VIRTUALIA JOURNAL
UFOP DEFIL Programa de Pós-graduação
https://periodicos.ufop.br/virtualia-journal
Editor responsável: Prof. Dr. Rodrigo Cid

IA, para quê te quero? Sobre agência e responsabilidade em tempos tecnológicos.

Sabrina Alves Aragão Lima
Universidade Estadual do Ceará
https://orcid.org/0009-0004-3470-3684
http://lattes.cnpq.br/9429320549107935
sabrina.aragao93@gmail.com

Resumo: Este artigo investiga a possibilidade de atribuir agência moral à Inteligência Artificial (IA). Primeiramente, discute-se a viabilidade dessa agência e, caso ela exista, qual seria a sua natureza. Em seguida, analisa-se os desafios da responsabilização tendo em vista a opacidade dos sistemas autônomos, dentre eles o *Deep Learning*. Por fim, colocam-se em diálogo as propostas de David Gunkel e Mark Coeckelbergh. Enquanto o primeiro defende a paciência moral, deslocando o foco da capacidade de agir para a de ser afetado, o segundo desenvolve uma abordagem socio-relacional, na qual o status moral é constituído nas relações estabelecidas com as IAs. Conclui-se que a visão de Coeckelbergh se mostra mais adequada para pensar a ética na IA, pois a verdadeira responsabilidade pressupõe transparência e capacidade de resposta.

Palavras-chave: Agência moral. Inteligência artificial. Ética.

Abstract: This paper examines the possibility of attributing moral agency to Artificial Intelligence (AI). First, it discusses the feasibility of this agency and, in case it exists, what its nature would be. Then, the challenges of accountability are analyzed, taking into account the opacity of autonomous systems, such as Deep Learning. Finally, the paper puts into dialogue David Gunkel's and Mark Coeckelbergh's proposals. While the former defends moral patiency, moving the focus from the capacity to act to how AIs can be affected, the latter develops the socio-relational approach, in which moral status is built upon the relationships established with AI. It is concluded that Coeckelbergh's proposal is more appropriate for thinking about ethics in AI, since true responsibility presupposes transparency and responsiveness.

ISSN 3086-3899

……..

Keywords: Moral agency. Artificial intelligence. Ethics..

CRediT-IA (+): Declaro que o uso de ferramentas de Inteligência Artificial foi estritamente instrumental e que assumo a responsabilidade humana integral pelo conteúdo do artigo. Simetricamente, autorizo o uso instrumental de ferramentas de Inteligência Artificial pelos pareceristas.
Ferramenta de IA utilizada: Deepseek
Versão / modelo: DeepSeek-V3
Finalidade do uso: O uso de Inteligência Artificial teve como finalidade exclusiva o aprimoramento da qualidade textual do artigo. A ferramenta foi empregada pontualmente como assistente linguístico para revisão gramatical, correção ortográfica, adequação do tom acadêmico e melhoria da fluidez e clareza da escrita, sem qualquer intuito de gerar conteúdo inédito, análises científicas ou dados para a pesquisa.
Tipo de uso: Trata-se de um uso assistido de IA de caráter secundário e estritamente editorial/estilístico. A ferramenta Deepseek atuou na fase de pós-redação como suporte para ajustes na estrutura, reestruturação da coesão de frases e sugestões de sinonímia e estilo, sendo todas as alterações propostas devidamente checadas, pela autora.
Limites do uso: A Deepssek não foi utilizada na concepção e delimitação da ideia central do artigo, na definição da sua estrutura de seções, na pesquisa e seleção das referências bibliográficas, nem na construção da fundamentação teórica. Da mesma forma, a formulação de toda a linha argumentativa, a análise crítica dos conceitos e a elaboração das conclusões foram realizadas apenas pela autora.

Veja o modelo completo de Declaração CRediT-IA, criado pelo Virtualia Journal.

"A Máquina não pode, não deve, nos tornar infelizes."
Isaac Asimov

……..

Introdução

O desenvolvimento tecnológico impulsiona a criação de ferramentas e sistemas que prometem otimizar tarefas e resolver problemas do cotidiano. No entanto, esse processo corriqueiramente vai modificando as relações, modos de produção e até mesmo nossa forma de pensar, o que evidencia dilemas éticos que demandam análise sobre seus limites e consequências. Dentre essas tecnologias, a Inteligência Artificial (IA) se destaca pela capacidade de agir em sociedade, como os sistemas autônomos, em específico o *Deep Learning*, cuja atuação, muitas vezes imperceptível, já está presente em setores que vão desde diagnósticos médicos até o uso de armas letais autônomas. Essa autonomia, contudo, não é neutra, sendo moldada por vieses, valores institucionais e decisões de projeto que nem sempre são acessíveis aos usuários ou aos afetados por suas decisões.

A dificuldade de compreender como sistemas autônomos chegam às suas decisões é uma característica estrutural dos modelos atuais de aprendizado profundo, em que bilhões de parâmetros operam de forma que nem seus próprios criadores conseguem explicar seus critérios de decisão (Coeckelbergh, 2023). Um exemplo é o sistema COMPAS, usado em tribunais estadunidenses para avaliar risco de reincidência. Investigações revelaram que ele classificava erroneamente réus negros, classificando-os como de alto risco, em uma frequência muito maior do que réus brancos, ainda que com históricos similares. Mas seu funcionamento interno permanecia opaco, protegido por segredo comercial[1].

[1] Caso de Paul Zilly. H. FRY. Hello world: Being human in the age of algorithms. NY: W. W. Norton, 2018, pp. 71-72

Assim, percebemos que conceitos antes bem definidos, como responsabilidade e agência, tornam-se cada vez mais obscuros com a presença dos sistemas de IA. Os conceitos filosóficos tradicionais de responsabilidade foram pensados em um contexto no qual as ações humanas eram as únicas preocupações morais, pois envolviam intenção, controle e conhecimento do que se faz. Os sistemas autônomos contemporâneos, no entanto, agem sem intenção, mas produzem consequências moralmente relevantes. Coloca-se então a pergunta central deste artigo: como atribuir consideração moral a entidades que desafiam as definições tradicionais de agência?

Para desenvolver essa proposta, recorro a dois filósofos contemporâneos cujas obras apontam nessa mesma direção: o estadunidense David Gunkel e o belga Mark Coeckelbergh. Gunkel critica os critérios tradicionais de agência moral, apontando uma limitação conceitual provocada pelo avanço tecnológico, o que gera uma lacuna ética sobre causas e consequências no cotidiano. Como alternativa, ele propõe a paciência moral, que desloca o foco da capacidade de agir para a possibilidade de ser afetado. Coeckelbergh, por sua vez, desenvolve uma abordagem sócio-relacional, influenciada pela hermenêutica ontológica heideggeriana, na qual o status moral emerge da interação entre humanos e sistemas autônomos e se torna possível por meio de maior explicabilidade e transparência, já que apenas humanos explicam coisas para outros humanos. Ambas as perspectivas convergem ao deslocar a pergunta do que a IA é para como nos relacionamos com ela, ainda que por caminhos distintos.

Antes de avançar, duas delimitações metodológicas são necessárias. A primeira diz respeito ao escopo da análise: não parto de projeções futuristas sobre superinteligência, mas da IA algorítmica contemporânea, tal como ela se apresenta em sistemas de recomendação, diagnósticos automatizados, veículos autônomos e assistentes virtuais. A segunda delimitação é de ordem disciplinar: embora a filosofia

da mente ofereça um caminho para discutir estados mentais e intencionalidade, a análise aqui proposta é ética. Isso significa que o interesse recai sobre as implicações morais da atuação da IA, e não sobre a determinação de sua eventual intencionalidade ou consciência.

O artigo está organizado em três momentos. No primeiro, mostro como o conceito tradicional de agência moral é tensionado pelo avanço da IA, apresentando as posições antagônicas de Johnson (2006) e de Anderson e Anderson (2010), e discutindo o chamado *responsibility gap* de Matthias (2004). No segundo, examino as condições aristotélicas de responsabilidade e as limitações da abordagem distributiva proposta por Floridi (2016), apontando as dificuldades de atribuir responsabilidade em sistemas complexos. No terceiro, coloco Gunkel (2012) e Coeckelbergh em diálogo para mostrar como suas perspectivas se complementam.

1. Agência sem agente

Na história da filosofia ocidental, sempre houve uma preocupação com a boa orientação das ações humanas em sociedade, essa é uma constante que remonta aos pré-socráticos e perpassa toda a tradição, desde as reflexões gregas sobre a virtude até as investigações contemporâneas sobre a justiça e o comportamento moral. (Russell, 2015). O conceito de agência moral, por exemplo, à primeira vista, aparece bem fundamentado e delimitado nessa tradição. Durante séculos, a figura humana sempre esteve como referência central. Em Kant, esse conceito esteve relacionado ao "ser racional finito" (Tavares; Ferro, 1997, p. 80), para quem o valor moral de uma ação reside na intenção (vontade) de agir conforme o dever, e não nas consequências ou resultados que essa ação produz. Isso significa que uma ação só tem genuíno valor

moral quando é realizada por respeito à lei moral, e não por inclinações ou interesses pessoais. Além disso, a liberdade aparece como condição da agência, sendo entendida como autonomia, ou como uma capacidade de se autodeterminar segundo princípios racionais que o próprio agente reconhece como universalmente válidos.

Essa compreensão conceitual situava o humano em uma posição distinta entre o mundo físico e o mundo da razão. "Essa separação já implicava uma oposição entre o humano e as coisas mecânicas" (Nascimento; Santos, 2018, p. 66-67). Enquanto estas obedecem à leis de forma determinista, sendo governadas por uma causalidade externa, o humano possuiria autonomia para orientar sua conduta pela razão prática. Nesse sentido, a moral não poderia ter sua fonte em algo exterior ao homem, como ocorria nos modelos éticos anteriores, mas deveria residir no próprio individuo, de modo que a razão assumiria a função de legislar as normas que orientam a ação humana. A agência moral, era expressão de liberdade e não de causalidade mecânica. Essa distinção, a meu ver, ainda hoje orienta nossa maneira de pensar a agência, mesmo quando nos deparamos com sistemas que desafiam essa separação.

Alguns sistemas de IA nos faz questionar se ainda podem ser vistos apenas como ferramentas. Diferentemente das máquinas tradicionais, que se limitam a tarefas predefinidas, os sistemas autônomos atuais são utilizados para em processos de decisões que geram impactos morais, especialmente quando ocorrem erros não previstos. A confiança depositada nesses sistemas, no entanto, baseia-se principalmente em sua eficiência e capacidade de processamento, o que pode levar a uma dependência excessiva (over-reliance). Essa dependência se torna ainda mais preocupante porque a opacidade desses sistemas dificulta a compreensão plena de seu funcionamento e limita a supervisão humana, de modo que os usuários tendem a delegar decisões sem ter clareza sobre os critérios que as orientam (Taddeo; Floridi,

2018). Diante dessa lacuna ética, seria possível atribuir agência a esses sistemas a ponto de responsabilizá-los por seus atos?

Diversas correntes filosóficas debatem essas questões sustentando posições antagônicas. A primeira é defendida por Johnson (2006), para quem os sistemas computacionais são entidades morais, mas não agentes morais. Ela critica tanto a visão que nega qualquer relevância moral aos computadores quanto a que os trata como agentes morais plenos, pois negar que computadores sejam agentes morais não equivale a negar que tenham importância moral. Seu argumento central é que os computadores têm intencionalidade, mas não possuem estados mentais nem intenção de agir. Eles estão predispostos a se comportar de certas maneiras em resposta a determinadas entradas de dados, mas não têm livre-arbítrio. A intenção de agir, segundo a autora, é o que distingue um agente moral de um mecanismo, pois está ligada à liberdade, algo que sistemas computacionais não possuem.

A partir dessa análise, a autora propõe uma tríade de intencionalidade composta por projetista, sistema e usuário. O projetista cria o sistema com certa intencionalidade, o sistema a incorpora em seu funcionamento e o usuário a ativa ao utilizá-lo, de modo que nenhum dos três age sozinho. Para ilustrar essa dinâmica, a autora compara um sistema computacional a uma mina terrestre. Uma mina explode anos depois de ser plantada e mata uma criança. Nenhum humano pretendia aquele resultado, mas a explosão não é um evento natural isolado, é o resultado da intencionalidade do projetista (que concebeu a mina para explodir sob pressão), do soldado (que a instalou no local) e da própria mina (que incorpora essa intencionalidade). O mesmo raciocínio se aplica a um sistema computacional, pois sua ação é sempre parte de uma cadeia que começa na intencionalidade humana. Portanto, o sistema é um componente da agência

moral humana, não um agente independente. A IA é um instrumento moralmente relevante, mas não uma autora de decisões morais.

Em oposição a essa visão, Anderson (2010) defendem a posição contrária à de Johnson. Para eles, não apenas é possível, mas necessário criar sistemas autônomos que se comportem eticamente. Para chegar a essa conclusão, eles examinam seis abordagens possíveis, apontando as limitações das cinco primeiras. A primeira é a programação rígida para prevenir ações antiéticas, mas é impossível antecipar todas as circunstâncias possíveis. A segunda é transferir o ônus da responsabilidade ao usuário, mas usuários podem ignorar advertências. A terceira é aprender com pesquisas de opinião ou decisões passadas, o que reflete preconceitos históricos e populares. A quarta é usar uma única teoria ética existente, como o utilitarismo ou a deontologia. A quinta é impor uma hierarquia de princípios, como as leis de Asimov, nas quais deveres inferiores nunca podem sobrepor os superiores. Diante da insuficiência dessas cinco alternativas, os autores propõem uma sexta via: sistemas devem ser programados para descobrir, por meio de aprendizado de máquina e diálogo com especialistas em ética, quais características de uma situação são moralmente relevantes e quais deveres provisórios se aplicam àquele contexto, de modo que possam tomar decisões balanceando obrigações conflitantes (Anderson; Anderson, 2010). Um exemplo prático é o robô de cuidado a idosos desenvolvido pelos autores, que precisa decidir quando lembrar um paciente de tomar remédio e quando aceitar sua recusa, equilibrando os deveres de prevenir dano e respeitar a autonomia.

Diferentemente de Johnson, que situa a IA como uma entidade moral desprovida de agência plena, os Andersons defendem que a estrutura da ética pode ser computada e que as máquinas podem se tornar agentes morais autônomos. Para eles, programar sistemas para agir eticamente não apenas assegura seu

comportamento, como também podem tornar a ética humana mais clara, ao resolver inconsistências e revelar princípios até então não formulados. Mais do que isso, tais sistemas poderiam servir como modelos ideais para inspirar os próprios humanos a se comportarem de maneira mais ética, uma vez que a maioria das pessoas está longe de ser um agente ético exemplar, tendendo a favorecer a si mesmas em vez de agir por princípios universais (Anderson; Anderson, 2010). Um robô programado para aplicar consistentemente um princípio de justiça, por exemplo, não se cansaria nem se frustraria como um cuidador humano, revelando assim o que seria uma ação imparcial e forçando uma reflexão sobre os próprios hábitos morais.

Ao observar ambas as posições, percebo que elas talvez se limitem a uma dualidade focada em definir se a IA é ou não um agente moral. A ideia implícita parece ser a de que, uma vez resolvida essa questão com base em certos critérios éticos tradicionais, tornar-se-ia mais fácil decidir como tratá-la em sociedade. Entretanto, antes de determinar se esse sistema pode ocupar o lugar que tradicionalmente foi ocupado pelo humano, talvez interessa reconhecer que sua presença já está alterando práticas coletivas, independentemente de qualquer definição conclusiva sobre sua agência. Por exemplo, sistemas de recomendação já definem o que vemos nas redes sociais e influenciam decisões de consumo. Algoritmos de triagem priorizam currículos em processos seletivos, automatizam diagnósticos médicos e até sugerem sentenças em tribunais (Coeckelbergh, 2023).

Esses usos demostram que a IA já atua como mediadora em vários aspectos de nossas vidas, e que seus efeitos talvez não dependam de uma definição conclusiva sobre sua natureza, seja ela baseada em parâmetros técnicos, em eficiência operacional ou na ausência de habilidades morais. Uma vez que a "moral não pode ser reduzida

ao seguimento de regras nem se limita às emoções humanas, ainda que estas últimas sejam indispensáveis para os julgamentos morais" (Coeckelbergh, 2023, p. 53).

A reflexão que considero relevante, portanto, não é decidir se a IA possui ou não agência, mas compreender como sua atuação nos convoca a repensarmos as categorias filosóficas com as quais tradicionalmente pensamos a moral. Uma vez que a IA já atua como mediadora em aspectos centrais da sociedade, a pergunta que se impõe talvez não seja mais se ela merece ser considerada um agente moral segundo parâmetros que tomam o humano como modelo de comparação, mas sim que tipo de responsabilidade estamos dispostos a assumir diante das ações realizadas por esses aparatos.

Com essa situação, surge uma lacuna, ou melhor, *responsibility gap (*Matthias, 2004), conceito que descreve situações nas quais sistemas autônomos causam danos sem que haja um agente claramente responsabilizável. De um lado, os desenvolvedores e instituições não podem ser responsabilizados, pois o sistema aprende e se adapta de maneiras imprevisíveis, escapando ao controle do projeto inicial. Por outro lado, os usuários tampouco podem ser considerados responsáveis, já que não compreendem plenamente as decisões que o sistema toma. O próprio sistema, por sua vez, não se enquadra nos critérios tradicionais de imputação moral, uma vez que lhe falta intencionalidade, consciência ou livre-arbítrio. O que se evidencia, portanto, é uma zona de indeterminação ética em que a ação ocorre sem acompanhamento claro de uma responsabilização.

Uma das hipóteses que se pode formular é que essa nova configuração da responsabilidade não afeta todos os agentes da mesma maneira. Para as instituições que implementam essas tecnologias, a opacidade algorítmica talvez funcione como uma blindagem, dificultando a atribuição de falhas. Para aqueles que sofrem os efeitos de decisões automatizadas, a mesma opacidade pode se traduzir em ausência de

reparação. A lacuna de responsabilidade, nesse quadro, tenderia a resguardar uns enquanto desampararia outros.

É precisamente nesse ponto que a noção de zona de deformação moral, desenvolvida por Elish (2019), se torna relevante. Assim como a zona de deformação de um carro é projetada para absorver o impacto de uma colisão e proteger os ocupantes, a zona de deformação moral protege a integridade do sistema tecnológico, ainda que às custas do operador humano mais próximo. Esse mecanismo se torna evidente quando observamos, por exemplo, o caso dos carros autônomos, em que empresas oferecem a praticidade de assumir a direção dos veículos, mas, ao mesmo tempo, exigem que os motoristas humanos permaneçam no controle, sem, no entanto, arcar com a responsabilidade implicada pela tarefa de condução. O operador humano, nesse esquema, torna-se a zona de deformação moral do sistema, absorvendo a responsabilidade por falhas que não pode evitar, protegendo a empresa e a tecnologia, enquanto a verdadeira autonomia do sistema permanece intocada.

Essa dificuldade de responsabilização se aprofunda quando consideramos o modo como esses sistemas operam. Frequentemente descritos como caixas-pretas, eles tornam inacessível o processo interno que conduz de um dado de entrada a uma decisão ou ação (Pasquale, 2015). Em outras palavras, embora o resultado seja obtido o percurso que levou a esse resultado permanece obscurecido. No contexto da IA, isso significa que um algoritmo pode gerar uma ação sem que seja possível rastrear os critérios exatos que a motivaram. E se não sabemos como uma decisão foi tomada, como poderíamos responsabilizar alguém por ela?

2. Quem responde pela IA?

Na filosofia aristotélica, a responsabilidade está diretamente vinculada à distinção entre ações voluntárias e involuntárias, uma vez que, "apenas ações voluntárias recebem louvor ou censura; aquelas que são involuntárias são perdoadas, e às vezes até mesmo lamentadas" (Aristóteles, 1984, 1109b30-35), ou seja, uma ação só pode ser atribuída moralmente a alguém se houver controle e conhecimento por parte do agente. Isso significa que, para que uma pessoa seja culpada ou elogiada por seus atos, a ação deve partir dela de forma voluntária e consciente, ou seja, o agente deve estar ciente das circunstâncias do que está fazendo. Essa perspectiva estabelece duas condições: "(I) sob que condições alguém pode ser moralmente responsabilizado por suas ações e (II) quando alguém pode ser dito livre para executar suas ações" (Andreazza, 2014, p. 149).

Aristóteles, no entanto, reconhece que nem toda ação se enquadra perfeitamente nas categorias de voluntário e involuntário, existindo uma classe intermediária de ações que ele denomina ações mistas. Essas ações são voluntárias no momento da execução, mas involuntárias em sua origem, e sua caracterização revela a complexidade da teoria aristotélica da responsabilidade. O exemplo clássico apresentado por Aristóteles é o do marinheiro que lança a carga ao mar para salvar o navio. O ato de lançar a carga, por um lado, "parece ser involuntário, pois em situações normais (não impostas por fatores externos) o agente não lançaria a carga valiosa (não é uma ação escolhida por si mesma), mas por outro é voluntário, uma vez que o princípio motor está no agente quando este arremessa a carga para fora do navio" (Andreazza, 2014, p. 166). Aristóteles entende que "atos assim são 'mistos', mas assemelham-se mais a atos voluntários pela razão de serem escolhidos no momento

em que se fazem e pelo fato de ser a finalidade de uma relativa às circunstâncias" (Aristóteles *apud* Andreazza, 2014, p. 167). O que torna tais ações mais próximas do voluntário é o fato de serem escolhidas no momento em que se realizam, sendo a finalidade da ação relativa às circunstâncias particulares.

A relevância das ações mistas para a teoria da responsabilidade reside no fato de que, embora sejam voluntárias no momento da execução, algumas delas compartilham a não imputabilidade moral dos atos involuntários (Andreazza, 2014). Isso significa que, mesmo quando o agente age voluntariamente, pode haver circunstâncias favoráveis que justificam indulgência em vez de censura. Aristóteles, contudo, "entende que certas ações mistas são moralmente imputáveis (por exemplo, um pai que destrói a cidade inteira para satisfazer a exigência que um sequestrador impõe para libertar a sua filha) e que o indivíduo deve suportar o sofrimento (nesse caso a perda de uma filha) e não executar a ação" (Andreazza, 2014, p. 167). O reconhecimento dessas ações intermediárias demonstram que a teoria aristotélica não atua com uma divisão rígida entre responsabilidade plena e isenção total, mas admite níveis na atribuição de responsabilidade moral. Essa complexidade, contudo, não compromete o centro da teoria, pois Aristóteles mantém que a responsabilidade fundamental "está em nosso poder agir e também não agir" (Andreazza, 2014, p. 150).

Essa concepção de responsabilidade oferece um eixo conceitual relevante para pensarmos o desenvolvimento e o uso da inteligência artificial. Diferentemente das tecnologias anteriores, que se destinavam a mediar ou ampliar as capacidades sensoriais ou físicas dos usuários, os sistemas autônomos são apresentados como uma tecnologia potencialmente capaz de substituir as capacidades humanas de raciocínio e tomada de decisão (Constantinescu et al., 2021). Nesse contexto, à medida que "os seres humanos delegam o controle sobre a tomada de decisões para as tecnologias de IA,

parece que eles também abrem mão de conhecer como essas decisões são feitas" (Davenport, 2014; Hakli; Mäkelä, 2019; Loh; Loh, 2017 *apud* Constantinescu et al., 2021, p. 803). Mas, até que ponto isso desafia a própria noção de responsabilidade tal como concebida na tradição aristotélica, principalmente em casos de erros de sistemas de IA?

Com a tradição aristotélica, esse aspecto de distribuição de responsabilidade é bem definido ao situar o humano como ponto central ao avaliar em que situação não ocorreu o nível de conhecimento e o controle necessários. Com a IA, esse sentido talvez seja tensionado porque a condição epistêmica para a responsabilidade moral em Aristóteles exigiria que o agente conhecesse as circunstâncias particulares da ação que realiza. Isso significaria que, para ser responsável, o agente precisaria saber o que está fazendo, sobre quem está agindo, com quais instrumentos, com que finalidade e de que maneira. Com a IA, especialmente aquela baseada em aprendizado de máquina, a tecnologia não apenas substituiria o raciocínio humano, mas também tornaria opaco o processo decisório. Quando o humano delega uma decisão ao sistema, talvez ele deixe de conhecer como aquela decisão foi tomada, quais dados foram considerados ou quais critérios foram aplicados.

> Embora os sistemas de IA baseados em aprendizado de máquina possam exibir virtudes dianoéticas como *nous* ou episteme, que simplesmente exigem estudo, eles enfrentam um verdadeiro desafio quando se trata de virtudes dianoéticas como a phronesis, que exige prática. (CONSTANTINESCU et al., 2021, p. 807)

Nesse sentido, a IA pode, em certa medida, processar informações e identificar padrões de forma semelhante ao conhecimento teórico, mas lhe falta a capacidade de discernimento contextual exigida pela phronesis. Essa virtude, diferentemente do conhecimento teórico, só se adquiriria com experiência e prática cotidiana, elementos que um sistema artificial não possuiria. Sem phronesis, a IA não poderia exibir virtudes

éticas, pois não seria capaz de avaliar as particularidades de cada situação nem de agir voluntariamente no sentido aristotélico

Entretanto, mesmo restringindo a responsabilidade aos agentes humanos sua atribuição se mostra difícil, por dois motivos. Em primeiro lugar, os humanos não têm controle suficiente sobre a IA, o que torna problemático responsabilizar alguém por decisões que não pode nem mesmo acompanhar, como nos casos de sistemas de negociação de alta frequência ou defesas antimísseis que reagem em milissegundos. Em segundo lugar, há o problema que Coeckebergh (2019) chama de "muitas mãos", pois designers, programadores, empresas, usuários, reguladores, fabricantes de hardware e provedores de dados estão todos envolvidos no desenvolvimento e uso de um sistema de IA. Tornando-se quase impossível identificar um único responsável, já que a cadeia causal é longa, distribuída no tempo e no espaço, e frequentemente opaca.

Uma alternativa para esse impasse vem da abordagem de Floridi que parte do reconhecimento de que ações e decisões geradas por sistemas de IA surgem de uma teia de interações que envolve múltiplos agentes — programadores, empresas, usuários e instituições. A isso se dá o nome de agência distribuída, e a responsabilidade, nesse contexto, tende a acompanhar a mesma lógica de distribuição (Floridi, 2016). Para lidar com essa complexidade, o filósofo propõe um modelo de responsabilidade sem culpa, no qual a "atribuição de responsabilidade moral não depende da intencionalidade do agente ou do conhecimento prévio sobre sua ação, mas se concentra na identificação dos agentes causalmente envolvidos" (Floridi, 2016, p. 6). Esse modelo, ao responsabilizar todos os agentes da rede, exerce uma função preventiva, estimulando comportamentos mais cautelosos e atentos às consequências das interações.

Para ilustrar essa nova perspectiva da responsabilidade, Floridi recorre a dois casos da justiça holandesa. No primeiro, o dos três ciclistas, todos são igualmente responsáveis porque qualquer um deles poderia ter agido para evitar a irregularidade. No segundo, o dos quatro barcos, apenas o último barco é responsável porque era o único que poderia ter evitado a situação com facilidade. "A diferença entre os casos mostra que a atribuição da responsabilidade exige discernimento para identificar quem tinha capacidade de evitar o dano" (Floridi, 2016, p. 8-9). Aplicado à IA, esse princípio sugere que a responsabilidade pode ser distribuída de maneiras distintas, ou seja, se a falha é sistêmica e todos poderiam ter agido, a responsabilidade se espalha; se apenas um agente poderia ter evitado o dano, a responsabilidade recai sobre ele

Essa proposta, ainda que ofereça um caminho interessante, parece esbarrar em dificuldades de ordem prática. Para que a responsabilidade distribuída funcione, seria preciso conhecer, com um grau razoável de precisão, as contribuições de cada agente para o resultado final. Em sistemas complexos de IA, porém, esse conhecimento raramente está disponível. A história do desenvolvimento de um algoritmo pode envolver dezenas ou até centenas de programadores ao longo de anos, com partes de código que se tornam irreconhecíveis após sucessivas modificações. Os conjuntos de dados utilizados para treinamento passam por processos de coleta, seleção, limpeza e rotulagem que envolvem múltiplos agentes, muitas vezes em diferentes países e sob regimes legais distintos. Além disso, o software pode ser desenvolvido para uma finalidade e depois reaproveitado para outra completamente distinta, como no caso de sistemas de reconhecimento facial que migram da área médica para a vigilância policial. Nesse sentido, a responsabilidade não se reduz à atribuição causal, mas ainda assim a falta de rastreabilidade inviabiliza qualquer distribuição justa. (Coeckelbergh, 2019)

O segundo problema é que distribuir responsabilidade não significa distribuí-la de forma igualitária. Na ausência de critérios claros, uma partilha homogênea da culpa tende a beneficiar justamente os agentes mais poderosos. O caso do chatbot Tay, da Microsoft, ilustra bem essa questão. Diante dos comentários racistas e misóginos produzidos pelo bot, a mídia tendeu a distribuir a responsabilidade entre desenvolvedores, a empresa e os usuários. No entanto, os principais responsáveis eram aqueles que decidiram inserir o bot no Twitter, ou seja, os desenvolvedores e a Microsoft, não os usuários comuns[2]. Uma distribuição igualitária da responsabilidade, nesse caso, serviria aos interesses da empresa, que teria seu ônus diminuído pela dispersão da culpa. O problema mais profundo é que diferentes agentes têm diferentes poderes para ocultar suas falhas e direcionar a atenção para outros. Quando isso acontece, a responsabilidade distribuída se torna, na prática, uma responsabilidade evasiva onde ninguém é claramente responsável, e os mais vulneráveis acabam sendo os mais prejudicados.

O terceiro problema diz respeito à hierarquia das responsabilidades. Ainda que seja possível a responsabilidade distribuída, não se pode ignorar que alguns agentes têm muito mais controle sobre o sistema do que outros. O programador que escreve o código de sistema autônomo para um carro tem um nível de controle e, portanto, um nível de responsabilidade incomparavelmente maior do que o passageiro que utiliza o veículo. O executivo que decide lançar um produto de IA no mercado, mesmo ciente de seus riscos, tem uma responsabilidade diferente da do estagiário que testou o sistema superficialmente. A teoria aristotélica, por mais antiga que seja, oferece aqui um critério relevante. A responsabilidade está ligada à possibilidade de agir de outro modo e ao conhecimento das circunstâncias. Se um desenvolvedor podia, no momento

[2] https://revistazum.com.br/radar/inteligencia-artificial-racismo/

da programação, optar por uma arquitetura mais segura e não o fez, sua responsabilidade é maior do que a daquele que simplesmente seguiu ordens sem ter acesso às alternativas. Se um gerente sabia dos riscos e mesmo assim autorizou a implantação, sua responsabilidade é maior do que a do usuário final que não tem nenhuma capacidade de modificar o sistema.

Problematizar o modelo distributivo de Floridi, no entanto, não significa defender um retorno à responsabilização individualizada, mas observar que apesar de incluir a IA nas preocupações morais, é possível notar que ela ainda não oferece uma forma de responsabilização que acompanhe a essa nova dinâmica das relações sociais, assim como ainda não "é claro onde a IA termina e a outra tecnologia começa" (Coeckelbergh, 2019, p. 2058), ou seja, quando algo não funciona como o planejado, torna-se difícil direcionar a responsabilidade de forma precisa devido ao número de pessoas envolvidas e tecnologias interligadas. É nesse impasse que as perspectivas de Gunkel e Coeckelbergh podem ajudar.

3. Dois olhares sobre a ética da IA

Se fizermos uma análise do conceito de agência, perceberemos que ela possui um peso excludente entranhado em sua estrutura. Historicamente, os critérios que definem quem conta como agente foram moldados por grupos hegemônicos que se autodeclararam homens plenos, excluindo mulheres, escravizados e populações não brancas do reconhecimento moral pleno. Com o tempo, essa exclusão passou a ser questionada e transformada por meio de lutas sociais que reivindicaram mudanças, sempre com o objetivo de construir uma sociedade mais justa, ética e inclusiva. O

mesmo mecanismo de exclusão, agora se repete com a inteligência artificial. (Gunkel, 2012)

Contudo, ainda há grupos excluídos dessa consideração moral, como os animais e as máquinas com seus sistemas de IA. De fato, existem movimentos em defesa dos animais e leis que os protegem, mas será que a mesma lógica de inclusão moral pode ser aplicada às máquinas inteligentes? O primeiro obstáculo é que as IAs não lutam, não reivindicam e não organizam movimentos sociais. A lógica histórica da expansão moral, baseada na pressão política e na resistência dos excluídos, não se transfere diretamente para entidades que não possuem esse ímpeto. Diante disso, somos direcionados a um outro caminho, que questiona os próprios fundamentos usados para definir agência. Quando isso é feito, percebe-se que o "conceito de agente moral é incoerente e obsoleto" (Dennett, 1998, *apud* Gunkel, 2012, p. 47), de modo que, se aplicado com inflexibilidade, nem mesmo os humanos conseguiriam se qualificar como agentes morais plenos. Nessas circunstâncias, surge a proposta da paciência moral, que "não se preocupa em determinar o caráter moral do agente ou em avaliar o significado ético de suas ações, mas sim com a vítima, o destinatário ou o receptor de tais ações" (Gunkel, 2012, p. 48).

O conceito de paciência moral, nesse quadro, aponta para um redirecionamento ético que encontra inspiração no modelo dos direitos animais. A questão se desloca da identificação de quem é um sujeito moral legítimo para a capacidade de sofrer. Essa mudança de perspectiva demostra uma "revolução copernicana na filosofia moral" (Gunkel, 2012, p. 49), pois desafia o privilégio antropocêntrico e abre a comunidade moral para entidades antes marginalizadas. A inclusão das IAs, então, deixaria de depender de sua capacidade de lutar ou reivindicar e passaria a ser pensada a partir das relações que estabelecemos com elas. Crianças, animais, idosos em estado

vulnerável e talvez as IAs poderiam ser reconhecidos como pacientes morais justamente por estarem inseridos em relações de cuidado, responsabilidade ou dano potencial.

No entanto, como provar que máquinas sofrem? Essa pergunta se desdobra em algumas dificuldades (Gunkel, 2012). A primeira delas é que já existem máquinas que simulam dor. Engenheiros construíram mecanismos que agrupam possíveis emoções, como o robô de treinamento dental Simroid[3] que grita de dor quando estudantes o machucam. Se tomarmos a simulação como critério, isso aparentemente validaria a paciência moral. A segunda dificuldade é epistemológica. Não é possível saber como um outro ser sofre, pois a dor é uma experiência subjetiva e privada. Além de existir uma contradição em exigir testes de dor para aquilo que, por definição, não pode ser testado (Dawkins, 2005). A terceira dificuldade diz respeito à própria definição de dor que sequer sabemos definir o que é dor de maneira conclusiva (Dennett, 1998 *apud* Gunkel, 2012). O melhor que podemos fazer é dar conta das suas causas e efeitos, mas a dor em si mesma não aparece. Desta forma, não há uma teoria que explique a dor de forma definitiva e, portanto, nenhum computador ou robô poderia expressar tal emoção de maneira verdadeira. A quarta dificuldade e a principal de ponto de vista ético e moral, está na tentativa de construir máquinas capazes de sentirem dor para demonstrar sua paciência moral, esse próprio ato seria eticamente suspeito, pois causaria sofrimento desnecessário. Sendo assim, a paciência moral, com essas limitações, não oferece uma solução definitiva para o estado moral das máquinas, mas confirma problemas conceituais na própria estrutura ética, colocando em questão até mesmo os avanços alcançados os direitos dos animais.

[3] https://www.morita.com/group/en/products/educational-and-training-systems/training-simulation-system/simroid/

Mas até que ponto faz sentido utilizar parâmetros pensados para humanos e animais na tentativa de incluir as IAs na mesma categoria moral? A tentativa de estender a paciência moral às IAs parece esbarrar em uma dificuldade significativa, pois, ao determinar como critério o sofrimento ou consciência, está-se novamente submetendo-as a uma análise que, por definição, elas não podem preencher. Talvez, em vez de buscar para as máquinas um lugar fixo em categorias já estabelecidas, seja mais fecundo investigar não o que a IA é, mas o que ela faz e como isso atravessa as experiências cotidianas daqueles que com ela se relacionam. A pergunta, então, talvez não seja sobre a dor ou a consciência das máquinas, mas sobre a responsabilidade pelos efeitos que elas já produzem no mundo, independentemente de sentirem ou não. Em outros termos, "questão não é sobre a ética na IA ou por ela, mas sobre nossa ética para com a IA" (Coeckelbergh, 2023, p. 51). Nesse sentido, a IA seria uma preocupação ética, e não uma potencial agente ética em si, uma vez que, até o momento, essa tecnologia não age por si só, embora também não seja apenas uma ferramenta, mas uma mediadora ativa em sistemas morais complexos. Atribuir responsabilidade direta a esses sistemas por possíveis erros parece, assim, uma saída apressada. O que importa, antes, é compreender como nos dispomos a responder pelo que essas tecnologias fazem e pelo que fazemos com elas, especialmente quando seu funcionamento escapa ao nosso controle e entendimento.

No que se refere à paciência moral, Coeckelbergh acrescenta um elemento importante ao debate, a perspectiva fenomenológica inspirada na hermenêutica ontológica heideggeriana. Embora ele não explicite uma adesão à posição de Gunkel, é possível identificar uma aproximação parcial entre os dois. Ambos partilham da ideia ao afirmarem que a consideração moral nasce das relações, não de qualidades inerentes. No entanto, esse reconhecimento não é fruto de um teste empírico, mas se

dá em um processo interpretativo no qual sentidos são construídos na interação entre humano e máquina. Não há necessidade de provar que a IA possui alguma capacidade moral para então tratá-la com consideração, pois a própria relação já é suficiente.

> Por exemplo, se tratamos nosso gato com carinho, não é por conta do nosso envolvimento moral com o gato, mas porque já estabelecemos uma relação social com ele. Ele já é um animal de estimação e um companheiro antes que realizemos o trabalho filosófico de lhe atribuir uma consideração moral - se é que sentiríamos algum dia a necessidade de fazer tal exercício. E, se damos um nome a um bicho, como nosso cachorro, (em contraste com os outros animais sem nome que comemos), já lhe conferimos um status moral particular independentemente de suas qualidades objetivas. Utilizando uma abordagem relacional, crítica e não dogmática, poderíamos argumentar que, de maneira similar, o status moral das IAs será atribuído por seres humanos e dependerá de como elas estiverem inseridas em nossa vida social, na linguagem e na cultura humana. (COECKELBERGH, 2023, p. 60)

O que torna essa abordagem relevante nestas problemáticas levantadas neste artigo, é o deslocamento da pergunta para o plano da experiência. Não se trata de descobrir o que essa tecnologia é em si mesma, mas de observarmos como ela se manifesta em nosso mundo, que significados e valores lhe atribuímos no cotidiano e quem se importa com seu bem-estar ou integridade. Além disso, não a observamos de um lugar neutro, já que ela surge antes mesmo de qualquer reflexão explícita sobre seu *status moral*[4].

[4] O termo status moral (também chamado de posição moral) pode se referir a dois tipos de questões. A primeira se ocupa do que a IA é capaz de fazer moralmente falando – em outras palavras, se ela pode ter o que os filósofos chamam de agência moral, e se então poderia ser um agente moral pleno. O que isso significa? A segunda é que tipo de capacidades morais uma IA tem ou deve ter? (COECKELBERGH, 2023, p. 50-51)

A fronteira entre humanos e máquinas torna-se cada vez mais híbrida à medida que desenvolvemos sistemas colaborativos fundamentados na complementaridade de nossas capacidades. A IA atualmente é um dispositivo presente desde atividades corriqueiras como, assistentes virtuais que organizam nossas agendas ou algoritmos de recomendação de leituras ou mídias, até triagem de currículos em processos seletivos, análise de risco em empréstimos bancários e monitoramento de espaços públicos com reconhecimento facial. Numa escala maior de uso, ela chega ao campo militar, especificamente quando a empresa Shield AI, apresentou recentemente o conceito da aeronave X-BAT, um caça não tripulado controlado por um software de IA que pode operar sem comunicação com o mundo exterior[5] . O veículo é projetado para atuar como um autônomo, tomando decisões de voo e combate em tempo real ao lado de jatos pilotados por humanos, com capacidade de viajar mais de 3.700 quilômetros para realizar missões de ataque ou reconhecimento[6] . Diante disso, qual o limite desse uso? E mais, em que tipo de decisões queremos delegar tarefas a esses sistemas?

> Essa abordagem 'socio-relacional' ao status moral encoraja-nos a tentar entender as diferentes maneiras como as máquinas nos aparecem: às vezes como 'simples máquinas', às vezes como quasi-animais ou quasi-humanos. Além disso, também nos mostra que não existe um robô-em-si (coisa-em-si) e nenhum status moral 'puro', independente do observador. (COECKELBERGH, 2012, p. 44)

O que a abordagem socio-relacional evidencia é que o status moral não é uma qualidade, ou uma característica própria dos sistemas, mas algo que se constitui nas relações e práticas sociais. Isso significa que a mesma entidade técnica pode ser tratada

[5] https://www.channelnewsasia.com/business/silicon-valley-backed-shield-ai-enters-fighter-jet-race-new-wingman-drone-5418771?cid=cna_flip_070214

[6] https://www.tecnoandroid.it/2025/11/03/shield-ai-svela-x-bat-il-vtol-autonomo-con-autonomia-record-1653365/

de formas moralmente distintas, dependendo do contexto e das expectativas que mobilizamos. Quando destaco esse caráter relacional e interpretativo, porém, não estou sugerindo um relativismo moral. Uma vez que, essa perspectiva não elimina a possibilidade de avaliação crítica; ela apenas desloca o foco de critérios muito rígidos e metódicos para a análise das práticas, narrativas e relações que moldam a atribuição de status moral (Coeckelbergh, 2012). O 'aparecer', portanto, não é uma impressão subjetiva ou arbitrária, mas um fenômeno intersubjetivo, enraizado em histórias, lutas e práticas compartilhadas que podem ser avaliadas e contestadas.

As relações de gênero ao longo da história, por exemplo, ilustram bem essa dinâmica. Durante séculos, a mulher foi considerada naturalmente inferior ao homem, e essa suposta inferioridade serviu como justificativa para sua exclusão de direitos civis, políticos e educacionais. O que mudou essa situação não foi a descoberta de uma suposta essência feminina, mas transformações sociais profundas como: movimentos feministas que reivindicaram igualdade, mudanças na legislação que garantiram direitos, e a ressignificação cultural do papel da mulher na sociedade (Butler, 1990). O status moral da mulher, portanto, não é algo que se descobre, mas algo que se transformava nas relações sociais e nas lutas por reconhecimento. A abordagem relacional oferece, assim, critérios históricos e sociais para a consideração moral, entre os quais se destacam as lutas sociais que reivindicaram mudanças, as narrativas que ressignificaram quem merece consideração, as práticas culturais que herdamos e as formas de linguagem que utilizamos para descrever os entes. Ao não se prender a um sentido fixo, essa abordagem mantém a flexibilidade necessária para analisar como as relações se configuram em cada contexto, sem deixar de reconhecer que há assimetrias relevantes entre os agentes envolvidos.

Essa compreensão relacional do status moral da IA encontra desdobramentos na forma como entendemos a responsabilidade. Uma vez que, pensar em "IA responsável" é um tanto estranho, por entrar na lógica de entender essas máquinas como se fossem entidades independentes, ou detentores de deveres, quando na verdade a responsabilidade deveria permanecer vinculada às redes de atores humanos, que as criam, treinam e operam (Coeckelbergh, 2023). Portanto, fazer uma distinção entre responsabilidade causal e responsabilidade socio-relacional é fundamental para evitar que a máquina se torne um escudo ético para seus criadores. Enquanto a abordagem causal se concentra em determinar quem contribuiu para um resultado, a perspectiva relacional pergunta a quem se deve responder e quem pode legitimamente demandar explicações. Em outras palavras, "nós não temos apenas responsabilidade como agentes, uma responsabilidade causal e moral por nossas ações, mas também uma responsabilidade para com os outros" (Coeckelbergh, 2023, p. 2438).

Essa distinção não elimina a necessidade de investigar cadeias causais ou de identificar agentes que contribuíram para um evento, mas recoloca essas investigações em um quadro mais amplo, no qual a pergunta fundamental não é apenas 'quem causou o dano?', mas também 'quem deve explicações àqueles que foram afetados?'. A pergunta, então, não seria somente 'quem é o agente responsável?', mas também 'a quem se é responsável?' (Coeckelbergh, 2023, p. 2440). Enquanto a abordagem causal se concentra em rastrear agentes e suas contribuições, a perspectiva relacional busca compreender a responsabilidade como a capacidade de dar explicações perante o outro, e não apenas como agir com consciência.

Trata-se, assim, de uma relação dialógica na qual o paciente da responsabilidade é tão relevante quanto o agente. Essa dimensão, contudo, permaneceu invisível nas

discussões tradicionais sobre tecnologia, que sempre privilegiaram o agente e suas intenções (Gunkel, 2012). Essa diferença se torna particularmente relevante no contexto da IA, uma vez que os sistemas autônomos não apenas dificultam a identificação de agentes causais, mas também tornam mais complexa a compreensão de como um erro ocorreu para fins de atribuição de responsabilidade, na medida em que o conhecimento disponível se desloca da busca por causas para correlações e probabilidades (Coeckelbergh, 2023).

Podemos citar por exemplo, o acidente do carro autônomo da Uber em 2018, ocorrido em Tempe, Arizona. A colisão aconteceu quando Elaine Herzberg, uma pedestre que atravessava a rua fora da faixa de pedestres empurrando sua bicicleta, foi atingida por um veículo autônomo em modo de teste, que contava com uma motorista de segurança ao volante. De acordo com a investigação do NTSB[7], o sistema de percepção do veículo detectou a pedestre 5,6 segundos antes do impacto e a rastreou até a colisão, oscilando entre classificá-la como 'objeto desconhecido', 'veículo' e 'bicicleta' (Elish, 2019, p. 14). A motorista de segurança, flagrada pelas câmeras do carro olhando para baixo, teve sua conduta imediatamente associada à hipótese de distração, seja pelo uso do celular ou pela atenção voltada à tela de monitoramento.

O caso repercutiu amplamente na mídia e entre as autoridades, que rapidamente a designaram como a principal responsável pelo acidente. O chefe de polícia de Tempe, por exemplo, declarou que a "Uber provavelmente não seria culpada, mas que a motorista de segurança poderia enfrentar acusações criminais" (Elish, 2019, p. 14). A cobertura midiática concentrou-se na conduta da motorista, enquanto as falhas no software de detecção e a desativação do sistema de freios

[7] A National Transportation Safety Board, é uma agência federal dos Estados Unidos, criada pelo Congresso em 1967, para investigar acidentes significativos em todos os modos de transporte, incluindo aviação, rodovias e etc. Disponível em: https://www.ntsb.gov/Pages/home.aspx

automático pela Uber permaneceram em segundo plano. O veículo em questão era um Volvo XC90 equipado com sistema de freios automático, mas esse sistema havia sido desativado pela Uber para que o veículo pudesse operar com seu próprio software de direção autônoma. Se o sistema de freios do Volvo não tivesse sido desativado, ele teria acionado os freios e parado antes de atingir a pedestre. A motorista de segurança, no entanto, "pode estar enfrentando acusações criminais por homicídio culposo no trânsito" (Somerville e Shepardson 2018 apud Elish, 2019, p. 53).

Sob a perspectiva relacional, a responsabilização não se direcionaria a uma questão de distribuir culpa conforme a contribuição causal de cada agente. Trata-se, antes, de examinar como as posições ocupadas por cada um moldam não apenas a ocorrência do erro, mas também a forma como ele será interpretado, explicado e julgado. O relatório da NTSB revela que a motorista de segurança, por estar fisicamente presente no local e ter sua distração registrada pelas câmeras, foi a única a responder criminalmente, apesar dos erros cometidos pela empresa Uber. A Uber, por sua vez, por ter desativado o sistema de freios automático do Volvo e definido as condições do teste, seria questionada sobre as decisões que ampliaram o risco do acidente. Os desenvolvedores, igualmente, seriam chamados a esclarecer por que o software oscilou na classificação da pedestre. (Elish, 2019)

Cada um desses agentes, portanto, ocuparia uma posição específica na teia de relações que produziu o acidente, sendo possível ordená-los conforme sua influência sobre o erro e sua capacidade de responder àqueles que sofreram suas consequências. A pedestre e sua família ocupariam o ponto mais exposto, pois foram eles que sofreram as consequências mais graves. Em seguida, a motorista de segurança estaria em uma posição secundária de vulnerabilidade, por estar fisicamente presente no local e não conseguir controlar a tempo o carro. Acima dela, os desenvolvedores ocupariam uma

posição intermediária, uma vez que poderiam tentar esclarecer por que o software oscilou na classificação da pedestre, embora não tenham participado diretamente das decisões de colocar esse projeto nas ruas, já que são apenas funcionários da empresa. No topo da cadeia, a Uber se encontraria na posição de maior influência, primeiro pela decisão de desativar o sistema de freios automático do Volvo e segundo por definir as condições do teste, decisões que ampliaram o risco do acidente e revelaram uma cultura de segurança inadequada. A pergunta que orientaria a investigação, então, não seria 'quem' causou o quê? Mas 'o que' cada um desses agentes deve explicar à família da vítima?'

A explicabilidade, nessas situações, possibilitaria que o indivíduo ou instituição não perca de vista que, embora a ação tenha sido executada por um agente artificial, a decisão de utilizá-lo e o modo como o fez permanecem sob sua responsabilidade. Diferentemente de um modelo moral tradicional que se concentra no que o agente sabe ou deveria saber, a abordagem relacional desloca o foco para o que o paciente pode legitimamente perguntar. O ato de perguntar, aqui, não seria uma busca por informação técnica, mas uma demanda por reconhecimento e prestação de contas. Se a família da vítima, por exemplo, viesse a perguntar por que o carro autônomo não freou, talvez não estivesse buscando apenas um relatório técnico sobre o funcionamento do algoritmo, mas um reconhecimento, por parte da Uber, de que suas decisões tiveram consequências sobre uma vida. Ser transparente nesse caso é saber responder e explicar por seus atos diante daqueles que sofreram as consequências. Sem essa disposição, a explicabilidade se reduziria a um exercício técnico desprovido de significado ético.

Se um agente humano toma uma decisão com base em uma recomendação da IA e não é capaz de explicá-la, isso compromete a responsabilidade por duas razões:

primeiro, porque o agente não sabe o que está fazendo; segundo, porque falha em responder àquele que é afetado pela decisão e que tem o direito de exigir uma explicação. Em contextos em que o paciente não está em condições de fazer essa exigência, outros podem fazê-lo em seu nome, garantindo que a demanda por transparência não se perca entre quem age e quem sofre as consequências (Coeckelbergh, 2019). Essa dinâmica pode ser observada em diferentes situações cotidianas envolvendo IA. No âmbito da saúde, um médico que utiliza um sistema de diagnóstico baseado em IA para recomendar um tratamento precisa ser capaz de explicar ao paciente por que aquela decisão foi tomada, ainda que não domine os detalhes técnicos do algoritmo. Em processos de seleção de pessoal, um gestor que confia em um sistema de triagem automatizada para classificar candidatos deve poder justificar ao candidato rejeitado os critérios que levaram à sua exclusão. No setor financeiro, um analista que utiliza um modelo de pontuação de crédito precisa responder ao cliente sobre os fatores que influenciaram a negativa de um empréstimo.

Nesse contexto, estabelecer esses esclarecimentos também ajuda a definir o tipo e o grau de transparência que podem ser razoavelmente exigidos. Embora pesquisas técnicas avancem no sentido de tornar a IA mais compreensível, e medidas legais possam garantir um direito à explicação, ainda não há consenso sobre o que constitui uma boa explicação ou sobre o quanto dela é realmente necessário. A literatura sobre explicação sugere que as pessoas não esperam cadeias causais completas, mas respostas que façam sentido no contexto social e que levem em conta o que o outro precisa saber (Miller, 2019 *apud* Coeckelbergh, 2019, p. 2063). Isso significa que a explicabilidade não exige transparência absoluta, nem que as máquinas expliquem por si mesmas, mas que sejamos aptos a não apenas ser conduzidos por decisões artificiais,

mas preservar nossa capacidade de interpretar e dar sentido ao nosso próprio modo de ser, co-constituído na relação com essas tecnologias.

Em suma, a abordagem relacional como via de responsabilização sugere um caminho para um uso técnico mais ético, na medida em que se inscreve no horizonte da existência humana. Essa virada fenomenológica, ao priorizar a relação que o individuo estabelece com a tecnologia em vez de analisar a constituição técnica do objeto em si, parece resgatar a crítica de Heidegger (2006) à metafísica tradicional. Sob essa perspectiva, o filósofo alemão apontava que o pensamento ocidental tendeu a reduzir a realidade a objetos isolados e mensuráveis — os entes —, esquecendo-se de investigar a própria abertura que permite a esses objetos fazerem sentido, isto é, a pergunta pelo ser. Desse modo, o sistema tecnológico não se apresentaria como uma entidade independente ou neutra; mas surgiria, antes, em um fluxo vivo de nossas experiências, horizonte de significados, e linguagem. A compreensão dessa tecnologia, portanto, talvez não possa prescindir da análise do próprio mundo que a acolhe e que ela, por sua vez, contribui para transformar.

Se a tecnologia participa de nossas narrativas e media nossa relação com o mundo, a ausência de explicabilidade não pode ser reduzida a um problema técnico de caixa-preta. Ela afeta, mais profundamente, a capacidade humana de manter-se como sujeito interpretativo, que tem interesse em compreender a sua volta de uma forma não tão metódica. Por essa razão quando uma decisão tomada por um sistema artificial não pode ser adequadamente compreendida ou comunicada, perde-se não apenas o controle sobre o processo, mas também a possibilidade de integrar aquela decisão à própria história da experiência coletiva. A opacidade da IA, nesse quadro, constitui um problema à própria capacidade do humano de se reconhecer como autor de sua existência e do efeito de sua ação na sociedade.

Talvez, nesse ponto, a ideia de explicabilidade adquira um caráter existencial, na medida em que somos seres que refletimos sobre nossa própria existência por meio do sentido e da linguagem. Ser privado da capacidade de compreender os critérios que moldam nossas vidas, portanto, equivale a uma fragmentação da própria narrativa que nos constitui. Se a IA co-escreve essas narrativas da vida cotidiana, conforme sugere a abordagem relacional, então a capacidade de explicar suas operações seria o que permitiria que essas histórias permaneçam, em alguma medida, sob manejo humano. Não se trataria, assim, de cobrar uma IA totalmente explicável em seus mecanismos internos, mas de perguntar o que a explicabilidade significa quando pensada a partir da experiência daqueles que são prejudicados por suas operações. Essa pergunta, contudo, talvez nos conduza menos à transparência técnica e mais à responsabilidade ética.

Considerações finais

O percurso traçado ao longo deste artigo partiu de uma análise teórico-conceitual de caráter qualitativo sobre o conceito de agência em tempos de avanço tecnológico, especificamente da inteligência artificial. A questão central reside no fato de que os critérios filosóficos tradicionais, desde os gregos até a filosofia moderna, não são suficientes para pensar a agência numa sociedade em que sistemas de IA e humanos agem de forma integrada. Nesse contexto, as condições éticas para a atribuição de responsabilidade permanecem imprecisas, especialmente quando os sistemas produzem erros que não podemos entender como se deu a decisão, o que possibilita uma lacuna de responsabilidade ou *responsibility gap* (Matthias, 2004).

Esse impasse levou à investigação de abordagens centradas no debate sobre a agência moral da IA. De um lado, autores como Johnson (2006) reconhecem à IA um estatuto de entidade moral, mas recusam-lhe uma agência plena. De outro, os Andersons (2010) defendem que sistemas autônomos podem e devem ser programados para atuar eticamente. A meu ver, porém, esse debate talvez já tenha sido superado, na medida em que as IAs já atuam de forma constante e produzem efeitos no mundo antes mesmo de qualquer definição conclusiva sobre sua agência ou consideração moral. Ainda que se possa reconhecê-la como agente no sentido operacional, ela segue sendo arresponsável, ou seja, não pode ser responsabilizada nos mesmos termos que um agente humano (Coeckelbergh, 2023). A segunda abordagem avança para uma discussão sobre que tipo de responsabilização desenvolver a partir desse movimento de múltiplas agências humanas e não humanas. Floridi (2016) propõe, nesse contexto, o conceito de responsabilidade distribuída, que busca incluir as IAs na cadeia de responsabilização por possíveis erros. A proposta tem o mérito de reconhecer que as ações mediadas por IA envolvem uma rede de agentes cujas contribuições se entrelaçam de maneira complexa, o que torna inadequada a atribuição de responsabilidade a um único agente. No entanto, a responsabilidade distribuída parece não oferecer critérios claros para atribuir responsabilidades de forma justa, especialmente quando os agentes humanos ocupam posições sociais e institucionais muito diferentes.

A partir de Gunkel (2012), foi possível observar o viés excludente que atravessa a própria estrutura do conceito de agência moral. O autor mostra que os critérios que definem quem conta como agente foram historicamente moldados por grupos hegemônicos, o que torna necessário questionar essa posição antropocêntrica. Para avançar nessa direção, propõe -se o conceito de paciência moral, ou seja, para aquele

que sofre os efeitos da ação, abrindo assim a possibilidade de estender a consideração moral a entidades como as IAs. Coeckelbergh (2019), por sua vez, desenvolve uma abordagem sócio-relacional, segundo a qual o status moral não é algo que se descobre ou se define por critérios fixos, mas algo que se atribui nas relações que estabelecemos com as IAs.

Entre as abordagens examinadas, alinho-me à perspectiva relacional de Coeckelbergh, pois a responsabilidade não se reduz a agir com consciência ou a ter boas intenções, mas envolve, antes, a capacidade de resposta perante o outro. Trata-se, mais precisamente, da disposição para explicar e justificar decisões para aqueles que são afetados por elas, o que recoloca a pergunta sobre responsabilidade não como uma questão de atribuição de culpa, mas como um exercício de prestação de contas àqueles que ocupam a posição de vulnerabilidade diante de erros realizados por sistemas autônomos.

Coeckelbergh mostrou, assim, a importância tanto da explicabilidade quanto da via interpretativa para o desenvolvimento de uma relação ética com a IA. Uma questão, porém, permanece aberta, pois a própria multiplicação de IAs e a crescente delegação de decisões podem tornar essa relação estruturalmente difícil de ser mantida. Mesmo que estejamos dispostos a responder perante o outro, o avanço da IA sobre essa lacuna ética talvez torne essa disposição insuficiente, na medida em que a delegação aumenta, mas a capacidade de prestar contas parece não acompanhar esse crescimento. O problema da responsabilidade, portanto, segue em aberto, e talvez seja esse o ponto a partir das quais investigações futuras possam partir, perguntando não apenas como responder, mas como manter abertas as condições para que a resposta seja possível.

Referências

ANDERSON, Michael; ANDERSON, Susan Leigh. *Robot be good*. Scientific American, New York, v. 303, n. 4, p. 72-77, 2010.

ANDREAZZA, T. M. *Aristóteles e o problema da responsabilidade moral: predeterminismo ou indeterminismo?* Pensando – Revista de Filosofia, [S. l.], v. 5, n. 10, p. 149-169, 2014.

ARISTÓTELES. *Ética a Nicômaco*. In: ARISTÓTELES. The complete works of Aristotle. Editado por J. Barnes. Princeton: Princeton University Press, 1984.

BUTLER, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1990.

COECKELBERGH, Mark. *Who cares about robots? A phenomenological approach to the moral status of autonomous intelligent machines*. In: The society for the study of artificial intelligence and simulation of behaviour, Birmingham, UK, 2-6 July 2012. Anais [...]. Birmingham: [s. n.], 2012.

_____. *Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability*. Science and Engineering Ethics, [S. l.], v. 26, p. 2051-2068, 2019. DOI: 10.1007/s11948-020-00242-0.

_____. *Ética na Inteligência Artificial*. Rio de Janeiro: Ed. PUC-Rio, 2023.

_____. *Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sence-making*. AI & Society,[S. I], v.38, n.6, p. 2437-2450, 2023. DOI: 10.1007/s00146-021-01375-x

CONSTANTINESCU, Mihaela; VOINEA, Cristina; USZKAI, Radu; VICĂ, Constantin. *Understanding responsibility in Responsible AI*. Dianoetic virtues and the hard problem of context. Ethics and Information Technology, [S. l.], v. 23, n. 4, p. 803-814, 2021. DOI: https://doi.org/10.1007/s10676-021-09616-9.

DAWKINS, Marian Stamp. *Em busca da consciência animal*. 2005. Disponível em: https://www.cahiers-antispecistes.org/pt-pt/marian-stamp-dawkins-em-busca-da-conscienciaanimal/. Acesso em: 20 abr. 2026.

ELISH, Madeleine Clare. *Moral crumple zones: Cautionary tales in human-robot interaction*. Engaging Science, Technology, and Society, [S. l.], v. 5, p. 40-60, 2019.

FLORIDI, L. *Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions*. Philosophical Transactions of the Royal Society A, London, v. 374, n. 2083, 2016.

GUNKEL, David J. *A Vindication of the Rights of Machines*. In: The society for the study of artificial intelligence and simulation of behaviour. Birmingham, UK, 2-6 July 2012. Anais [...]. Birmingham: [s. n.], 2012.

HEIDEGGER, Martin. *A Questão da Técnica*. In: *Ensaios e Conferências*. Petrópolis: Vozes, 2006

JOHNSON, Deborah G. *Computer Systems: Moral Entities but not Moral Agents*. Ethics and Information Technology, [S. l.], v. 8, n. 4, p. 195-204, 2006. DOI: 10.1007/s10676-006-9111-5.

MATTHIAS, Andreas. *The responsibility gap: Ascribing responsibility for the actions of learning automata*. Ethics and Information Technology, [S. l.], v. 6, n. 3, p. 175-183, 2004.

NASCIMENTO, Francisco Eliandro; SANTOS, Francisco Rogelio dos. *A estrutura da moral kantiana* Griot: Revista de Filosofia, vol. 17, núm. 1, 2018, pp. 61-84 DOI: https://doi.org/10.31977/grirfi.v17i1.808 Acesso em: 20 abr. 2026.

PASQUALE, Frank. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press, 2015.

RUSSELL, Bertrand. *História da filosofia ocidental*. Tradução de Hugo Langone. Rio de Janeiro: Nova Fronteira, 2015.

TAVARES, M.; FERRO, M. *Análise da obra fundamentação da metafísica dos costumes de Kant*. 2. ed. Lisboa: Editorial Presença, 1997.

TADDEO, M; FLORIDI, L. *How AI can be a force for good*. Science, Washington, D.C., v. 361, n 6404, p. 751-752, 2018. DOI: 10.1126/science.aat5991