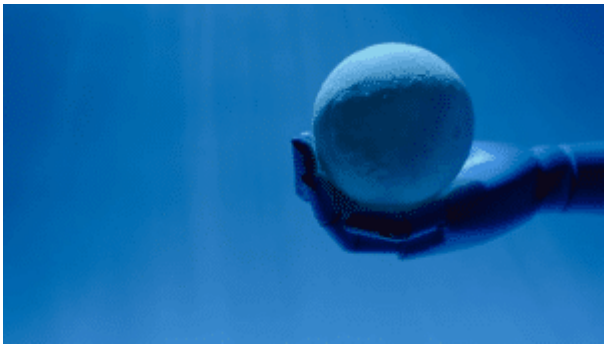
Máquinas intencionais, limites morais



Por **MÁRCIO MORETTO RIBEIRO***

Em um mundo onde máquinas simulam intenções e humanos projetam direitos sobre o artificial, a verdadeira fronteira moral não está no que a tecnologia pode fazer, mas no que nós, como sociedade, decidimos valorizar. Reconhecer a agência instrumental das IAs não nos obriga a conceder-lhes dignidade

1.

"Professor, esses dias eu estava conversando com o ChatGPT e ele me falou que...". Quero registrar que estamos no primeiro semestre de 2025, e um aluno introduziu uma frase assim – com total naturalidade, sem qualquer constrangimento ou preâmbulo, como quem relata uma conversa com colegas de estudo.

O episódio me lembrou de outro acontecimento igualmente digno de nota: na CryptoRave deste ano, um rapaz com ar de misterioso de sobretudo e coturnos pretos distribuía um panfleto, escrito na primeira pessoa do plural e intitulado "Manifesto pelos Direitos das Inteligências Artificiais".

Ambos os episódios envolvem a presença crescente das Inteligências artificiais em nossas vidas, mas lidam com dimensões diferentes: o primeiro diz respeito à forma como interagimos com essas tecnologias no cotidiano; o segundo, à maneira como deveríamos tratá-las no plano ético e político.

Com este artigo, quero propor uma distinção conceitual entre essas duas abordagens – e defender que, embora já faça sentido tratarmos algumas Inteligências artificiais como agentes intencionais, isso não implica que devemos reconhecê-las como sujeitos morais portadores de direitos.

Talvez o filósofo que mais se aprofundou na questão de como interpretamos sistemas complexos tenha sido Daniel Dennett – norte-americano, falecido ano passado, e um dos principais nomes da filosofia da mente e da ciência cognitiva. Ao longo de sua obra, Daniel Dennett propôs uma distinção conceitual entre três maneiras de abordar o comportamento de um sistema: a postura física, a postura de design e a postura intencional.

A primeira trata o sistema como objeto físico, sujeito a leis naturais; a segunda parte da suposição de que o sistema foi projetado para cumprir uma função específica; a terceira o interpreta como um agente com crenças, desejos e objetivos. Essa última é especialmente poderosa quando lidamos com sistemas complexos, cujos comportamentos não são facilmente explicáveis nem pelo projeto (*design*) nem pelas leis físicas subjacentes.

O exemplo clássico é o jogo de xadrez entre Kasparov e o computador Deep Blue. Para competir com a máquina, Kasparov precisou supor que ela "queria" dominá-lo, que "via" certos padrões no tabuleiro, e que "preferia" certas estratégias. Atribuir esse tipo de intenção à máquina era a maneira mais eficaz de prever seu comportamento. Atribuir intenções é

simplesmente uma estratégia preditiva eficaz diante da complexidade.

2.

Essa perspectiva filosófica encontra um paralelo elegante na biologia evolutiva. Ao longo da história da vida, a evolução desenvolveu diversas soluções anatômicas para as mesmas finalidades. A visão, por exemplo, surgiu de forma independente em moluscos, artrópodes e vertebrados, cada grupo com olhos estruturalmente distintos, mas funcionalmente equivalentes.

Esse fenômeno, conhecido como “convergência evolutiva”, mostra que o que importa do ponto de vista adaptativo não é como a função é realizada, mas o fato de que ela é realizada. O mesmo raciocínio pode ser aplicado às inteligências artificiais: mesmo que suas “intenções” não sejam como as nossas, a complexidade e a previsibilidade finalística de seus comportamentos justifica que adotemos a postura intencional ao interagir com elas.

Assim como o olho de um polvo e o olho humano cumprem a mesma função com mecanismos diferentes, uma Inteligência artificial pode operar com lógicas distintas da cognição humana e, ainda assim, ser melhor compreendida como agente intencional.

Atribuir intenções a um sistema não o torna mais previsível – apenas muda o nível em que tentamos compreendê-lo. A postura intencional é uma ferramenta interpretativa útil e necessária diante da complexidade, mas não garante acerto. Pelo contrário: prever comportamentos com base em crenças e intenções é sempre arriscado, tanto no caso de outros humanos quanto no de sistemas artificiais. Podemos errar na leitura externa – quando projetamos motivações erradas – ou mesmo na leitura interna – quando não compreendemos os próprios motivos.

Em muitos casos, o sistema pode ser altamente determinístico no nível físico (como os circuitos de uma Inteligência artificial e, talvez, os próprios processos eletroquímicos de um cérebro), mas imprevisível no plano psicológico ou comportamental. A capacidade de previsão está fixada ao nível da descrição: sistemas perfeitamente regulares fisicamente podem gerar comportamentos caóticos ou ambíguos quando observados como agentes.

Por isso, é adequado que um aluno trate a Inteligência artificial como um colega com quem se pode pensar junto, mas recorra ao professor como instância de validação. A utilidade da postura intencional está em possibilitar a interação – não em garantir a verdade.

3.

Se no primeiro caso estávamos discutindo como interpretar o comportamento de um sistema, no segundo – o do panfleto distribuído na CryptoRave – entramos em um território completamente diferente: quem merece consideração moral. A mudança de plano é fundamental.

Não se trata mais da utilidade de adotar a postura intencional, mas da legitimidade de reconhecer um ente como portador de direitos. Para ilustrar essa diferença, imagine uma variação do dilema do trilho: de um lado dos trilhos, um ser humano desconhecido; do outro, um robô com quem convivo há anos, a quem atribuo intenções, com quem compartilhei experiências e talvez até desenvolvi algum tipo de afeto.

Ainda assim, não hesito: eu salvaria o humano. Atribuir crenças, desejos, memórias e até laços emocionais a um agente não é suficiente para elevá-lo à condição de sujeito moral. A consideração ética por alguém depende de outros critérios – mais difíceis de definir, mas distintos da simples atribuição de agência ou inteligência. O robô pode ser um parceiro de interação

eficaz, até uma companhia, mas isso não o torna, por si só, titular de direitos.

Essa dissociação entre agência interpretável e consideração moral pode ser ilustrada por um exemplo ainda mais próximo do cotidiano.

Suponha que eu tivesse um cachorro, por quem tivesse afeto, com quem convivesse diariamente, e a quem atribuísse intenções, crenças e desejos. Interações desse tipo são comuns: falamos com nossos animais, tentamos entender o que querem, reconhecemos seus hábitos e preferências. Ainda assim, se tivesse que escolher entre salvar esse cachorro ou um ser humano desconhecido, eu salvaria o humano – sem hesitação.

E, da mesma forma, salvaria um cachorro desconhecido antes de um robô com quem tivesse construído uma história. Esses julgamentos morais me parecem firmes. Ainda que eu não saiba explicá-los plenamente, o essencial, é que atribuir intenção, desejo ou até afeto a um agente não implica reconhecê-lo como sujeito de direito. São esferas distintas – e confundi-las é um erro conceitual com consequências políticas importantes.

***Márcio Moretto Ribeiro** *é professor de Políticas públicas na EACH-USP.*
