Polymatheia
Revista de Filosofia

COMO PARÂMETROS COMPUTACIONAIS PODEM LIMITAR O ESCOPO DE DECISÕES ÉTICAS EM INTELIGÊNCIAS ARTIFICIAIS?

Bruno de Brito Rosa*

Resumo: Os Ataques Adversários (*Adversarial Attacks*) suscitam discussões e investimentos na área de Tecnologia da Informação. Porém, podem provocar problemas éticos profundos, pois esses Ataques Adversários desencadeiam alucinações em máquinas equipadas com Inteligências Artificiais Generativas por meio da entrada (*input*) de dados adulterados ou apenas mal interpretados. O objetivo é fazer um levantamento bibliográfico, não somente dos aspectos técnicos, mas também de propostas de implementações éticas em Inteligências Artificiais. Os resultados indicam que os maiores desafios estão envoltos nos possíveis danos irreparáveis decorrentes das distorções que os Ataques Adversários acarretam, exigindo novas formas de interpelar o problema ético em Inteligências Artificiais.

Palavras-chave: Filosofia da informação, Ética normativa, Inteligência Artificial, Ataques Adversários, Metaética.

HOW CAN COMPUTATIONAL PARAMETERS LIMIT THE SCOPE OF ETHICAL DECISIONS IN ARTIFICIAL INTELLIGENCE?

Abstract: Adversarial Attacks spark discussion and investment in the field of Information Technology. However, they can cause profound ethical problems, as these Adversarial Attacks trigger hallucinations in machines equipped with Generative Artificial Intelligence through the input of adulterated or just misinterpreted data. The aim is to carry out a bibliographical survey, not only of the technical aspects, but also of proposals for ethical implementations in Artificial Intelligence. The results indicate that the greatest challenges lie in the possible irreparable damage resulting from the distortions caused by Adversary Attacks, requiring new ways of addressing the ethical problem in Artificial Intelligence.

Key-words: Information Philosophy, Normative ethics, Artificial Intelligence, Adversarial Attacks, Metaethics.

INTRODUÇÃO

Ainda que os números exatos dos investimentos contra os ataques cibernéticos não sejam publicados, é possível perceber o quanto esse mercado tem crescido através dos

* Possui graduação em Filosofia pela Universidade do Vale do Rio dos Sinos (2021). Tem experiência na área de Filosofia, com ênfase em Filosofia, atuando principalmente nos seguintes temas: filosofia antiga, forma de vida, filosofia como forma de vida e estoicismo. Mestrando do PPG de Filosofia da UNISINOS.

Polymatheia
Revista de Filosofia

investimentos que as empresas desse ramo acumulam em suas capitalizações[1]. Com o advento das Inteligências Artificiais, novos problemas surgiram e o combate a um tipo específico, a saber, dos Ataques Adversários (*Adversarial Attacks*) aparenta ser uma das *últimas fronteiras* para uma confiança mais robusta na tomada de decisão por máquinas equipadas com Inteligências Artificias Generativas. Esses ataques serão mais especificados durante o artigo, mas basicamente é uma distorção de interpretação que as Inteligências Artificiais podem sofrer através de entrada de dados (*input*) corrompidos ou mesmo de informações mal interpretadas devido sua organização particular. A proposta desse estudo pode se vincular, então, somando a discussão um ponto de vista mais crítico e demonstrando através de estudos como pode haver uma distorção na realidade comunicada que compõe a Inteligência Artificial (IA). É importante que durante o desenvolvimento e a comunicação sobre como uma IA funciona haja clareza, demonstrando pontos forte e fraquezas, possibilitando melhorias reais no projeto (*design*).

O objetivo principal do artigo é possibilitar uma visão ampla das discussões em Ética da IA possibilitando, também, uma introdução de termos técnicos básicos provindos da Ciência da Computação. O aprofundamento da discussão posteriormente seria o fruto mais bem-vindo nesse contexto amplo. Objetivos secundários podem ser descolados do principal, tais como: (i) compreender a relação entre Ataques Adversários (*Adversarial Attacks*) numa perspectiva crítica da IA; (ii) investigar o panorama ético atualmente nas discussões e seus estilos principais de abordagem (*top-down* e *bottom-up*); e (iii) examinar a proposta de P. Railton no contexto Metaético e como isso acarreta numa concepção de implementação/aprendizagem de Ética para IAs. O método utilizado é a análise qualitativa dos dados levantados, assim como uma sumarização de termos técnicos que serão base para uma comparação hermenêutica dentro do contexto aproximativo das duas bases.

A segunda seção aborda o campo mais técnico, porém conduzindo agora para a especificidades da IA. O primeiro ponto, justamente, é o conceito e as variantes de IAs. Para essa construção foi realizado um levantamento histórico da evolução, no entanto, não foi focado tanto em reconstituir essa história, porque os próprios autores que formaram a

1 Os investimentos tem aumentado cada vez mais, conferir a matéria publicada sobre em: <https://www.bvp.com/atlas/cybersecurity-trends-in-2024>.

Polymatheia
Revista de Filosofia

base bibliográfica dessa parte do artigo já fizeram um trabalho mais completo sobre esse levantamento. Mais relevante é usar esses trabalhos para poder evoluir para a discussão filosófica. Posteriormente, a abordagem sobre algoritmos se mostrou imperativa, devido a função essencial que esses parâmetros reguladores das máquinas possuem, bem como a função de aprendizagem de máquina (*machine learning*) que é o cerne da evolução das IAs atualmente. O final dessa seção é dedicado a demonstrar o que são Ataques Adversários (AA) e como esse tipo de perigo para IAs se mostra como uma barreira que limita não apenas a própria Ciência da Computação, mas pode enfraquecer até mesmo a confiabilidade em sistemas de IAs autônomos.

Por fim, a última parte do artigo, antes das considerações finais, é dedicado para a abordagem do problema ético no contexto da IA, bem como sondar a função da Metaética na discussão sobre o tema. A primeira subseção é dedicada a abordar as duas variantes principais em disputa, a saber, o estilo *top-down*, no qual, grosso modo, princípios são elegidos para a aplicação da máquina antes dos problemas surgirem e o estilo *bottom-up*, no qual se parte da realidade da situação para que a máquina posteriormente ou imediatamente tome a decisão frente a dilemas morais, por exemplo. Dentro ainda dessa subseção o foco foi abordar a proposta do Equilíbrio Reflexivo (ER) que reflete uma postura mais *bottom-up*, assim sendo base para uma possível comparação com o método proposto por Railton posteriormente. Porém, o ER foi usado em sua nova roupagem dada por Savulescu e colaboradores do artigo, renomeado como CREP (*Collective Reflexive Equilibrium in Practice*) baseado no experimento público do *Moral Machine*, site desenvolvido pelo MIT (*Massachusetts Institute of Technology*) no qual colocam situações dilemáticas para quem está respondendo. Ou seja, a intenção dos autores é usar consensos morais coletados ao redor do mundo para basear uma programação que reflita as escolhas humanas de modo geral, mas que também leve em consideração dados científicos. O forte dessa abordagem seria o raciocínio por consistência, criando um sistema coerente e com possibilidade de revisão de padrões malsucedidos.

Nesse momento os textos relevantes como base foram os artigos de Amitai e Oren Etzioni, Denis Coitinho, Edmond Awad e parceiros, John Tasioulas, Savulescu e parceiros e Virginia Dignum. As duas últimas subseções foram dedicadas a sondar: (i) a proposta de

Polymatheia
Revista de Filosofia

realismo moral de Peter Railton; e (ii) a projeção dos conceitos de Railton para a Ética das IA's. Railton é identificado como um metaeticista no campo do realismo moral reducionista, ou seja, a objetividade das propriedades morais é redutível a propriedades naturais e fatos e valores são compatíveis dentro da sua proposta. Os artigos que baseiam essas afirmações são *Moral Realism* (1986) e *Ethical Learning, Artificial and Natural* (2020).

Através da pesquisa foi possível constatar que há uma influência sim das teorias metaéticas para a postulação e defesa dos filósofos sobre qual a melhor maneira de implementar uma Ética em IA. No entanto, as propostas ainda não levam em consideração algumas limitações das máquinas, o que pode gerar uma consequência não desejada no futuro, causada por Ataques Adversários, caso não sejam eliminados ou ao menos reduzidos significativamente.

INTELIGÊNCIA ARTIFICIAL, ALGORITMOS E ATAQUES ADVERSÁRIOS (*ADVERSARIAL ATTACKS*)

Seguindo a linha do artigo conexo com a ciência da computação e com a própria história dos conceitos computacionais o objetivo é cada vez mais propiciar um entendimento sobre o potencial e o limite da tecnologia.

A proposta desta parte do artigo é expor o que se entende por Inteligência Artificial (IA) e suas duas variantes, IA estreita (*narrow AI - ANI*) e IA geral (*general AI - AGI*), será dispensado da análise o que os especialistas[2] chamam de superinteligência artificial, já que supera ainda mais a perspectiva atual. Após, será exposto o significado de algoritmo e sua função geral para a ciência da computação. Por fim, o objetivo é colocar o problema dos Ataques Adversários (*adversarial attacks*) como limite para uma aplicação massiva das IA gerais.

INTELIGÊNCIAS ARTIFICIAIS

2 LEE, Kai-fu. AI superpowers: China, Silicon Valley, and the new world order. Boston: Houghton Mifflin Harcourt, 2018.

Polymatheia
Revista de Filosofia

É datado fortemente o verão de 1956 na *Dartmouth College* (Hanover, New Hampshire), no qual teóricos da computação se juntaram para debater sobre o tema e sobre a possibilidade dos avanços tecnológicos, como responsável sobre a nomenclatura *Artificial Intelligence*; é conhecido que em 1950, na revista *Mind*, Alan Turing já tinha publicado um artigo pautado sobre a possibilidade de uma máquina pensar. (BRINGSJORD; GOVINDARAJULU, 2018) Evidente que antes mesmo de Turing, outros pensadores como Ada Lovelace e até os gregos antigos pensaram em robôs, *golens*, autômatos artificiais, etc. (SILVEIRA, 2018, p. 23), cabe salientar que a proposta atual não é recuar a estes problemas que giram em torno da filosofia da mente e metafísica. Destaca-se que o problema atual é lidar com o desenvolvimento da Inteligência Artificial em consideração a sua aplicabilidade e sobre a necessidade de pautar o desenvolvimento sobre uma ética possível em IAs.

Para poder prosseguir com a exposição alguma definição do que é IA deve ser proposta, para tanto nesse artigo será utilizado o que Bringsjord e Govindarajulu (2018) afirmam em seu artigo na SEP a respeito da definição de IA proposta por Russel e Norvig[3] (1995, 2002, 2009):

> Todas essas respostas pressupõem que a IA deve ser definida em termos de seus objetivos: uma definição candidata, portanto, tem a forma "IA é o campo que visa construir..." Todas as respostas se enquadram em um quarteto de tipos dispostos em duas dimensões. Uma dimensão é se o objetivo é igualar o desempenho humano ou, em vez disso, a racionalidade ideal. A outra dimensão é se o objetivo é criar sistemas que raciocinam/pensam ou, em vez disso, sistemas que agem[4] (BRINGSJORD; GOVINDARAJULU, 2018).

O que justifica essa escolha e não de outra é porque ela representa muito bem as relações que os humanos estabelecem entre comportamento humano e razões e de que maneira isto impacta nas expectativas que se criam em torno das ações que as máquinas implementarão a partir de um modelo de engenharia. Ou seja, poderemos esquivar da definição pétrea de IA, que exigiria uma longa discussão, e ainda reflete bem os

3 Disponível em: https://aima.cs.berkeley.edu/

4 "These answers all assume that AI should be defined in terms of its goals: a candidate definition thus has the form "AI is the field that aims at building…" The answers all fall under a quartet of types placed along two dimensions. One dimension is whether the goal is to match human performance, or, instead, ideal rationality. The other dimension is whether the goal is to build systems that reason/think, or rather systems that act." (Tradução nossa)

Polymatheia
Revista de Filosofia

posicionamentos que eticistas defendem que a IA ou deve ser guiada por princípios gerais (*top-down*) ou deve ser guiada por uma compreensão mais próxima a humana (*bottom-up*).

Dentro das perspectivas mais comentadas sobre Inteligências Artificiais se destacam duas formas, de um lado tem-se a IA estreita (*narrow AI*) e, por outro lado, a IA geral (*general AI*). A primeira forma (*weak AI* ou IA fraca em português) tem uma concentração em tarefas específicas e é amplamente aplicada nos dias atuais (2023). Por exemplo, assistentes pessoais, canais de *streamings*, *chatbots* e sua fraqueza advém dessa capacidade limitada por um objetivo específico, já a IA geral (ou general AI) provem de uma intenção de réplica da inteligência humana e ainda não existe tal projeto consolidado, é uma projeção do futuro da IA, Tasioulas descreve da seguinte forma essas IAs:

> A IA geral refere-se a máquinas inteligentes capazes de replicar uma ampla gama de capacidades intelectuais humanas e até mesmo de superá-las. Essas formas de IA, embora conhecidas de personagens de ficção científica, como o C3PO de Guerra nas Estrelas, estão, na melhor das hipóteses, em um futuro muito remoto. Se houve algum progresso significativo na IA nos últimos anos, ele ocorreu na IA estreita. Trata-se de máquinas que reproduzem ou superam as capacidades humanas em relação a uma gama limitada de tarefas, como, por exemplo, dirigir um carro, fazer diagnósticos médicos ou traduzir idiomas[5] (TASIOULAS, 2019, p. 63).

ALGORITMO

Após a breve descrição de IA, um fator importante de destacar o fator algorítmico. Isso porque é pelo algoritmo que a IA define seu raio de ação, de uma maneira mais simples, segundo Angius et al. (2021), o algoritmo é um conjunto de regras que guia na execução de uma tarefa e sobre a origem do termo, os autores ainda afirmam que: "A palavra "algoritmo" tem origem no nome do matemático persa do século IX, *Abū Jaʿfar Muḥammad ibn Mūsā al-Khwārizmī*, que forneceu regras para operações

5 "General AI refers to intelligent machines that are able to replicate a broad range of human intellectual capacities, and even to surpass them. These forms of AI, al- though familiar from science fiction characters such as Star Wars' C3PO, lie at best in the very remote future. To the extent that there has been any significant progress in AI in recent years, it has occurred in narrow AI. These are machines that replicate, or exceed, human capabilities with respect to a limited range of tasks, e.g. car-driving, medical diagnosis or language translation." (Tradução nossa)

Polymatheia
Revista de Filosofia

aritméticas usando algarismos arábicos."[6]

Um dos fatores que se faz útil a definição desse termo da ciência da computação é seu escopo necessariamente finito. Dito de outro modo, não seria viável a criação de um algoritmo, ou seja, de uma regra ou instrução, que se alargasse em infinitas possibilidades. Essas afirmações são retiradas de dois extratos importantes sobre algoritmos que Angius et. al (2021) escreve na SEP, a (1) se refere a abordagem clássica:

> Como em Kleene (1967), a finitude afeta tanto o número de instruções quanto o número de etapas computacionais implementadas. Como na determinação de Markov, o princípio de definição de Knuth exige que cada etapa computacional sucessiva seja especificada sem ambiguidade; e a (2) é quando se refere à abordagem informal: Algoritmos são imperativamente dados, pois comandam transições de estado para realizar operações especificadas. Por fim, os algoritmos operam para atingir determinados objetivos sob algumas disposições, ou pré-condições, geralmente bem especificadas.[7]

Popularizou-se a explicação de que algoritmos capturam o gosto das pessoas em rede sociais, por exemplo, o que acontece é uma concatenação de algoritmos para responder mais eficientemente a um propósito[8].

Um dos fatores que tem mudado a amplitude dos algoritmos é o aprendizado de máquina (*machine learning*). Essa mudança se dá, principalmente, por ter refinado a forma com que os algoritmos extraem informações sem ser necessário um modelo matemático prévio. (FONTANA, 2020, p. 3) Existem três formas de compreender a atuação desses algoritmos de aprendizagem automática, a primeira se refere aos algoritmos supervisionados:

> Os algoritmos de aprendizagem supervisionada relacionam uma saída com uma entrada com base em dados rotulados. Neste caso, o usuário alimenta ao algoritmo pares de entradas e saídas conhecidos, normalmente na forma de

6 "The word "algorithm" originates from the name of the ninth-century Persian mathematician Abū Jaʿfar Muḥammad ibn Mūsā al-Khwārizmī, who provided rules for arithmetic operations using Arabic numerals." (Tradução nossa)

7 "As in Kleene (1967), finiteness affects both the number of instructions and the number of implemented computational steps. As in Markov's determinacy, Knuth's definiteness principle requires that each successive computational step be unambiguously specified."; e a (2) é quando se refere a abordagem informal: "Algorithms are imperatively given, as they command state transitions to carry out specified operations. Finally, algorithms operate to achieve certain purposes under some usually well-specified provisions, or preconditions" (Tradução nossa)

8 Disponível em: https://canaltech.com.br/inteligencia-artificial/o-que-e-um-algoritmo-226839/.

Polymatheia
Revista de Filosofia

> vetores. Para cada saída é atribuído um rótulo, que pode ser um valor numérico ou uma classe. O algoritmo determina uma forma de prever qual o rótulo de saída com base em uma entrada informada (FONTANA, 2020, p. 3).

Nesse tipo de algoritmo supervisionado, existem dois tipos, o primeiro é o que se chama algoritmos de classificação. Um primeiro passo capta as características (atributos) dos dados. Após, o algoritmo utiliza o auxílio de informações (etapa de aprendizagem) que caracterizam alguns dados para criar um padrão e a partir de então estabelecer um modelo (aprendido) a seguir. O que caracteriza essa tipagem é a limitação dos valores de saída a partir da própria base de dados. Outro algoritmo aplicado nesse tipo (supervisionado) é o algoritmo de regressão, que funciona basicamente como o de classificação, porém age de maneira a prever um número ou dado futuro a partir das características já estabelecidas, ou seja, se existe um banco de dados o próximo a ser acrescentado pode ser rotulado a partir da base existente, basicamente podendo ter qualquer valor de saída[9].

O segundo grande grupo de algoritmos de aprendizagem automática é denominado algoritmo não-supervisionado, como não se tem os "rótulos" que tinham o anterior, apenas montantes de dados agrupados por similaridade ou simplicidade: "Com base em um número grande de dados, o algoritmo busca padrões e similaridades entre os dados, permitindo identificar grupos de itens similares ou similaridade de itens novos com grupos já definidos." (FONTANA, 2020, p. 4) Nesse grupo há dois tipos de algoritmos, os de transformação e os de agrupamento. Nos algoritmos de transformação, os dados existentes são convertidos para facilitar a compreensão humana da máquina ou para tornar mais eficiente algum processo interno. Já os algoritmos de agrupamento (*clustering*): "particionam os dados em grupos com características similares com base em critérios pré-estabelecidos, permitindo encontrar padrões entre os dados fornecidos." (FONTANA, 2020, p. 4)

O último grupo de algoritmos a ser destacado é o de aprendizagem por reforço. Fundamentalmente esse tipo de algoritmo é recompensado por sua eficiência, gerando assim uma busca por maximizar essa recompensa. Esse modo de aprendizagem é mais simples, mas não menos complexa, pois necessita uma interação continua entre agente,

9 Disponível em: https://blog.geekhunter.com.br/aprendizado-de-maquina-e-seus-algoritmos/.

Polymatheia
Revista de Filosofia

ambiente e ação. (FONTANA, 2020, p. 4)

Um ponto que vale destacar é o procedimento de *Deep Learning* (DP) ou aprendizado profundo. É um estágio avançado do processo de captura de padrões. A premissa permanece a dos algoritmos de aprendizagem, mas com uma arquitetura que pretende simular uma rede neural (*Neural Network*): "Por esse motivo, as redes treinadas por métodos de aprendizagem profunda são frequentemente chamadas de redes neurais, embora a semelhança com as células e estruturas neurais reais seja superficial."[10] (RUSSEL; NORVIG, 2019, p. 750) O que é denominada por rede neural (NN) é um modelo de camadas que visa criar padrões mais complexos, essa complexidade, segundo Russel e Norvig (2019, p.750) é resultado do alongamento dos caminhos (*path*), com constantes associações e interações entre os grupos de informações, para se chegar a um dado de saída mais refinado.

Durante a pesquisa pareceu muito importante ressaltar o papel do algoritmo para as Inteligências Artificiais, pois é imprescindível entender o funcionamento de escolha dos padrões ou regras que tal dispositivo vai responder para, então, poder tecer alguma crítica ou advertência de forma mais adequada.

ATAQUES ADVERSÁRIOS (*ADVERSARIAL ATTACKS*)

Sob certo ponto de vista a empreitada até aqui foi para o leitor compreender melhor o que são os ataques adversos. Entender melhor a configuração da máquina e seus defeitos, para posteriormente refletir sobre uma possibilidade ética para os sistemas computacionais de IA, esse é um dos objetivos do artigo.

Os Adversarial Attacks[11] (ataques adversários ou antagônicos[12]) é o nome dado a uma forma de *input* (entrada) contaminado com informações (*pixels* por exemplo) que interferem na compreensão de Inteligências Artificiais. Essa interferência causa uma

10 "For this reason, the networks trained by deep learning methods are often called neural networks, even though the resemblance to real neural cells and structures is superficial." (Tradução nossa)

11 Em alguns textos também é referenciado o termo *adversarial examples* (RUSSEL; NORVIG, 2019; ATHALIE et al., 2023).

12 *Antagónicos* é como está na tradução para o português de Portugal no documento oficial de regulamentação da União Europeia.

Polymatheia
Revista de Filosofia

distorção, como no exemplo retirado do AIMA:

> Por exemplo, pode ser possível alterar apenas alguns pixels em uma imagem de um cachorro e fazer com que a rede classifique o cachorro como um avestruz ou um ônibus escolar, mesmo que a imagem alterada ainda se pareça exatamente com um cachorro[13] (RUSSEL; NORVIG, 2019, p. 769 e 770).

Mais precisamente, os ataques adversos (AA) ocorrem no nível de *deep learning*[14] da IA atingindo a base da rede neural (NN). O que resulta é caracterizar o método de rede neural e aprendizagem profunda como instável (*unstable*):

> No entanto, agora há evidências empíricas esmagadoras de que as técnicas atuais de DL geralmente levam a métodos instáveis, um fenômeno que parece universal e presente em todos os aplicativos listados acima (13-21) e na maioria das novas tecnologias de Inteligência Artificial (IA) (COLBROOK; ANTUN; HANSEN, 2023, p. 1)[15]

Um dos dilemas que foi encontrado em redes neurais (NN) computacionais, é o que os pesquisadores denominam por compensação (*trade-off*), foi notado que para estabilizar o funcionamento de um programa, você deve sacrificar uma parte de sua precisão, e vice versa. (COLBROOK; ANTUN; HANSEN, 2022, p. 1) Alguns dos exemplos que Wang et al. (2023) dão em seu artigo sobre exemplos notáveis de ataques adversários, um exemplo é datado de 2019, no qual um sistema de crédito não identificou corretamente um cartão de crédito falso, após a alteração de alguns dados na transação, o texto alerta ainda para o risco dos usuários e das instituições caso haja recorrência de tais atos. (WANG et al., 2023, p. 4) Mas esse é apenas um exemplo que se encontra na área da economia, existe outros exemplos, mas a partir de agora dois serão abordados em relação ao carro autônomo. Wang et al. (2023, p. 4) comentam sobre dois casos, um em 2020 e outro 2021, no qual veículos não tripulados pararam, equivocadamente, em placas que

13 For example, it may be possible to alter just a few pixels in an image of a dog and cause the network to classify the dog as an ostrich or a school bus — even though the altered image still looks exactly like a dog. (Tradução nossa)

14 Aqui já faz parte de uma IA geral (general AI). Não parece ser necessário atribuir mais um nível de complexidade para explicar o que seria *black box* (caixa preta) e a *white box* (caixa branca) que definem níveis de explicabilidade do algoritmo, dado que os AA atingem ambos,

15 "However, there is now overwhelming empirical evidence that current DL techniques typically lead to unstable methods, a phenomenon that seems universal and present in all of the applications listed above (13–21) and in most of the new artificial intelligence (AI) technologies." (Tradução nossa)

Polymatheia
Revista de Filosofia

possuíam pequenos adesivos alterando sua imagem. O outro exemplo é um no qual os pesquisadores provaram que pode haver engano (*mistake*) na interpretação de imagens em carros autônomos, podendo provocar acidentes graves: "Os pesquisadores mostraram que uma pequena perturbação adicionada a uma placa de trânsito poderia fazer com que o carro autônomo identificasse erroneamente a placa, o que poderia levar a erros perigosos na estrada."[16] (WANG et al., 2023, p. 4) Foi demonstrado, também, a fragilidade em reconhecimento de imagens em um caso em que se inseriu alterações de imagem para testar a acurácia de uma IA, que falhou ao identificar uma tartaruga com um rifle em cerca de 96% dos casos, estes testes foram executados cerca de 1000 vezes. (ATHALIE et al., 2018)

Existe um dado impressionante sobre os AA, no qual esse tipo de ataque foi simulado em rede *wi-fi* com sinais de -18db e a precisão dos ataques passou de 90%. (WANG et al., 2023, p. 5) Diversas táticas estão sendo testadas para tentar entender e combater os AA, algumas tentativas se centram em softwares que contribuem para estabilizar um sistema de NN (e. g., *FIRENET*, *AUTOMAP* no artigo de COLBROOK et al., 2023, p. 6); outras buscam aprender com os adversários (*adversarial learning*) (WANG et al., 2023, p. 20); alguns programas se focam em monitorar (*monitoring*) os AA no sistema (WANG et al., 2023, p. 20) e outros métodos que buscam compreender e mitigar a margem de erro das IAs, mas nenhum ainda atingiu eficácia total.

A visão geral sobre o tema dos AA é que é um problema sério e que precisa ser sumarizada nas discussões sobre ética da IA. É preciso assegurar certa compreensão da limitação dessas tecnologias, tanto para estabelecer um método de configuração ética da máquina quanto para não projetar na máquina tenha uma responsabilidade maior do que sua capacidade real de distinção da realidade. A seção seguinte traçará um apanhado ético e metaético no contexto da IA.

ÉTICA DAS INTELIGÊNCIAS ARTIFICIAIS E A PROPOSTA DE PETER RAILTON PARA O APRENDIZADO ÉTICO DE SISTEMAS DE

16 "The researchers showed that a small perturbation added to a traffic sign could cause the self-driving car to misidentify the sign, potentially leading to dangerous mistakes on the road." (Tradução nossa)

Polymatheia
Revista de Filosofia

INTELIGÊNCIAS ARTIFICIAIS

A preocupação humana sobre qual decisão uma máquina tomará em uma situação crítica envolvendo a vida e a morte de pessoas e demais animais, por exemplo, aflora as discussões na área da ética. Não é difícil de desprender perguntas que questionam como uma máquina irá "aprender" o conteúdo ético ou, sendo menos pretencioso, como ela irá aplicar na realidade algo que se assemelhe a uma ação ética e que corresponda adequadamente a expectativas específicas dos humanos sobre as máquinas. É relevante para esse estudo pensar a ética como um conjunto de saberes que atravessam a realidade de decisões reais (primeira ordem) somado a reflexões sobre o que orientam essas tomadas de decisões (segunda ordem), assim Alexander Miller abre seu livro afirmando que:

> Primeiro, há questões de primeira ordem sobre qual parte no debate, se houver, está certa e por quê. Depois, há questões de segunda ordem sobre o que as partes no debate estão fazendo quando se envolvem nele. Grosso modo, as questões de primeira ordem são da alçada da ética normativa, e as questões de segunda ordem são da alçada da metaética[17] (MILLER, 2003, p. 1).

Enquanto as questões normativas buscam a resposta de como se deve agir, as perguntas que a metaética busca compreender é como chegamos ao conhecimento moral (epistemologia), ou como se justificam formas de conhecimento moral, ou se existem propriedades morais (ontologia), ou como os discursos morais sustentam fatos morais ou não (semântica) entre outros questionamentos que residem nessa profundidade reflexiva.

A intenção de diferenciar esses dois campos de ação é importante por dois fatores dentro dessa pesquisa: (i) investigar até que ponto uma IA seria capaz de assimilar princípios para tomada de decisões; e (ii) entender como a orientação metaética, usando exemplo do realismo moral de Peter Railton, contribui para a formação de sua teoria ética para a IA. Será dedicado essa seção a abordar propostas mais próximas do estilo *bottom-up* de raciocínio moral, pois parece o sistema mais promissor e ao mesmo tempo que traz mais complexidade para refletir sobre como devemos lidar com a perspectiva de máquinas

17 "First, there are first order questions about which party in the debate, if any, is right, and why. Then, there are second order questions about what the parties in the debate are doing when they engage in it. Roughly, the first order questions are the province of normative ethics, and the second order questions are the province of metaethics." (Tradução nossa)

Polymatheia
Revista de Filosofia

tomarem decisões impactantes na vida das pessoas.

ÉTICA DA INTELIGÊNCIA ARTIFICIAL

Um dos problemas éticos que mais chamam a atenção talvez seja o de carros autônomos. Uma das reflexões possíveis faz referência aos *Trolley problems*, amplamente discutidos na Filosofia desde o século XX. (THOMSON, 1985) A grande diferença é que se elevou um nível dentro da reflexão, pois agora se trata realmente de decidir qual seria a escolha mais aceitável dentro de um dilema ético. Enquanto nos problemas originais as situações hipotéticas definiam mais a maneira que se argumentava e de que maneira algumas teorias demonstravam fragilidades, agora se trata de lidar com teorias que de fato impactarão no mundo.

A discussão se faz presente desde o momento que não se tem um consenso e não se encontra forma de provar qual teoria ética está correta em absoluto. Por isso a qualidade da abordagem e eficiência de cada teoria em relação aos problemas que uma IA enfrentará no dia-a-dia será o fator que determinará qual sistema ético deve ser implantado. (DIGNUM, 2019, p. 71).

Dentro dos debates que ocorrem em torno do tema é possível destacar duas com grandes destaques[18]. A primeira é da ordem principialista (*top-down*), no qual regras ou comandos pré-estabelecidos determinam como que o robô executará uma tarefa, isso torna a ação da IA altamente previsível. A segunda abordagem seria inversa da anterior (*bottom-up*), ou seja, partir de dados coletados da situação e permitir que a IA escolha a opção mais apropriada para aquele momento. (TASIOULAS, 2019, p. 64)

Os debates ainda se alongam, se em cada uma dessas duas categorias postularem visões diferentes sobre como chegar em suas próprias conclusões. O exemplo é nítido na maneira de cima para baixo (*top-down*) de encarar o problema ético das IA. Se uma teoria utilitarista fosse aplicada, a melhor situação de escolha de uma IA seria aquela em que se maximizasse o bem-estar. O kantismo também é uma maneira *top-down* de compreender o

[18] Dignum (2019, p. 72) aborda outra vertente que seria o hibridismo (*Hybrid approachs*), que seria uma mescla das teorias *top-down* e *bottom-up*. Essa pesquisa se concentrará nas duas destacadas (*top-down* e *bottom-up*) por haver referências mais específica sobre elas.

Polymatheia
Revista de Filosofia

problema, pois elevaria os princípios como imperativos categóricos como resposta moral para todas as situações que podem ocorrer. E, por fim, qualquer variante de consequencialismo seria outro exemplo de teoria ética que busca imprimir nas decisões particulares previsões de resultados melhores esperados. (ETZIONI; ETZIONI, 2017, p. 406)

Sobre a abordagem de cima para baixo (*bottom-up*) partiria da observação das situações reais, seguido pelo processamento da máquina, para então tomar uma decisão. Um bom exemplo na perspectiva *bottom-up* é do equilíbrio reflexivo (ER). O procedimento do Equilíbrio Reflexivo original provém daquilo que Ralws propôs em sua Teoria da justiça, no qual sugere que ao invés de buscar uma verdade última em preceitos éticos se busque uma coerência de razões que leve em conta crenças morais, princípios éticos e fatos científicos para tomada de decisão. Nesse sistema coerentista de raciocínio moral a consistência ocupa o valor que antes a verdade obtinha sob fatos morais. (COITINHO, 2023, p. 59 e 60) Um ponto forte dessa abordagem ética é que uma gama de fatores contribui para formar uma decisão refletida, pois uma das suas características é a revisionabilidade. (COITINHO, 2023, p. 62)

A versão do Equilíbrio Reflexivo que é defendida e que se relaciona com a ética da IA pode ser descrito pela proposta de Savulesco, Gyngell e Kahane (2021) contida no artigo intitulado *Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies*. Os autores denominam como *CREP* (*Collective Reflexive Equilibrium in Practice*) que é um Equilíbrio Reflexivo aplicável na prática a partir da coletividade. A intenção foi usar os dados coletados pela *Moral Machine*[19], desenvolvida pelo *MIT*[20], no qual uma série de escolhas e opções eram disponibilizados como escolha de resposta, esse experimento alcançou cerca de 40 milhões de pessoas ao redor do mundo. Com essa consulta pública a intenção foi coletar dados que correspondem a ações que humanos tomariam em situações dilemáticas, houve uma filtragem dos dados, assim:

> Uma vez que tenhamos um conjunto de "dados robustos" de atitudes públicas que tenham sido "lavados" dessa forma preliminar para minimizar o viés e

19 Sobre os detalhes do experimento, recomendamos o artigo *The moral machine experiment* de Awad et al., publicado na Nature, em novembro de 2018, vol. 563.
20 Disponível em: https://www.moralmachine.net/.

Polymatheia
Revista de Filosofia

> aumentar a confiabilidade, e que reflitam, na medida do possível, as verdadeiras preferências, a segunda etapa do CREP é procurar a coerência entre essas intuições e os princípios morais[21] (SAVULESCU; GYNGELL; KAHANE, 2021, p. 8).

A abordagem de equilibrar intuições, razões e princípios é uma qualidade do procedimento. Uma série de críticas foram postuladas ao longo do tempo sobre o procedimento do ER, que apenas citaremos duas, tais como a desconfiança na credibilidade inicial das crenças e da possibilidade de a coerência de um sistema de crenças refletir apenas um subjetivismo (ou ainda o perigo do relativismo). (COITINHO, 2023, p. 63 e 64) Esse exemplo do Equilíbrio Reflexivo é uma das formas mais proveitosas de demonstrar as discussões no campo das escolhas que faremos na hora de pensar sobre o projeto (*design*) que se desenvolverá para IAs. Agora um desafio para o ER, CREP e ERP[22] é como justificar as tomadas de decisões caso a máquina continue sofrendo interferências dos Ataques Adversários, gerando a interpretação errônea de imagens simples (que seria identificado por qualquer humano típico facilmente).

Dando prosseguimento, um fator que se precisa destacar é refletir sobre o que possibilitou o estilo *bottom-up* em máquinas foi, justamente, a *machine learning*. A capacidade exponencial de cálculo de possibilidades e de "aprendizagem", ou ainda, capacidade de assimilação de padrões complexos, possibilitando projetar se um sistema de IA não seria capaz de ponderar durante o trajeto a sua escolha em uma situação dilemática.

CONTEXTO METAÉTICO DO REALISMO MORAL REDUCIONISTA DE PETER RAILTON

Um exemplo de estilo de baixo para cima (*bottom-up*) seria a proposta defendida P. Railton. Os argumentos de Railton se encontram no artigo *Ethical learning, Natural and*

21 "Once we have a set of 'robust data' of public attitudes that have been 'laundered' in this preliminary way to minimize bias and increase reliability, and to reflect as far as possible true preferences, the second step of CREP is to look for coherence between these intuitions and moral principles." (Tradução nossa)

22 Para Coitinho (2023, p. 70) o Equilíbrio Refletivo Prudente (ERP) solucionaria alguns problemas relacionados ao ER, que "De forma geral, o ERP contará com a expertise do agente com sabedoria prática para bem deliberar, isto é, para deliberar adequadamente sobre os meios necessários para alcançar um fim bom, o que pode ser visto como chegar a crenças razoáveis."

Polymatheia
Revista de Filosofia

Artificial, publicado dentro do livro *Ethics and Artificial Intelligence*[23] organizado por S. Matthew Liao, no ano de 2020. Outro artigo importante de Railton que será utilizado é o *Moral Realism*, de 1986, a referência, no entanto, é do compilado de artigos que o autor publicou como livro em 2003[24].

Peter Railton[25] é um reconhecido eticista que trabalha com questões de metaética, principalmente se associando a vertente do realismo moral. Há uma gama de opções dentro dessa compreensão metaética, sendo a postura de Railton uma forma de reducionismo naturalista: "[...] que as propriedades morais são objetivas, embora relacionais; que as propriedades morais se sobrepõem às propriedades naturais e podem ser reduzidas a elas; [...]"[26] (RAILTON, 2003, p. 5) Um fator importante desse reducionismo é perceber a convergência entre fatos e valores, isso porque se houvesse uma distinção abrupta entre eles causaria um problema para uma proposta fiável em ética, enfraquecendo o argumento da objetividade das propriedades morais e sobre a maneira de assimilar o conteúdo moral, ainda que o autor admita que o problema o intrigue. (RAILTON, 2003, p. 9)

Outro ponto que é preciso destacar é a relação de independência das propriedades morais e o *feedback* entre o ambiente e a formação moral da pessoa, fator que é importante para refletir posteriormente sobre a IA e o aprendizado moral. A primeira instância que sustenta o realismo dos valores morais é da independência das propriedades morais, que é destacada por Railton (2003, p. 9) como: "(i) independência: ele existe e tem certas características determinadas, independentemente do fato de pensarmos que ele existe ou tem essas características, independentemente, inclusive, de termos boas razões para pensar

23 "This volume emerged in part from a conference in October 2016 on the ethics of artificial intelligence at New York University hosted by the NYU Center for Mind, Brain and Consciousness and the NYU Center for Bioethics." (LIAO, 2020, X)

24 RAILTON, Peter. *Facts, values, and norms*: Essays toward a morality of consequence. Cambridge University Press, 2003.

25 Professor da Universidade de Michigan, departamento de Filosofia: "Professor Railton's main research has been in ethics and the philosophy of science, focusing especially on questions about the nature of objectivity, value, norms, and explanation." Disponível em: https://lsa.umich.edu/philosophy/people/faculty/prailton.html

26 "[...] that moral properties are objective, though relational; that moral properties supervene upon natural properties, and may be reducible to them; [...]" (Tradução nossa)

Polymatheia
Revista de Filosofia

assim."[27] Ainda que exista independência, propriedades morais não são algo *sui generis*, algo que precisasse uma explicação ou modelo além mundo. Os interesses objetivos *brotam* (*supervenience*) das relações naturais e sociais humanas e o valor moral se materializa pela prática humana. (RAILTON, 2003, p. 16) O outro ponto de apoio é o *feedback*, que é caracterizado como a capacidade do humano de modelar seus padrões morais a partir da interação com o mundo. (RAILTON, 2003, p. 10) Para o filósofo norte-americano esses dois fatores contribuiriam para elevar a objetividade das propriedades morais e superar a visão mais relativista e subjetivista da moralidade.

Para sustentar a possibilidade de evolução moral, Railton, recorre à racionalidade individual, que vai selecionando, muito por tentativa e erro, mas que também envolve certa reflexividade ou explicação criteriosa (*criterial explanation*), assim:

> Os padrões de crenças e comportamentos que não exibem muita racionalidade instrumental tendem a ser, até certo ponto, autodestrutivos, um incentivo para mudá-los, enquanto os padrões que exibem maior racionalidade instrumental tendem a ser, até certo ponto, recompensadores, um incentivo para continuá-los[28] (RAILTON, 2003, p. 19).

Essa maneira de interpretar a realidade moral não deixa de comportar o aprendizado, sendo conectado diretamente à chave desejos/interesses que regem o comportamento humano, consciente ou inconscientemente. (RAILTON, 2003, p. 20)

APRENDIZADO MORAL ARTIFICIAL E NATURAL EM PETER RAILTON

Partindo agora para a aproximação do realismo moral reducionista de Railton e a IA. Logo de início no artigo de 2020, Railton afirma que:

> Em uma primeira aproximação, podemos caracterizar a sensibilidade a preocupações éticas como uma capacidade robusta e confiável de detectar e

27 "(i) independence: it exists and has certain determinate features independent of whether we think it exists or has those features, independent, even, of whether we have good reason to think this." (Tradução nossa)

28 "Patterns of beliefs and behaviors that do not exhibit much instrumental rationality will tend to be to some degree self-defeating, an incentive to change them, whereas patterns that exhibit greater instrumental rationality will tend to be to some degree rewarding, an incentive to continue them." (Tradução nossa)

Polymatheia
Revista de Filosofia

> responder adequadamente a características eticamente relevantes de situações, ações, agentes e resultados[29] (RAILTON, 2020, p. 45).

Essa caracterização atravessaria a agência humana e se enquadraria também para sistemas de IA gerais, já que a evolução de sistemas artificiais não supervisionados poderiam pesar razões e responder apropriadamente em relação a situações, ações, agentes e resultados. Esse salto entre o momento que uma IA é apenas instrumentalizada na sociedade e depois é vista como uma peça social de grande valor é nítido quando o norte-americano afirma: "Precisaremos encontrar maneiras de coordenar, cooperar, colaborar e competir de forma pacífica e produtiva com sistemas artificiais, vistos como partes independentes cujo comportamento não podemos simplesmente ditar."[30] (RAILTON, 2020, p. 46)

Um dos problemas que é detectado dentro da perspectiva relevante para esse artigo é a maneira com que Railton (2020, p. 47) associa a aprendizagem de máquina (*machine learning*) com a aprendizagem humana. Para o norte-americano se deveria olhar para a maneira que o humano aprende e assim simular em IAs. Nesse ponto é importante notar o arcabouço metaético destacado anteriormente, isso porque novamente Railton reduzirá a capacidade as propriedades morais a uma maneira de captar do ambiente natural e social humano características eticamente relevantes. Nesse sentido, seria mais um ponto que reforça o estilo *bottom-up* para máquinas, que coletariam das "experiências" do mundo para otimizar sua própria tomada de decisão, inclusive chega a afirmar uma situação de aprendizado envolvendo carros autônomos:

> Imagine a diferença em aprender a dirigir com segurança em uma gama aberta de situações se uma comunidade de carros autônomos compartilhar dados individuais de direção, em vez de cada um usar os dados que adquire para obter

29 "To a first approximation, we can characterize sensitivity to ethical concerns as a robust, reliable capacity to detect and respond appropriately to ethically relevant features of situations, actions, agents, and outcomes."(Tradução nossa)

30 "We will need to find ways to coordinate, cooperate, collaborate, and compete peacefully and productively with artificial systems, seen as independent parties whose behavior we cannot simply dictate."(Tradução nossa)

Polymatheia
Revista de Filosofia

> quaisquer vantagens estratégicas que puder sobre os outros[31] (RAILTON, 2020, p. 49).

A ideia de Railton (2020, p. 50) é que a maneira como a criança aprende é um bom exemplo para aprimorar o aprendizado de máquina. A argumentação de Railton paira sobre a perspectiva de aprendizagem ética em crianças aconteceria de forma espontânea, sem instruções diretas, por exemplo, uma mãe não precisa explicar teoricamente para o filho porque é errado ser cruel, simplesmente fica evidente que é errado ser cruel com pessoas e demais animais. Railton ainda complementa, sendo coerente com o fator já mencionado sobre o *feedback* e o aprendizado moral:

> E as formas probabilísticas de aprendizado transformam essa experiência aparentemente "passiva" em algo mais do que uma "mera associação". Em vez disso, é uma forma de experimentação ativa, com a formação contínua de expectativas com base nas associações observadas e no feedback contínuo das discrepâncias entre essas expectativas e os resultados reais.[32] (RAILTON, 2020, p. 51).

Por mais que a análise que Railton faz do aprendizado moral em crianças, muito bem elaborado, não faz muito sentido dentro do contexto de IA's. Porque se fosse, deveríamos nos questionar o que é infância e se uma máquina nasce ou possui infância ou inocência. Pouco provável que sejam questões relevantes para este artigo, por isso essa instância apresentada por Railton fica como uma menção de como o arcabouço metaético de um filósofo pode contribuir para a aplicação de raciocínio moral em máquinas inteligentes. Isso porque ele reforça a origem das propriedades morais na interação humana (mundo social) e a importância do *feedback* para o aprimoramento de tais aptidões morais.

A ideia de investigar como o humano delibera moralmente e como o raciocínio humano funciona, recusando o *principialismo* ou uma forma alinhamento (*align*) com padrões principialistas (RAILTON, 2020, p. 64), aparenta trazer uma estratégia mais

31 "Imagine the difference in learning to drive safely in an open-ended array of situations if a community of self-driving cars shares individual driving data rather than each using the data it acquires to gain whatever strategic advantages it can over the others." (Tradução nossa)

32 "And probabilistic forms of learning turn this seemingly "passive" experience into more than "mere association." Instead it is a form of active experimentation, with the continuous formation of expectations on the basis of observed associations and continuous feedback from discrepancies between such expectations and actual outcomes." (Tradução nossa)

Polymatheia
Revista de Filosofia

eficaz, mas ainda um passo além do que realmente é possível.

Na resenha que Dora Kaufman (2021, p. 160) realiza do livro de S. Mattew Liao, é categórica ao afirmar sobre o escrito de Railton e dos outros quatro artigos iniciais do compilado: "Os cinco ensaios atribuem um agenciamento inexistente nos sistemas atuais de IA, que são "meros" modelos estatísticos de probabilidades aos quais não pode ser atribuída a condição de agente moral, pressuposto corroborado por vários filósofos." Nesse contexto, o próprio Railton (2020, p. 65) admite não possuir *expertise* sobre aprendizados de máquina, por isso que esse artigo parece complementar à discussão. Porque, de um lado, mostra como Railton está baseando sua proposta de Ética da IA dentro de um contexto de fundamentação metaética, um reflexo de seu arcabouço filosófico e de seus estudos anteriores e, por outro lado, a crítica resumida a *inexistência* de agenciamento moral em IAs parece não ser suficiente para colaborar com uma discussão frutífera sobre o tema. Dado as perspectivas e os temas abordados é possível partir para a reflexão crítica da pesquisa.

CONSIDERAÇÕES FINAIS

O campo da tecnologia digital ainda é muito recente. Muito será desenvolvido e poderá ser polido os *serrilhados* dos conceitos. Isso é perceptível na pesquisa sobre os conceitos básicos apresentado na primeira seção do artigo. A apropriação por parte da Filosofia dos conceitos computacionais ainda está em fase inicial. Isso pode ser exemplificado com o conceito de *Big data* em computação, até o surgimento do escândalo da *Cambridge Analitics*[33], pouco se acreditava que existia um tipo de poder tão grande associado a esse conjunto de dados. Os algoritmos agora fazem parte da sabatina de estudos de qualquer pensador que queira criticar ou mostrar os benefícios de tal mecanismo. Agora, veja o exemplo dos Ataques Adversários (*Adversarial Atacks*), na própria regulamentação europeia, o qual se espera uma amplitude e uma segurança para

33 "A empresa teria tido acesso ao volume de dados ao lançar um aplicativo de teste psicológico na rede social. Aqueles usuários do Facebook que participaram do teste acabaram por entregar à Cambridge Analytica não apenas suas informações, mas os dados referentes a todos os amigos do perfil." Matéria completa disponível em: https://g1.globo.com/economia/tecnologia/noticia/entenda-o-escandalo-de-uso-politico-de-dados-que-derrubou-valor-do-facebook-e-o-colocou-na-mira-de-autoridades.ghtml

Polymatheia
Revista de Filosofia

colocar em funcionamento sistemas inteligentes artificiais, cita apenas uma vez (p. 20) um problema que pode causar uma catástrofe. Um exemplo que pode ser usado é, se na implementação dos veículos não-tripulados, que funcionaria melhor se uma frota vasta de tais veículos fosse aplicada, pois assim a comunicação entre os veículos aprimoraria muito a rota, a velocidade e a precisão do andamento em trânsito (HARARI, 2016, p. 383), caso ainda sejam propensos a Ataques Adversários poderia causar a morte ou lesão de milhões de pessoas de uma vez só.

Não é nem um alarmismo, é apenas uma consequência de desinformação ou ignorância de um fator de risco. Isso se torna mais real e perigoso quando se pensa que a limitação da computação está numa limitação matemática, ou seja, praticamente é necessário que um paradoxo matemático (COLBROOK et al., 2020) seja superado para que seja possível uma IA geral, por exemplo. Nesse contexto é que as propostas de Railton e dos filósofos que defendem o estilo *bottom-up* em ética se enfraquece. Como confiar em um sistema que pode ser iludido? Se, a justificativa para termos veículos não-tripulados se baseia na desconfiança sobre o humano de seguir padrões e a lei de trânsito, o que garante que não serão os mesmos inconfiáveis que burlarão com ataques adversos sistemas inteligentes autônomos? Não basta apenas o desejo e o furor das empresas em implementarem sistemas assim, parece que, ao menos por enquanto, é jogar com a sorte. Alguém poderia afirmar que o humano também está no ambiente das probabilidades assim como as IAs, no entanto, apenas saber que sistemas inteligentes quanto mais complexo, menos precisos e, portanto, mais vulneráveis a ataques desse tipo, acenderia uma luz de alerta. Semelhante a dar chaves para um condutor que vai a uma festa e se sabe de antemão que ele tem tendências alcoólatras. Isso sem contar toda a possibilidade de jogos comerciais que podem provocar instabilidades em empresas ligadas a esse tipo de tecnologia com objetivo de sabotar a concorrência, bem como sofrer com ações de ativistas[34].

Enfim, parece que o realismo moral defendido por Railton ainda é insuficiente para responder as necessidades de regramentos em máquinas, mesmo que fosse possível

34 Sobre ataques desse tipo, disponível em: https://www.terra.com.br/noticias/rs-sob-ataque-ciberguerra-abala-provedores-de-internet-no-estado,26bce91f417cdb2671903a1c4ee18bb8gvkmqke3.html.

Polymatheia
Revista de Filosofia

tal maneira de programação, o Equilíbrio Reflexivo parece ser ainda mais eficiente se fosse o caso, pois traria o consenso e a consistência mais fiáveis, principalmente em relação a possibilidade de revisão de padrões prejudiciais para a sociedade. Enquanto não surge uma alternativa robusta de tecnologia capaz de lidar com situações complexas com alta confiança, algo mais *realista* talvez seja supervisionando ou mesmo a limitando a massificação do uso de tecnologias inteligentes autônomas que podem causar danos irreparáveis as pessoas.

Existe um caminho longo de desenvolvimento de tecnologias inteligentes digitais, talvez hoje o maior benefício em termos de reflexão filosófica seja essa volta sobre nós mesmos que somos provocados a explicar junto a escolha sobre um estilo ou outro de pensar a ética, sobre como justificamos o conhecimento moral, de que maneira o discurso moral atua entre os humanos e toda a gama de questionamentos metaéticos que surgiriam para fundamentar uma implementação ética em IAs. Isso porque simplesmente afirmar que Direitos Humanos, por exemplo, devem ser respeitados e que outros princípios devem ser agrupados em torno das IAs não é suficiente, parece ser necessário uma construção mais rigorosa sobre como chegamos a tais princípios e fundamentos. E esse rigor passa por tentar compreender cada vez mais em que estágio está a tecnologia digital e o que a Filosofia pode contribuir para aprimorar ou criticar os pontos frágeis em tal campo do saber. Ao passo que essa pesquisa projeta ainda mais a possibilidade de novas discussões na área.

REFERÊNCIAS

ADRIAANS, Pieter. Information. *In* The Stanford Encyclopedia of Philosophy (Winter 2023 Edition). Edward N. Zalta & Uri Nodelman (eds.). Disponível em: <https://plato.stanford.edu/archives/win2023/entries/information/>. Acesso em: 18 Dez. 2023.

ANGIUS, Nicola; PRIMIERO, Giuseppe; Turner, Raymond. The Philosophy of Computer Science. *In* The Stanford Encyclopedia of Philosophy (Spring 2021 Edition). Edward N. Zalta (ed.), Disponível em: <https://plato.stanford.edu/archives/spr2021/entries/computer-science/>. Acesso em: 05 dez. 2023.

ARTIFICIAL Intelligence: A modern approach 4th ed. *In* Berkley.edu. Califórnia,

Polymatheia
Revista de Filosofia

[2023?]. Disponível em: https://aima.cs.berkeley.edu/. Acesso em: 08 Dez. 2023.

ATHALYE, Anish et al. Synthesizing robust adversarial examples. *International conference on machine learning*. PMLR, 2018. p. 284-293.

AWAD, Edmond et al. The Moral Machine Experiment. *Nature*. v. 563, 2018, p. 59-64.
BBC. Entenda o escândalo de uso político de dados que derrubou valor do Facebook e o colocou na mira de autoridades. *In* G1. São Paulo, 20 mar. 2018. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/entenda-o-escandalo-de-uso-politico-de-dados-que-derrubou-valor-do-facebook-e-o-colocou-na-mira-de-autoridades.ghtml>. Acesso em: 28 Dez. 2023.

BENEVENIDES, Bernardo. Os algoritmos de machine learning. *In* GEEKHunter, [São Paulo], 2018. Disponível em:https://blog.geekhunter.com.br/aprendizado-de-maquina-e-seus-algoritmos/. Acesso em: 08 Dez. 2023.

BESSEMER VENTURE PARTNERS. Cybersecurity trends in 2024. Disponível em: <https://www.bvp.com/atlas/cybersecurity-trends-in-2024>. Acesso em: 25 out. 2024.

BRINGSJORD, Selmer. GOVINDARAJULU, Naveen Sundar. Artificial Intelligence. The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.) Disponível em: <https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/>. Acesso em: 05 dez. 2023.

BROOKS, David. The Philosophy of Data. New York Times. Nova Iorque, 04 de fev. 2013. Disponível em: <https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html>. Acesso em: 05 Dez. 2023.

COITINHO, Denis. Prudent Reflective Equilibrium. **Ethics in Progress**, v.14, n.1, 2023, p. 46-63.

COLBROOK, Matthew J.; ANTUN, Vegard; HANSEN, Anders C. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. *Proceedings of the National Academy of Sciences*, v. 119, n. 12, p. e2107151119, 2022.

Comissão Europeia. Orientações éticas para uma IA de confiança. Direção-Geral das Redes de Comunicação, Conteúdos e Tecnologias. Serviço das Publicações: 2019. Disponível em: <https://data.europa.eu/doi/10.2759/2686>. Acesso em 29 Dez. 2023.

DIGNUM, Virginia. *Responsible Artificial Intelligence*: How to Develop and Use AI in a Responsible Way. Cham: Switzerland, 2019.

DOS SANTOS HASHIMOTO, Caroline Cavilha Cavilha. Dataísmo, a religião do século

XXI e sua manifestação do sagrado no filme I Am Mother (2019). *Epígrafe*, v. 10, n. 1, p. 537-554, 2021.

ETZIONI, Amitai; ETZIONI, Oren. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics***.** v. 21, n. 4, 2017, p. 403-418.

EVTIMOV, Ivan et al. Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945, v. 2, n. 3, p. 4, 2017.

HARARI, Yuval Noah. *Homo Deus*: uma breve história do amanhã. São Paulo: Editora Companhia das Letras, 2016.

KAUFMAN, Dora. Resenha do livro: "Ethics of artificial intelligence", de Matthew Liao. *TECCOGS*: Revista Digital de Tecnologias Cognitivas, n. 23, 2021.

LEE, Kai-fu. *AI superpowers*: China, Silicon Valley, and the new world order. Boston: Houghton Mifflin Harcourt, 2018.

LIAO, S. Matthew (Ed.). *Ethics of artificial intelligence*. Oxford University Press, 2020.

LISBOA, Alveni. O que é um algoritmo? *In* CANALTECH. [São Paulo], 8 out. 2022. Disponível em: <https://canaltech.com.br/inteligencia-artificial/o-que-e-um-algoritmo-226839/>. Acesso em: 08 Dez. 2023.

MILLER, Alexander. *An introduction to contemporary metaethics*. Maiden: Polity Press, 2003.

RAILTON Peter. *In.* LSA Philosophy University of Michigan. Michigan, [2023?]. Disponível em: https://lsa.umich.edu/philosophy/people/faculty/prailton.html. Acesso em: 29 Dez. 2023.

RAILTON, Peter. Ethical Learning, natural and artificial. *In*: LIAO, S. Matthew (Ed.). Ethics of Artificial Intelligence. New York: Oxford University Press, 2020, p. 45-78.

RAILTON, Peter. *Facts, values, and norms: Essays toward a morality of consequence*. Cambridge University Press, 2003.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial intelligence a modern approach*. London, 2019.

SAVULESCU, Julien; GYNGELL, Christopher; KAHANE, Guy. Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies. *Bioethics*, 2021, p. 1-12.

SHAFER-LANDAU, Russ. *Moral realism*: A defence. Nova York: Clarendon Press,

2003.

SILVEIRA, Paulo Antônio Caliendo Velloso da. Ética e inteligência artificial: da possibilidade filosófica de agentes morais artificiais. Porto Alegre: Edipucrs, 2020.

SULLINS, John. Information Technology and Moral Values. The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), Disponível em:<https://plato.stanford.edu/archives/fall2023/entries/it-moral-values/>. Acesso em: 25 nov. 2023.

SZEGEDY, Christian et al. Intriguing properties of neural networks. *arXiv preprint arXiv*:1312.6199, 2013.

TASIOULAS, John. First Steps Toward an Ethics of Robots and AI. *Journal of Practical Ethics*, v. 7, n. 1, 2019, p. 61-95.

THOMSON, Judith Jarvis. The trolley problem. *Yale LJ*, v. 94, p. 1395-1415. 1984.
WANG, Yulong et al. Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. *arXiv preprint arXiv*:2303.06302, 2023.

WHAT'S is Artificial Intelligence?. *In* IBM.COM. Disponível em: <https://www.ibm.com/topics/artificial-intelligence>. Acesso em: 05 dez. 2023.