Luís Gabriel Provinciatto*
DOI: 10.32334/oqnfp.2023n53a982

Ética na Inteligência Artificial

COECKELBERGH, Mark. *Ética na inteligência artificial*.
Traduzido por Clarisse de Souza et al. São Paulo: Ubu; Rio de Janeiro: Editora PUC-Rio, 2023. 192pp.

Não restam dúvidas de que um dos temas que atualmente mais suscita discussão – em vários níveis, por inúmeras áreas e com diferentes propósitos – é a inteligência artificial. De fato, ela existe há mais de meio século, mas foi potencializada e amplamente difundida por meio da inteligência artificial generativa, como o caso do ChatGPT, o chatbot desenvolvido pela OpenIA e lançado em 2022, que atuou na redação do prefácio à edição brasileira de *Ética na inteligência artificial*, de Mark Coeckelbergh.

Lançada originalmente em inglês, em 2020, a recém-traduzida obra de Coeckelbergh traz ao público de língua portuguesa uma abordagem introdutória acerca da possibilidade de se ler a inteligência artificial (IA) a partir da perspectiva filosófica. A obra destaca e aborda mais precisamente a relação da IA com questões de ordem ética, moral e política. O que o/a leitor/a encontrará ao longo da obra, entretanto, não é um manual de condutas ético-políticas aceitáveis sobre o uso adequado da IA. Na verdade, trata-se muito mais de trazer à tona um conjunto de perguntas e problemas relativos ou decorrentes da IA que reivindicam uma abordagem filosófica do que propor uma normatização ética da IA ou algo do tipo. Em outras palavras, o principal objetivo de Coeckelbergh é evidenciar como a IA não se restringe à tecnologia da informação e suas áreas correlatas, mas também se entrelaça com a filosofia e sua tradição.

* Pontifícia Universidade Católica de Campinas (PUC-Campinas).
Contato: luis.provinciatto@puc-campinas.edu.br

Recebido em: 09/06/2024 Aceito em: 15/07/2024

Para levar a cabo tal propósito, Coeckelbergh organiza a obra em doze capítulos, que podem ser assim articulados em três principais conjuntos. O primeiro conjunto, que engloba os quatro primeiros capítulos, pauta a discussão em aspectos epistemológicos e antropológicos, com vistas a traçar um cenário teórico conceitual para abordar as questões práticas relacionadas à IA. O segundo conjunto, composto pelos capítulos cinco e seis, define brevemente o que é IA e discute a importância da ciência de dados no processo de aprendizado de máquina, apontando para os problemas éticos que estão aí implicados. Por fim, o terceiro conjunto, composto pelos seis capítulos restantes, aborda diferentes desdobramentos e impactos éticos e políticos da IA na vida cotidiana, tais como a privacidade de dados, a atribuição de responsabilidade moral às máquinas, as propostas de políticas em torno à IA no âmbito público e no setor privado, bem como a possibilidade de articular as discussões sobre IA com outro tema de grande magnitude, a saber, a mudança climática.

À luz desse enquadramento, vejamos brevemente os principais pontos discutidos em cada um dos capítulos, antecipando que todos eles são ricos em exemplos, o que não os torna bastante fluídos e acessíveis a diferentes públicos.

O primeiro capítulo se inicia com uma exposição do conflito existente entre ser humano e IA, destacando que uma das capacidades desta é a de aprender por si mesma, o que leva ao primeiro grande problema: será possível pensar a IA a partir do paradigma do instrumento/ferramenta, visto como algo que complementa a tarefa e é dominado pelo ser humano? A resposta é clara: "entramos em uma Segunda Era da Máquina, em que as máquinas não apenas complementam os seres humanos, como na Revolução Industrial, mas também os substituem" (Coeckelbergh, 2023, p. 18). Torna-se necessário pensar a IA a partir de outro paradigma, um que abarque não apenas um "caráter antropológico" da máquina, mas também um "âmbito ético", pois agora a máquina não apenas substitui o ser humano, como toma decisões em seu lugar: "quantas decisões e *quanto* dessas decisões queremos delegar à IA? E quem é responsável quando algo dá errado?" (Coeckelbergh, 2023, p. 15). Esses dois problemas iniciais se apresentam como os balizadores das discussões que se seguem nos demais capítulos.

O segundo capítulo aborda duas concepções estreitamente ligadas à IA: a de superinteligência, que propõe uma superação da inteligência humana pela máquina ou pela IA, e a de transumanismo, que se refere ao aprimoramento/aperfeiçoamento da espécie humana por meio da tecnologia. Ao conceito de superinteligência se associa o de *singularidade tecnológica*, "um momento na história humana no qual o progresso tecnológico exponencial poderia trazer

uma mudança tão dramática que nós não compreenderíamos o que estaria acontecendo" (Coeckelbergh, 2023, p. 22). À luz desse conceito, pode-se pensar tanto em um *transumano* quanto em uma *inteligência artificial geral* que seja igual ou superior à humana. O problema de fundo aqui reside na presença de uma premissa humanista que, por um lado, assume a inteligência humana como parâmetro de definição do conhecimento e, por outro, dá origem a uma narrativa de competição entre ser humano e máquina, tal como narrado no romance *Frankenstein* de Mary Shelley.

Como ir além dessas narrativas de competição e de hipervalorização, seja do humano, seja da máquina? Uma possibilidade é debruçar-se sobre a IA e suas premissas, conhecendo-a criticamente. Isso começa a ser feito no terceiro capítulo, no qual Coeckelbergh faz um rápido mapeamento da filosofia no século XX para mostrar que, ao discutir a possibilidade de uma IA geral – que, de fato, ainda não é um problema relevante para a ética –, se descobrem diferentes posições acerca da natureza do ser humano. Destacam-se três tensões: a decorrente dos séculos XVIII e XIX entre Iluminismo e Romantismo, ou seja, entre o otimismo no que a ciência poderia fazer pela humanidade e o desencantamento em torno ao mistério da vida provocado por tal desenvolvimento; aquela entre humanismo e transumanismo, que discute o que é o ser humano e o que ele deve se tornar; por fim, o debate entre humanistas e pós-humanistas, pautado no lugar ocupado pelo ser humano nas ontologias e éticas modernas. O capítulo se encerra oferecendo outra possível chave de leitura para superar essas narrativas de competição: a abordagem pós-fenomenológica, que entende que a tecnologia é parte da existência humana, mediando nossa relação com o mundo.

Nesse cenário, o quarto capítulo de apresenta como central para toda a obra, pois nele se levanta a questão sobre o status moral da IA, pondo o seguinte problema: uma IA pode ter uma plena capacidade de agir[1] moralmente? Para construir uma resposta, precisa-se compreender, por um lado, o que constitui propriamente a moralidade e, por outro, o que é fator decisório no momento mesmo da decisão. Coeckelbergh afirma que, embora não possa ser reduzida à emoção, a decisão também não é apenas algo perfeitamente racional: "se uma IA geral é realmente possível, então não havemos de querer um tipo de 'IA psicopata' que seja perfeitamente racional, porém insensível às preocupações humanas porque é desprovida de emoções" (Coeckelbergh,

1 Os tradutores optam por traduzir o termo inglês *agency* por agência, que aqui significa essa *capacidade de agir.*

2023, p. 53). A isso, no entanto, segue-se outro problema: por que se deveria imitar a moralidade humana? Isso expõe novamente o fundo humanista da discussão e, ao mesmo tempo, suscita o debate acerca de uma moralidade não baseada em princípios humanos. Seriam estes que tornariam a máquina moralmente responsável por suas decisões? Entretanto, como uma máquina pode *aprender a escolher*? É isso o que a caracteriza como "inteligente"?

Como se percebe, antes de avançar para as discussões dos desdobramentos éticos da IA, faz-se necessário um breve entendimento sobre o que é o aprendizado de máquinas, um dos elementos centrais e mais característicos da IA. Antes de chegar ao aprendizado de máquinas, tema do sexto capítulo, Coeckelbergh caracterizada brevemente a IA no quinto capítulo como *forte* e *fraca*. A IA forte é aquela capaz de levar a cabo qualquer tarefa cognitiva também desempenhada por um ser humano, enquanto a IA fraca é capaz de operar em domínios específicos, como seleção e classificação de imagens. Além disso, o autor mostra as diferentes possibilidades paradigmáticas de abordar a IA, destacando os paradigmas simbólico, conexionista e corporificado. Essa caracterização geral traça o panorama de aplicação da IA, possibilitando a análise dos primeiros impactos sociais do avanço tecnológico.

A descrição sobre o aprendizado de máquinas também é realizada tendo em vista retornar ao problema posto no quarto capítulo. Por essa razão, o autor aqui também privilegia a caracterização geral, apresentando os três diferentes tipos de aprendizado de máquinas – supervisionado, não supervisionado e por reforço – e a importância da coleta, limpeza, cruzamento, seleção e formas de utilização de dados. Aqui se nota a inegável importância da IA: a quantidade de dados processados, a celeridade e o cruzamento de dados possíveis são algo nunca antes imaginável. Ao mesmo tempo, nota-se que, mesmo com uma quantidade massiva de dados, falta à IA o "entendimento da relevância", "compreensão, experiência, sensibilidade e sabedoria" (Coeckelbergh, 2023, p. 86). Além disso, falta-lhe um dado empírico bastante evidente: "sem programadores e cientistas de dados, a tecnologia simplesmente não funciona" (Coeckelbergh, 2023, p. 86).

À luz desse panorama, Coeckelbergh começa a abordar alguns dos principais desdobramentos éticos da IA, dedicando o sétimo capítulo à questão da privacidade e proteção de dados. A discussão se inicia abordando o modo como esses dados são coletados, trazendo o exemplo de aplicativos que exigem o consentimento dos usuários, mas nem sempre fornecem informações claras sobre o destino de seus dados após a coleta. Isso traz à tona outro ponto: a vulnerabilidade dos usuários, que, por um lado, estão expostos a

diversos riscos e, por outro, servem como "'mão de obra' digital gratuita" (Coeckelbergh, 2023, p. 95), pois fornecem dados para os bancos de dados, correndo "o risco de se tornarem a força de trabalho explorada e não remunerada que produz dados para a IA" (Coeckelbergh, 2023, p. 95).

O problema exposto no oitavo capítulo é o da possível atribuição de responsabilidade moral às "máquinas inteligentes". Recorrendo à definição aristotélica de responsabilidade, Coeckelbergh afirma que as máquinas, em geral, podem ser *agentes*, mas não *agentes morais*. Em outras palavras, elas são aptas para agir, mas não para serem responsáveis pela ação, pois todas as ações tomadas por elas são, no fundo, programadas por alguém que exerce certo controle sobre elas. No entanto, será que isso também se aplica à IA? Será que o próprio sistema de funcionamento das IAs é transparente o suficiente para ser explicável? A questão da transparência pode ser facilmente resolvida em alguns casos, como quando um chatbot, por exemplo, utiliza um mecanismo de árvore genealógica de resposta. Porém, no caso de redes neurais, explicar uma determinada decisão se torna impossível. Não há explicação possível, nem transparência. O máximo que se pode dizer é sobre a estrutura de funcionamento, mas não sobre a decisão em si. Esse é o "problema da *caixa preta*" (Coeckelbergh, 2023, p. 110), que leva à seguinte questão: "se o custo de um sistema com um desempenho é a falta de transparência, devemos ainda assim usá-lo ou não?" (Coeckelbergh, 2023, p. 113).

Ademais, como são constituídos esses bancos de dados? Tal pergunta expõe o problema do viés nos bancos de dados, tema do qual Coeckelbergh se ocupa no nono capítulo. Após concluir que não é possível construir um banco de dados neutro, o autor se propõe a mostrar como minimizar o impacto de determinados vieses, como os de raça, cor, classe social, gênero, etc. Discute-se também o fato de que qualquer algoritmo de seleção precisa ser "discriminatório": "a questão ética e política é se determinada discriminação é injusta e desleal" (Coeckelbergh, 2023, p. 123). Por isso, o próprio desenvolvimento da IA reivindica a necessidade de políticas, não apenas por causa do impacto causado nos campos de trabalho e na educação, por exemplo, mas também porque a coleta, limpeza, seleção e utilização de dados precisam ser reguladas e transparentes aos usuários leigos. O capítulo se encerra abordando a disparidade na "velocidade de transformação" entre o impacto das tecnologias – da IA, sobretudo – e as propostas de discussão éticas e políticas a esse respeito.

O décimo capítulo dá continuidade a esse ponto, pois aí Coeckelbergh esboça as linhas gerais de orientação para uma possível forma de intervenção prático-política a respeito da regulação da IA. A questão mais óbvia

é *o que deve/pode ser feito*, mas a esta se seguem algumas outras: "*por que* deve ser feito, *quando* deve ser feito, *quanto* deve ser feito, *por quem* deve ser feito e qual é a *natureza, extensão* e *urgência do problema*" (Coeckelbergh, 2023, p. 136). A partir disso, traçam-se cinco linhas de ações gerais: a) justificar as medidas propostas; b) tentar implementar políticas antes de a tecnologia estar totalmente desenvolvida; c) identificar a talvez não necessidade de novas medidas, utilizando as já existentes de forma eficiente; d) ter clareza a respeito de quem deve agir; e) ter clareza sobre o que e quanto deve ser feito. O conjunto de exemplos trazidos por Coeckelbergh demonstra certos pontos de convergência entre algumas políticas já adotadas no setor público e privado. Por fim, o autor também aponta para o fato de tais políticas deverem combinar aspectos técnicos, como fatores de proteção, com aspectos não-técnicos, ou seja, características que não são condizentes ao desenvolvimento técnico/tecnológico da IA. Isso permite levantar questões como a da explicabilidade e transparência, que, por sua vez, reivindicam a discussão filosófica de fundo.

O tema do desenvolvimento de políticas é concluído no capítulo onze, dedicado à discussão dos principais desafios presentes em sua formulação. Mesmo indicando que a principal proposta ética na IA consiste em adotá-la desde a concepção do projeto, Coeckelbergh reconhece que a maior questão está em tentar prever ou determinar quais problemas surgirão, já que eles ainda não existem de fato. Nesse cenário, uma possibilidade é dar voz e ouvir os diversos agentes e as muitas partes envolvidas em tais projetos, desde os desenvolvedores técnicos até os usuários. Com essa proposta, Coeckelbergh propõe justamente uma "inovação responsável construída mais de baixo para cima" (Coeckelbergh, 2023, p. 157), que contrasta com a maioria das situações, nas quais as políticas e planos éticos são desenvolvidos de forma abstrata apenas por especialistas.

Se todas as partes implicadas nos projetos tecnológicos devem ser consideradas, surge uma espécie de imperativo ético-político: "se defendermos o ideal da democracia e se tal conceito envolve inclusão e participação na tomada de decisão sobre o futuro de nossas sociedades, a escuta das vozes das partes interessadas não é uma opção, mas um requisito ético e político" (Coeckelbergh, 2023, p. 158). A isso se segue o (quase) evidente problema da concentração de poder pelas grandes empresas de processamento e armazenamento de dados, somada à diminuição da relevância dos interesses públicos sobre o assunto, o que levaria à seguinte conclusão: tal pauta existe, mas é pouco ou nada eficiente. Por essa razão, Coeckelbergh defende a necessidade de haver espaço para

se colocar tais questões, não fazendo delas meros acessórios para justificar o desenvolvimento da tecnologia a qualquer custo. Percebe-se aqui novamente a relevância da filosofia para o debate acerca da tecnologia.

O último capítulo é marcado por duas questões centrais: a primeira, se a ética na IA deve ser centrada no ser humano; a segunda, sobre a prioridade de problemas éticos, questionando se não há temas mais urgentes, como a mudança climática, a falta de água potável e as guerras, por exemplo. Nesse cenário, Coeckelbergh aponta para o fato de a IA trazer novos problemas e, além disso, poder agravar outros já existentes, de modo que as questões relativas à IA podem ser articuladas com outras relativas a outros temas. O exemplo melhor descrito pelo autor é o da relação entre a IA e a mudança climática, destacando os potenciais da IA para auxiliar na identificação de formas de mitigação dos impactos da transformação do meio ambiente na vida humana e não-humana. O problema decorrente é entender e assumir a IA como a *principal* solução, submetendo tudo a ela, o que reforçaria o paradigma tecnocrático: "um perigo da IA, portanto, é que ela permite esse tipo de pensamento e se torna uma máquina de alienação" (Coeckelbergh, 2023, p. 178).

Ao concluir a leitura de *Ética na inteligência artificial*, o/a leitor/a percebe que percorreu uma excelente introdução temática à leitura filosófica dos problemas decorrentes do desenvolvimento tecnológico, em geral, e da IA, em específico. Ao mesmo tempo, percebe que se encontra muito mais com um conjunto de perguntas-problema de difícil resolução do que com respostas a tais problemas. Isso, longe de ser uma falha, demonstra o seu total pertencimento à tradição filosófica, que ainda resiste à tentação de oferecer respostas técnicas para os complexos problemas da época técnica do mundo. A relevância da obra para uma discussão interdisciplinas sobre o tema também é evidente. *Ética na inteligência artificial* certamente figurará não apenas nos currículos de filosofia, mas também em áreas como tecnologia da informação, que se beneficiam do diálogo com a filosofia. Em suma, a obra de Mark Coeckelbergh se destaca por sua capacidade de instigar o debate e estimular a reflexão crítica sobre os impactos da IA na sociedade, abrindo caminho para a construção de um futuro mais justo e responsável.