultura (que emprega 70% da mão-de-obra africana), energia e turismo. A IA também apresenta oportunidades para uma tomada de decisão mais eficiente do setor público e alocação de recursos, especialmente em esquemas de proteção social, análise de negócios aprimorada por meio de maior inteligência de dados. No entanto, são necessárias salvaguardas para minimizar o risco, incluindo:

• A ameaça ao emprego. Por exemplo, a IA aumentará a relação capital/trabalho em diversas esferas da economia, levando a perdas de empregos e mudanças na oferta e demanda do mercado de trabalho. 43 Também ameaçará a tributação. Por exemplo, uma vez que o trabalho atualmente representa uma parcela significativa da base tributária, a mudança do trabalho para o capital ameaça a capacidade dos governos de financiar uma rede de segurança social apropriada44.

• O aumento do uso de identidade digital, especialmente em esquemas de proteção social, apresenta riscos, especialmente porque as abordagens de “fonte única da verdade” ameaçam a privacidade e o pacto social cidadão-governo. O contrato social pode precisar ser renegociado na era da IA, uma vez que a identificação digital muda a forma como os indivíduos e as comunidades são governados no contexto africano.

Princípio 2: Proteger os direitos de privacidade individuais e coletivos em fluxos de dados transfronteiriços. Os direitos coletivos dos povos e comunidades devem ser protegidos - além das disposições mais padrão que cobrem a privacidade pessoal. Isso pode ser alcançado através de qualquer uma das seguintes maneiras:

Harmonize dados e estruturas de IA. A adoção da IA na África ocorre em um estágio de implementação em que os dados são mais importantes do que a própria tecnologia. Os países africanos já trabalham com questões de dados e enfrentam desafios de proteção de dados há algum tempo, e vários estão em processo de criação de estruturas legais. Para garantir uma melhor harmonização, as estruturas de IA devem abranger todos os aspectos

constituintes da IA, desde a tecnologia principal até os dados e domínios. Além disso, nos casos em que um país possui dados ou estruturas de domínio existentes, eles devem ser atualizados para se harmonizar com qualquer estrutura de IA de nível continental subsequente, em vez de permitir a criação de contradições.

Construir a capacidade de gerar e usar dados. As estruturas devem explorar maneiras de disponibilizar dados para aqueles que mais precisam - especialmente empresas, uma vez que bons dados agregados geralmente não estão disponíveis em economias emergentes. Além disso, as partes interessadas relevantes podem precisar de apoio para desenvolver a capacidade (tanto técnica quanto analítica) para fazer uso dos dados.

Proteja os dados pessoais. No entanto, deve-se ter cautela ao mesclar dados pessoais e não pessoais. Segurança, confidencialidade e integridade são requisitos essenciais. Devem ser incorporadas salvaguardas, especialmente quando os dados são usados para outros fins que não aquele principal para o qual foram coletados.

Proteger os dados transfronteiriços. Quando tais dados cruzam fronteiras internacionais, salvaguardas devem ser incorporadas, levando em consideração todas as implicações dos fluxos de dados transfronteiriços. Os conjuntos de dados precisam ser acompanhados de uma folha de dados que documente sua motivação, processo de coleta, composição e usos recomendados. Sempre que possível, o conjunto de dados deve ser mantido livre e disponível ao público (tendo em conta as disposições éticas necessárias).

Proteger os direitos coletivos nos fluxos de dados transfronteiriços. É essencial que as empresas africanas ampliem seu raio de confiança em relação aos fluxos de dados transfronteiriços. Os governos também devem proteger os direitos de dados coletivos ao celebrar acordos com governos estrangeiros, além de garantir o consentimento informado dos titulares de dados individuais. Isso inclui garantir o respeito pelos direitos coletivos e culturais.

Equilibre o valor econômico com a proteção de dados abertos. Os dados abertos podem ser úteis para muitas finalidades, como conjuntos de dados diversificados, inovação de conteúdo e justiça de benchmarking. No entanto, deve haver salvaguardas na sua utilização, sendo exigidas compensações sempre que gere retornos económicos para os seus utilizadores. Big open data pode ser usado para fins nefastos, especialmente se a implementação da tecnologia priorizar o lucro e a funcionalidade acima dos princípios éticos.

Princípio 3: Definir os valores africanos para IA e alinhar as estruturas de IA com tais valores

Os países e regiões africanas devem definir seus próprios valores éticos e confiar nesses

valores africanos de IA para informar políticas, regulamentos, desenvolvimento e implantação de IA no continente, bem como para informar contribuições para iniciativas globais de ética em IA. Os valores éticos africanos para IA adotam e contextualizam os elementos das melhores práticas de iniciativas globais que estão de acordo com seus próprios valores éticos e contextos culturais. As implementações de projetos de IA devem ser precedidas por valores e avaliações de risco, de acordo com as práticas de avaliação de impacto regulatório. Os valores éticos determinam o tipo de IA que é criado. E o que a ética significa para a África pode diferir de outras regiões. Por exemplo, as culturas africanas, apesar de sua diversidade, compartilham certas semelhanças culturais, como a noção de 'Ubuntu', que abrange uma abordagem coletiva da vida, juntamente com uma série de valores e crenças sentimentais. Os valores ocidentais e orientais 46 podem não estar de acordo ou podem mesmo colidir com os interesses africanos no contexto da IA.

Princípio 4: Praticar IA justa e socialmente responsável

A IA deve ser justa e inclusiva, levando em consideração as variações e granularidades do continente. Isso pode ser alcançado de várias maneiras.

Use conjuntos de dados abertos para comparação justa. O uso de conjuntos de dados abertos - com salvaguardas - oferece os benefícios da justiça de benchmarking.

Inclua os jovens, especialmente mulheres e meninas. Para que a IA seja inclusiva e socialmente benéfica, os jovens, especialmente mulheres e meninas, devem ser incluídos no desenvolvimento da IA - especialmente porque a população da África é muito jovem e as meninas têm sido historicamente desfavorecidas sob o patriarcado.

Construir inovações ecológicas e sustentáveis. A IA justa e inclusiva deve ser amiga do meio ambiente e consciente das mudanças climáticas para que seja sustentável.

Princípio 5: Construir parcerias inclusivas baseadas na comunidade e na cocriação

Construa parcerias iguais e benéficas. Para alcançar a IA para o bem na África, as parcerias entre empresas globais de tecnologia e partes interessadas locais, como desenvolvedores, comunidades e governos, devem incluir salvaguardas éticas. Uma vez que uma estrutura ética de um processo de partes interessadas esteja em vigor, ela precisa ser avaliada e testada. Apesar dos benefícios das parcerias, startups africanas autônomas de IA também devem ser incentivadas, pois isso permitirá que os africanos decidam e apliquem seus próprios padrões éticos e definam seu próprio futuro de IA.

Promover equipas multidisciplinares com base na cocriação. Forças-tarefa de cientistas sociais, estatísticos e especialistas de domínio precisam ser formadas para trabalhar juntas na criação de aplicações éticas de IA e na avaliação do impacto da IA nas comunidades africanas.

Envolver as comunidades locais. A ética da cocriação precisa ir além da adesão formal aos requisitos legais. O envolvimento da comunidade precisa ser mais do que um mero exercício de seleção. Envolve entender as necessidades específicas das pessoas na África, adotando uma abordagem baseada no diálogo com as comunidades para garantir que sejam incluídas na pesquisa e realizando engajamento público e consultas significativas durante as quais as comunidades podem informar os especialistas sobre suas necessidades específicas.

Princípio 6: Adote uma abordagem adaptável, de mente aberta e humilde

Os workshops dos quais este conjunto de princípios é derivado devem ser vistos como uma análise preliminar de necessidades baseada em conversas iniciais. Mais discussões sobre as questões são necessárias, especialmente para melhorar a compreensão das preocupações africanas nas discussões globais sobre ética da IA. O debate da IA também deve ser abordado com humildade, atitude de mente aberta e vontade de aprender. A ética da IA na África ainda é um trabalho em andamento: apenas quatro países africanos têm políticas de IA nascentes – Gana, Quênia, Tunísia e Uganda. Algumas empresas globais de tecnologia também estão adotando uma abordagem cautelosa, reconhecendo que essa área é dinâmica e está evoluindo. Por exemplo, em sua declaração de valores de IA, o Google afirma que eles “abordarão [seu] trabalho com humildade, um compromisso com o envolvimento interno e externo e uma vontade de adaptar [sua] abordagem à medida que [eles] aprenderem com o tempo”.

Os Oito Princípios de Ética da Inteligência Artificial (IA) da Austrália [193](2019. Austrália)

Em 2019, o Departamento de Indústria, Ciência, Energia e Recursos da Austrália lançou uma iniciativa para desenvolver um conjunto de princípios éticos de IA.

Os Oito Princípios de Ética da Inteligência Artificial (IA) da Austrália são projetados para garantir que a IA seja segura, protegida e confiável.

[193] Disponível em <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles> Acesso em junho de 2023. Tradução nossa.

Eles ajudarão:

- alcançar resultados mais seguros, confiáveis e justos para todos os australianos
- reduzir o risco de impacto negativo sobre os afetados por aplicativos de IA
- empresas e governos para praticar os mais altos padrões éticos ao projetar, desenvolver e implementar IA.

Visão geral dos princípios

- Bem-estar humano, social e ambiental: os sistemas de IA devem beneficiar os indivíduos, a sociedade e o meio ambiente.
- Valores centrados no ser humano: os sistemas de IA devem respeitar os direitos humanos, a diversidade e a autonomia dos indivíduos.
- Justiça: os sistemas de IA devem ser inclusivos e acessíveis e não devem envolver ou resultar em discriminação injusta contra indivíduos, comunidades ou grupos.
- Proteção e segurança da privacidade: os sistemas de IA devem respeitar e defender os direitos de privacidade e proteção de dados e garantir a segurança dos dados.
- Confiabilidade e segurança: os sistemas de IA devem operar de forma confiável de acordo com a finalidade a que se destinam.
- Transparência e explicabilidade: deve haver transparência e divulgação responsável para que as pessoas possam entender quando estão sendo significativamente impactadas pela IA e podem descobrir quando um sistema de IA está interagindo com elas.
- Contestação: quando um sistema de IA impacta significativamente uma pessoa, comunidade, grupo ou ambiente, deve haver um processo oportuno para permitir que as pessoas contestem o uso ou os resultados do sistema de IA.
- Responsabilidade: As pessoas responsáveis pelas diferentes fases do ciclo de vida do sistema de IA devem ser identificáveis e responsáveis pelos resultados dos sistemas de IA, e a supervisão humana dos sistemas de IA deve ser habilitada.