QUAL É O PAPEL DA ÉTICA NA ÉTICA DA INTELIGÊNCIA ARTIFICIAL?

WHAT IS THE ROLE OF ETHICS IN THE ETHICS OF ARTIFICIAL INTELLIGENCE?

Monica Franco[1]

Resumo: O estudo filosófico da ética sobre as interações entre os seres humanos e as tecnologias que empregam a Inteligência Artificial se tornou um tópico popular e importante como nunca antes. O estudo dos problemas que concernem essas interações justifica-se por conta das transformações tecnológicas pelas quais as sociedades humanas têm passado, e que ainda podem acarretar profundas mudanças legais, políticas e sociais. O artigo tem como objetivo principal identificar o papel da ética, enquanto disciplina filosófica, com relação ao recente e promissor domínio de investigação sobre a ética da Inteligência Artificial. Após apresentar a posição do filósofo Peter Railton na ética aplicada à Inteligência Artificial, o artigo reconstitui as suas explicações na ética normativa e na metaética. Ao ilustrar as conexões entre a ética aplicada à Inteligência Artificial e a teoria moral, o artigo argumenta que o papel da ética na ética da Inteligência Artificial pode ser explicitado com o auxílio da engenharia conceitual. Essa abordagem mostra que conceitos como 'agente' e 'interesse', que são cruciais para a ética, como pode ser visto nas explicações de Railton, não devem ser pensados exclusivamente pela ética, ou mesmo somente pela filosofia, mas necessitam ser informados pelas ciências naturais e humanas. O artigo enfatiza a necessidade de realizar a tarefa filosófica de pensar a ética para além da ética, em um contexto interdisciplinar capaz de lidar com os complexos problemas práticos que impactam as sociedades humanas.

Palavras-chave: Ética da Inteligência Artificial. Teoria Moral. Engenharia Conceitual.

Abstract: The philosophical study of ethics concerning the interactions between human beings and technologies that deploy Artificial Intelligence has become a popular and important topic as never before. The study of problems concerning such interactions is justified because of technological transformations that human societies have gone through, and that may still lead to profound legal, political, and social changes. The main objective of this paper is to identify the role of ethics, as a philosophical discipline, concerning the recent and promising field of research on the ethics of Artificial Intelligence. After presenting the position of philosopher Peter Railton in applied ethics to Artificial Intelligence, the paper reconstructs his explanations in normative ethics and metaethics. By illustrating the connections between applied ethics to Artificial Intelligence and moral theory, the paper argues that the role of ethics in the ethics of Artificial Intelligence can be made explicit with the aid of conceptual engineering. This approach shows that concepts such as 'agent' and 'interest,' which are crucial for ethics, as can be seen in Railton's explanations, should not be thought exclusively by ethics, or even only by philosophy, but need to be informed by natural and human sciences. The paper emphasizes the need to accomplish the philosophical task of thinking ethics beyond ethics in an interdisciplinary context capable of dealing with the complex practical problems that impact human societies.

Keywords: Ethics of Artificial Intelligence. Moral Theory. Conceptual Engineering.

[1] Doutoranda pelo Programa de Pós-Graduação em Filosofia da Universidade Federal de Santa Catarina (PPGFIL-UFSC). Bolsista CAPES. E-mail: monica.franco.fm@gmail.com. ORCID: https://orcid.org/0000-0002-5705-8444.

Introdução

O filósofo Peter Railton defende que a aprendizagem humana fornece o modelo de aprendizagem ética para a elaboração de uma ética das máquinas (*machine ethics*), ou seja, para a aprendizagem ética artificial. Ele argumenta que mesmo em crianças, é possível encontrar um grau considerável de autonomia na aprendizagem causal do mundo e da mente das outras pessoas. A autonomia se manifesta especialmente na aprendizagem ética, pois as crianças costumam resistir às ordens de adultos que sejam percebidas como "desnecessariamente danosa[s] ou injusta[s]" (RAILTON, 2020, p. 56, tradução nossa). Se devemos conceder autonomia para que as máquinas aprendam a se comportar moralmente, como Railton defende, várias perguntas necessitam ser respondidas. A primeira delas é se as máquinas são capazes de possuir algum tipo de agência. Ou seja, se estamos delegando decisões autônomas às máquinas sobre assuntos que são importantes para os seres humanos, e decisões que têm implicações morais, estamos diante de um novo tipo de agência, aquela própria de *agentes artificiais*? Por conta disso, a primeira parte deste artigo apresenta a posição de Railton na ética aplicada à Inteligência Artificial, indicando a sua resposta para essa pergunta.

É importante salientar que a ética pode ser caracterizada como a disciplina filosófica que investiga a moralidade em três níveis diferentes, e que a ética aplicada constitui apenas um desses níveis. Os outros níveis compreendem o que costuma ser chamado de teoria moral. As teorias filosóficas oferecidas no âmbito da *ética normativa* se ocupam dos critérios para avaliar se uma ação é correta ou errada. Essas teorias costumam ser testadas na ética aplicada para investigar as soluções que elas podem prover para problemas filosóficos práticos, como aqueles a respeito das interações entre os seres humanos e as tecnologias que empregam a Inteligência Artificial. As teorias filosóficas oferecidas no âmbito da *metaética* investigam problemas filosóficos abstratos sobre a própria natureza da moralidade ou, de modo mais amplo, da normatividade. A metaética pode ser definida como uma investigação de segunda ordem, pois ela se sobrepõe à investigação de primeira ordem que é realizada pela ética normativa. Isso porque em vez de procurar estabelecer os critérios normativos que permitem justificar se uma ação é moralmente boa e correta, a metaética se detém sobre a necessidade de justificar o que, em primeiro lugar, são o moralmente bom e o correto. Desse modo, as teorias metaéticas almejam explicar aspectos fundamentais como a existência da moralidade, a possibilidade de obtenção de conhecimento moral, a linguagem moral, bem como a dimensão prática

da moralidade e sua conexão com a motivação moral.

Embora as relações entre a ética aplicada, a ética normativa e a metaética sejam assimétricas, e as implicações de um domínio para os outros não sejam óbvias, é importante salientar que todos os domínios da ética podem ser mobilizados em conjunto para pensar sobre a Inteligência Artificial. O objetivo principal deste artigo é identificar o papel da ética, enquanto disciplina filosófica, com relação ao domínio de investigação da ética da Inteligência Artificial. Por isso, na segunda parte deste artigo, serão ilustradas as conexões entre a ética aplicada à Inteligência Artificial e a teoria moral a partir das explicações oferecidas por Railton na ética normativa e na metaética.

No nível da teoria moral, há ainda a necessidade de pensar o status moral dos agentes artificiais. Que lugar eles ocupam na comunidade moral? Devemos, por exemplo, adotar uma concepção teleológica, como o consequencialismo, e dizer que é importante garantir a satisfação dos interesses de agentes artificiais? E como especificar quais seriam esses interesses? Ou devemos, em vez disso, aderir a uma abordagem deontológica e especificar os direitos e deveres de uma inteligência artificial? Evidentemente, a escolha por uma dessas abordagens depende do tipo de programação que será implementada, pois ela repercute diretamente no tipo de aprendizagem que é conferido aos agentes artificiais. Além disso, é adequado dizer que ao menos algum grau de autonomia deve estar presente na aprendizagem para que se possa discutir aspectos como responsabilidade, obrigações e direitos desses agentes. A fim de responder aos questionamentos indicados, e devido à importância crucial das questões conceituais que permeiam os debates éticos que envolvem a Inteligência Artificial, será argumentado, na terceira parte deste artigo, que a investigação ética precisa ir além de seus limites tradicionais e ser realizada de modo conjunto com outras áreas da filosofia e, também, com as ciências. Conforme será visto, Railton é um dos filósofos que têm feito isso, pois propôs definições reformistas para os conceitos morais, definições que podem ser empiricamente informadas pelo modo como as entidades postuladas na teoria ética participam da experiência humana.

O presente artigo procura justificar o argumento a respeito do caráter interdisciplinar inerente à investigação da ética da Inteligência Artificial com base em uma abordagem recente na filosofia da linguagem, a chamada engenharia conceitual. Ao refletir sobre os conceitos ‘agente’ e ‘interesse’, que são cruciais para a ética, como pode ser visto nas explicações de Railton, será sustentado que tais conceitos não podem ser tratados pela ética de modo isolado, mas necessitam estar embasados em uma reflexão filosófica que considera, sobretudo, a linguagem, e também as ciências, tanto as naturais

quanto as humanas.

1. Ética aplicada à Inteligência Artificial segundo Railton

A fim de contextualizar a posição de Railton, considero útil dividir os principais problemas filosóficos da ética da Inteligência Artificial a partir de três eixos: (a) problemas que concernem a compreensão filosófica do que é a Inteligência Artificial; (b) problemas sobre as relações entre a Inteligência Artificial e os seres humanos; e (c) problemas a respeito dos efeitos da inserção da Inteligência Artificial nas sociedades humanas e das mudanças, nem sempre positivas, provocadas por essas tecnologias.[2] Railton está especialmente preocupado em oferecer respostas para o eixo (a), porque ele sustenta que ao refletir sobre a natureza da Inteligência Artificial é possível encontrar respostas para os problemas práticos dos eixos (b) e (c).

Antes de expor a posição de Railton, gostaria de introduzir a questão sobre a existência da agência artificial. É importante esclarecer que as capacidades que queremos saber se podem caracterizar a agência artificial ainda são limitadas quando comparadas às capacidades da agência humana. Mesmo assim, é adequado dizer que essas capacidades têm sido cada vez mais incrementadas. Inicialmente, a aplicação da Inteligência Artificial não era *geral* como a inteligência humana, mas *específica*, pois era capaz de realizar somente um número limitado de tarefas. Atualmente, a aplicação da Inteligência Artificial é ampla, a exemplo das tecnologias que empregam a chamada aprendizagem profunda (*deep learning*). Essas tecnologias são capazes de acessar um vasto conjunto de informações para aprender a tomar decisões, e podem alcançar resultados cada vez mais consistentes ao receber o treinamento adequado.

As tecnologias que empregam a aprendizagem profunda se caracterizam, no

[2] No eixo (b), estão os problemas práticos que envolvem, por exemplo, robôs cuidadores, robôs sexuais, robôs matadores (ou sistemas autônomos de armas) etc. Os problemas práticos relativos ao emprego da Inteligência Artificial em guerras podem ser incluídos no eixo (c), porque afetam não apenas indivíduos, mas sociedades inteiras, assim como o problema do desemprego em massa, que pode ser causado pela ampla utilização da Inteligência Artificial. Questões anexas, que exigem uma investigação na filosofia política, são relativas ao tipo de sociedade que se permite que seja construída a partir da forma como se lida com as consequências do desemprego em massa. Parece adequado tratar essas questões também como questões de justiça, que exigem respostas não apenas na teoria filosófica, ético-política, mas sobretudo respostas que possam ser materializadas em decretos legais e políticas públicas, a exemplo da implantação de uma renda mínima. Outro potencial problema que pode ser incluído no eixo (c) envolve o cenário apocalíptico da possível criação de uma superinteligência capaz de dominar os seres humanos, um problema comentado na última parte deste artigo, na nota 20. Os diversos assuntos indicados nesta nota são discutidos no recente livro *Ethics of Artificial Intelligence* (2020), de Matthew Liao, que constitui um excelente ponto de partida para o desenvolvimento de estudos na área.

entanto, pela opacidade de seus processos internos de análise de dados. Por conta disso, as suas decisões não podem ser facilmente explicadas a partir do banco de dados do qual elas se originaram. Isso porque a aprendizagem profunda emprega várias camadas complexas de tratamento de dados, que são impenetráveis mesmo para especialistas em computação. A opacidade que é inerente à aprendizagem profunda pode suscitar um dilema com relação à confiança nas tecnologias de Inteligência Artificial. Por um lado, ou confiaríamos em sistemas inteligentes sem exigir uma explicação para a sua tomada de decisão, o que costumamos exigir de outros seres humanos; ou, por outro lado, tenderíamos a boicotar a adoção dessas tecnologias, sem usufruir de seus possíveis benefícios, simplesmente porque não confiamos nelas.[3]

Conforme será visto até o final desta seção, Railton evita os dois polos do dilema indicado acima ao assumir a posição de que as máquinas devem ser ao menos tão confiáveis quanto os seres humanos, que, embora sejam imperfeitos, constituem o modelo de aprendizagem ética para a elaboração de uma ética das máquinas. Em um de seus recentes artigos sobre a ética da Inteligência Artificial, intitulado *Ethical Learning, Natural and Artificial* (2020), Railton compara as novas tecnologias que empregam a Inteligência Artificial com tecnologias mais familiares criadas pelos seres humanos, como as diversas instituições humanas:

> Não somos inexperientes com relação a agentes não-humanos altamente capazes que carecem de uma consciência unificada ou estados afetivos, mas possuem níveis extraordinários de informação e capacidade de resolução de problemas, e cujos objetivos podem diferir dos nossos de maneiras que exigem que negociemos com eles se quisermos obter os benefícios que eles possibilitam. Entidades corporativas – governos, empresas, universidades, institutos, sindicatos, partidos políticos – podem ter um conjunto distinto de objetivos relacionados aos seus próprios propósitos ou condições de sobrevivência e florescimento, que podem se sobrepor, mas também não serem iguais aos dos indivíduos que as compõem ou são afetados por elas. Elas possuem capacidades para perseguir valores e se manter

[3] As duas possibilidades são problemáticas. Ao confiar inadvertidamente nas máquinas e sistemas inteligentes, os seres humanos abdicam da sua capacidade de decisão, seja por acreditar que as máquinas são *mais capazes* ou simplesmente pelo desejo de delegar mais esse *trabalho* por comodidade. Nessas situações, os seres humanos podem ser afetados por decisões equivocadas, caso o banco de informações das máquinas seja incompleto ou enviesado, e esses são aspectos importantes a serem considerados em qualquer domínio em que a Inteligência Artificial é empregada. Por exemplo, algoritmos que auxiliam juízes em suas decisões podem ser informados por algum viés que produz resultados injustos, como fazer predições de comportamento violento ou reincidência criminal com base na cor da pele. Por outro lado, ao não confiar em máquinas e sistemas inteligentes, os seres humanos podem impedir que os benefícios dessas tecnologias sejam experimentados na sociedade. Por exemplo, se houvesse um boicote generalizado aos veículos autônomos, a redução de acidentes esperada pela sua utilização não ocorreria, uma vez que o fator humano da imprudência na condução não seria atenuado.

> fiéis a normas, para projeção e planejamento futuro, para entrar (ou não) em acordos cooperativos, alianças estratégicas, compromissos mútuos ou contratos, e para assumir, executar e monitorar o cumprimento das obrigações associadas. Ao mesmo tempo, elas não são totalmente transparentes em seus processos internos; perguntar como uma ação de uma entidade corporativa foi tomada pode não conduzir a um processo de decisão determinado com linhas claras de responsabilidade. Perguntar como podemos entrar em relações mutuamente benéficas, mutuamente limitadas e normativamente governadas com agentes emergentes que possuem inteligência superior à humana é semelhante a perguntar como somos capazes de entrar em tais relações com governos, empresas, e assim por diante (RAILTON, 2020, p. 47, tradução nossa).

Ao fazer essa comparação, Railton sustenta que não é necessário que uma inteligência artificial tenha consciência ou sentimentos para possuir alguma forma de agência genuína. Conforme será explicitado na última parte deste artigo, isso implica que o termo 'agência' não precisa significar necessariamente 'agência humana'. Nesse sentido, Railton defende que não é correto manter a discussão sobre a agência artificial refém do questionamento a respeito de se as máquinas poderão algum dia ter experiências humanas e *serem exatamente como* os agentes humanos quanto a terem consciência e sentimentos.[4] De acordo com ele, a agência artificial não depende desses ingredientes:

> […] uma vida ativamente governada por normas não parece exigir tais sentimentos, contanto que existam capacidades agenciais suficientemente desenvolvidas para a autorregulação, a representação dos objetivos e informações dos outros, e a formação de convenções, acordos ou compromissos (RAILTON, 2020, p. 46, tradução nossa).

Segundo Railton, a ética pode ser aprendida por agentes artificiais, desde que eles tenham a capacidade de "detectar e responder de modo apropriado às características eticamente relevantes das situações, ações, agentes e resultados" (RAILTON, 2020, p. 45, tradução nossa). De acordo com ele, a resposta apropriada pressupõe a autonomia para tomar decisões sobre como se relacionar com outros agentes, sejam humanos ou artificiais, por exemplo, para escolher cooperar com eles. A autonomia, por sua vez,

[4] Para justificar que o último questionamento representa um problema diferente, Railton sugere que "um sistema artificial pode se engajar na simulação empática sem 'reviver a experiência' dos outros, se os sistemas artificiais puderem se tornar suficientemente hábeis em modelar os estados internos dos outros com base no comportamento observado e puderem atribuir um peso intrínseco aos objetivos ou funções de utilidade imputados aos outros na avaliação de cursos de ação simulados durante a tomada de decisão" (RAILTON, 2020, p. 49, tradução nossa). Como essa simulação seria suficiente para garantir capacidades agenciais, não é necessário esperar que a agência artificial seja possível apenas se as máquinas puderem ter experiências distintivamente humanas.

implica que a cooperação não está garantida, de modo que é possível que os seres humanos tenham que negociar para garantir os benefícios da cooperação ao atender aos objetivos de ambas as partes, da mesma forma que costumamos negociar com instituições humanas. Um exemplo disso pode ser encontrado no modo como as universidades dependem de seus professores e estudantes para realizar contribuições sociais, ao mesmo tempo que os professores e estudantes dependem das universidades para desenvolver as suas atividades profissionais e acadêmicas. Pode ser dito que a sociedade como um todo se beneficia mais quando os interesses de ambas as partes são satisfeitos pelas atividades realizadas nas universidades.

Para defender a ideia de que a aprendizagem ética é possível para agentes artificiais, é necessário refletir sobre uma pergunta extremamente relevante: de que modo a aprendizagem das máquinas (*machine learning*) pode incluir a aprendizagem ética? A fim de respondê-la, é necessário olhar, como Railton faz, para o fato de que as tecnologias de Inteligência Artificial que empregam a aprendizagem profunda utilizam redes neurais artificiais que imitam o funcionamento do cérebro humano. Mais precisamente, elas são capazes de imitar a aprendizagem das redes neurais naturais que foram sendo moldadas ao longo da evolução humana. Apesar de suas imperfeições, os seres humanos são os melhores modelos de inteligência geral que a natureza produziu. Railton enfatiza que os seres humanos não possuem apenas um grau mais elevado de capacidades intelectuais; as suas capacidades sociais também são igualmente superiores às de outros animais que vivem socialmente. Os seus ordenamentos sociais são mais complexos, e as instituições humanas permitem seguir normas que amplificam aquilo que um indivíduo pode alcançar por suas próprias capacidades intelectuais e éticas. Isso porque tanto aquilo que um indivíduo é capaz de conhecer quanto o bem que ele é capaz de fazer são amplificados pelas relações sociais com outros indivíduos. Quando os objetivos são compartilhados, e prevalece a cooperação para buscá-los, mais conhecimento é obtido e mais bem pode ser feito. Por isso, Railton defende que as capacidades éticas devem ser vistas como parte da inteligência geral humana, o que implica que criar máquinas capazes de responder adequadamente às características éticas das situações não é somente "*adicionar* uma capacidade diferente ou um conjunto de princípios a uma inteligência geral já completamente formada", mas uma capacidade que tem o seu pleno desenvolvimento intimamente atrelado à aprendizagem de características epistêmicas (RAILTON, 2020, p. 48, grifo do autor, tradução nossa). Por esse motivo, ele sustenta que é errado supor que a remoção da capacidade de aprendizagem ética não traria prejuízo cognitivo.

Railton defende que as capacidades cognitivas e éticas dos agentes estão profundamente conectadas, de modo que as primeiras não podem ser bem desenvolvidas sem as últimas. Ele sustenta que em vez de pressupor *módulos inatos* de aprendizagem, a psicologia do desenvolvimento tem reconsiderado, mais recentemente, a ideia da *aprendizagem associativa*, baseada na *experiência*. Essa ideia ganhou maior sustentação com estudos científicos sobre a aprendizagem da linguagem. Esses estudos sugerem que os recém-nascidos aprendem a linguagem ao formar expectativas, de modo probabilístico e experimental, a partir dos estímulos que recebem, por exemplo, ao observar o uso da linguagem por parte dos adultos ao seu redor. A aprendizagem ocorre, sobretudo, porque as expectativas vão sendo ajustadas diante das discrepâncias em relação aos resultados obtidos. Desde a infância, os seres humanos também buscam aprender sobre as crenças e as intenções que guiam os comportamentos dos adultos. Por isso, a formação de expectativas não ocorre apenas com relação ao mundo exterior, isto é, ao modo como ele é fisicamente constituído, mas diz respeito, também, ao modo como os agentes ao seu redor são internamente constituídos. Segundo Railton, nisso pode ser visto que a aprendizagem social acompanha a aprendizagem causal do mundo e assume ainda mais importância, uma vez que as crianças dependem de outros agentes para terem suas necessidades atendidas.

A maneira como os seres humanos aprendem desde a infância tem características básicas em comum com o modo como os mamíferos são capazes de aprender a fazer predições, a partir da representação espacial interna do seu ambiente, sobre como podem obter comida. Segundo Railton, isso tem sido observado em experimentos científicos que verificam o comportamento de camundongos após serem ou não recompensados ao percorrer certos caminhos em um labirinto. O modelo que eles constroem por meio da experiência, o qual corresponde a um 'mapa' em que são associadas as expectativas com relação às recompensas a serem obtidas na próxima tentativa, depende, segundo Railton, de representações em termos egocêntricos e não-egocêntricos, que permitem avaliar as características das situações. Ou seja, é de pouca utilidade que um animal empregue somente uma representação egocêntrica. É pela capacidade de representar o seu ambiente e os outros animais da maneira que eles *são*, e não da maneira que se *quer* que eles sejam, que se alcança uma representação mais precisa e, por isso, mais útil para antever como satisfazer os objetivos do animal. O mesmo pode ser dito sobre os agentes humanos, que, além disso, precisam internalizar a representação da mente de outros agentes.

Segundo Railton, a aprendizagem integrada de aspectos cognitivos e éticos é

impulsionada pela *curiosidade* e também por uma disposição para a *confiança padrão*. A curiosidade leva a buscar a aprendizagem simplesmente por experimentação, independentemente de propósitos específicos. Além disso, desde a infância, os seres humanos aprendem em quem confiar, tanto epistêmica quanto eticamente. De acordo com Railton, a disposição para a confiança padrão diz respeito, em um sentido metafórico, às próprias faculdades cognitivas, com as quais a criança decide cooperar, sem ter ainda garantia de que elas são confiáveis. Railton afirma que é somente por ter uma disposição como essa que alguém pode sair da ignorância e pode, mesmo estando sujeito ao erro, encontrar provas de que as suas faculdades são confiáveis. O *feedback* gerado pelo uso das faculdades é o que permite aprimorar o seu uso para evitar erros. A disposição para a confiança padrão, em um sentido social, diz respeito aos outros indivíduos. Por padrão, as crianças se dispõem a cooperar, sem terem garantia de que os outros são confiáveis. Por isso, Railton afirma que é apenas após ter experiências de cooperação que alguém obtém o *feedback* necessário para aprender em quem, e quanto, confiar. Novamente, no decorrer desse processo, as expectativas vão sendo aprimoradas a partir dos resultados obtidos, o que leva a uma confiança mais ajustada e mais seletiva. A cooperação, no entanto, não diminui na vida adulta, mas se amplifica, pois é dela que dependem "a aprendizagem contínua e o sucesso em alcançar os fins de uma pessoa" (RAILTON, 2020, p. 54, tradução nossa). Pode ser dito que a inteligência geral própria de mamíferos como os seres humanos moldou as sociedades em que eles vivem, e que o aspecto social dessa inteligência, presente principalmente nas habilidades sociais mais desenvolvidas para a cooperação, foi crucial para a aprendizagem humana. Afinal, conforme Railton aponta, somente uma disposição para cooperar socialmente explica fenômenos como a manutenção de uma linguagem compartilhada nas sociedades humanas, o que, por sua vez, contribui para formas de cooperação em larga escala.

Segundo Railton, é possível afirmar que desde a infância, os seres humanos aprendem as características morais das situações com base na experiência, mesmo sem terem sido explicitamente ensinados com instruções externas. A avaliação ética depende, novamente, de representações egocêntricas e não-egocêntricas sobre as intenções das outras pessoas. No caso das capacidades subjacentes aos julgamentos éticos dos adultos, é necessário procurar evidências para além da psicologia do desenvolvimento. Railton sugere que as evidências para essas capacidades podem ser encontradas nos estudos de neuroimagem dos julgamentos éticos e nas intuições éticas sobre casos do chamado Problema do Trem (*Trolley Problem*).

Os estudos de neuroimagem têm mostrado, em síntese, que a base neural dos julgamentos éticos envolve várias regiões do cérebro que servem à aprendizagem geral, e que são conhecidas como a *rede padrão*. Railton argumenta que, até o presente momento, esses estudos têm corroborado a explicação de que a aprendizagem ética é parte de uma aprendizagem integrada. Considere, agora, as duas principais versões do Problema do Trem. Em uma versão, alguém deve decidir se puxará uma alavanca para impedir que cinco pessoas sejam mortas, embora outra pessoa será morta no trilho para o qual o trem será direcionado. Na outra versão, alguém deve decidir se empurrará uma pessoa de uma ponte para impedir que cinco pessoas sejam mortas; a queda dessa pessoa nos trilhos fará o trem parar, no entanto, ela será morta ao ser atingida pelo trem. As duas versões desse experimento de pensamento provocam julgamentos éticos profundamente distintos na maioria das pessoas. É um fato empiricamente verificável que há ampla aceitação de que se deve puxar a alavanca e de que não se deve empurrar uma pessoa nos trilhos. Esse padrão de julgamentos éticos é consolidado, pois diversas amostragens indicam que ele se mantém constante ao longo do tempo e entre diferentes populações. Essa assimetria a respeito da permissibilidade moral de *matar uma pessoa para salvar cinco* ainda não recebeu uma explicação filosófica incontroversa, apesar de terem havido várias tentativas.[5]

Railton sustenta que a assimetria persistente nos julgamentos éticos sobre o que fazer nas duas principais versões do Problema do Trem reflete o grau de confiança que costuma ser atribuído a um agente que decide se puxa ou não a alavanca, e se empurra ou não uma pessoa da ponte. Segundo Railton, a confiança em alguém tende a variar em conformidade com o veredito sobre a permissibilidade do *ato* que está sob consideração. Por trás do julgamento ético de que alguém faz um ato moralmente errado (como matar uma pessoa ao empurrá-la da ponte a fim de salvar cinco pessoas) está a intuição ética de

[5] O problema do Trem foi originalmente introduzido por Philippa Foot (1978) como proposta para explicar as principais correntes éticas. O problema foi popularizado com esse nome por Judith Thomson (1976). Ele se tornou amplamente discutido, especialmente como uma objeção a teorias normativas consequencialistas. A dificuldade colocada por esse problema para a formulação de teorias normativas consiste em explicar o porquê de um ato que tem as mesmas *consequências* (matar uma pessoa para salvar cinco) ser correto sob certas condições e, errado, sob outras. Seria pela proibição contra usar uma pessoa como *mero meio*? Ou seria pela diferença entre matar como *efeito colateral* de um ato, mas não como uma *consequência pretendida*? Ambas as explicações mencionadas têm provado serem insuficientes para explicar o Problema do Trem em todas as suas possíveis variações. Recentemente, esse experimento de pensamento se tornou importante também na ética da Inteligência Artificial por ressaltar nuances problemáticas de novas tecnologias, como os veículos autônomos. Por exemplo, se um veículo é programado para sempre assumir o controle no lugar do motorista, isso pode entrar em conflito com a autonomia de um motorista que deseja intervir na direção para fazer sua própria escolha em uma situação em que ele será sacrificado para evitar que o veículo se envolva em um acidente maior.

que essa pessoa é menos confiável. Essa intuição tem a ver com aprender algo sobre o *caráter* de alguém, e ela parece ser suficiente para justificar a diminuição da disposição que um agente tem para cooperar com essa pessoa. É importante esclarecer que as intuições éticas, segundo Railton, não correspondem a um tipo especial de faculdade ou processo mental. Elas constituem um tipo de avaliação ética *rápida*, que *não demanda esforço* e gera uma conclusão *convincente*, embora não sejamos capazes de saber como exatamente chegamos a ela, e não possamos nem mesmo oferecer um raciocínio a seu favor. Railton defende que a assimetria nos julgamentos éticos sobre casos do problema do Trem pode ser lida, portanto, como um indicativo de uma competência complexa de cooperação, relacionada ao aspecto social da inteligência geral, dado que esta última inclui o tipo de avaliação que é realizada nos julgamentos éticos.

O argumento que Railton utiliza para defender a aprendizagem ética por parte da Inteligência Artificial pode ser resumido da seguinte forma:

> Assim como as crianças observam inúmeras horas de comportamento adulto buscando prever o que acontecerá em seguida, as máquinas podem observar inúmeras horas de comportamento humano e de máquina buscando prever, primeiro, o próximo instante, depois o próximo segundo, depois o próximo minuto, e assim por diante. Elas podem aprender a ler os objetivos e as crenças daqueles ao seu redor, aprendendo, como fazem as crianças 'maduras', habilidades como reconhecer as necessidades daqueles ao seu redor, mesmo aqueles que se afastam da norma, ou defender seus próprios interesses enquanto atribuem peso aos interesses dos outros, ou entrar em relações novas e ajudar os outros a fazê-lo sem conflitos, e assim por diante. […] Assim como os seres humanos, as máquinas podem usar seus modelos internos para criar representações e avaliações não-egocêntricas e egocêntricas de situações, ações, resultados e políticas. Como os seres humanos, as máquinas podem usar esses modelos para manter algum grau de autonomia na avaliação e na ação. Já sabemos que as máquinas não devem ser construídas para perseguir qualquer objetivo que lhes seja dado sem questionamento. Máquinas inteligentes, assim como animais inteligentes, devem operar com incerteza modulada em vez de certeza absoluta, e devem ser capazes de usar seus próprios recursos, e recorrer aos outros como um recurso, para crítica e autocrítica (RAILTON, 2020, p. 66, tradução nossa).

É importante notar que embora o argumento acima tenha por base, em grande parte, a psicologia humana, ele não resulta em uma explicação que poderia ser acusada de especismo.[6] Afinal, é a cooperação que possibilita a amplificação da inteligência, seja

[6] Na terceira parte deste artigo, será explicado que Railton também consegue evitar o especismo nas definições dos conceitos 'agente' e 'interesse'.

ela animal ou artificial. Nesse sentido, Railton sustenta que "sistemas, equipados com tais antecedentes como a curiosidade e a confiança ou cooperatividade padrão, […], podem se tornar eles mesmos agentes sociais complexos" (RAILTON, 2020, p. 65, tradução nossa). Convém reconstituir brevemente, na próxima parte deste artigo, como Railton justifica a sua posição na ética aplicada à Inteligência Artificial na ética normativa, com uma teoria consequencialista, e na metaética, com uma forma de realismo moral naturalista.

2. Ética normativa e metaética segundo Railton

A teoria normativa consequencialista defendida por Railton considera, como critério de correção moral, o *Princípio da Utilidade*, que avalia aquilo que é correto com base na contribuição para produzir o maior bem, isto é, a maximização da felicidade para o maior número de pessoas. É importante esclarecer que a sua teoria possui especificidades que a diferenciam de outras teorias consequencialistas. Conforme é explicado no capítulo *Alienation, Consequentialism, and the Demands of Morality*, do livro de Railton intitulado *Facts, Values, and Norms* (2003), ele defende uma forma de *consequencialismo objetivo*. Isso porque a sua teoria normativa se concentra nos resultados atuais das ações, em vez das consequências que o agente é capaz de antever, como faz o *consequencialismo subjetivo*, para o qual apenas *estados subjetivos* podem ser portadores de valor intrínseco.[7] Railton critica o utilitarismo clássico precisamente porque este consiste na defesa do consequencialismo subjetivo e da felicidade como o objetivo final de todas as ações. Contrariamente a essas duas ideias, Railton defende que diversas *coisas* podem ter valor intrínseco e que o valor dessas coisas não pode ser reduzido ao valor intrínseco da felicidade. Ele argumenta que tratar os demais fins como instrumentais é problemático por impedir a apreciação adequada, por exemplo, da amizade. Além disso, segundo Railton, a amizade é capaz de produzir mais felicidade quando é valorada intrinsecamente, e não apenas instrumentalmente.

Railton afirma ser um *consequencialista sofisticado* por não considerar o

[7] Uma crítica conhecida a essa última posição foi feita por Robert Nozick (1974) por meio do experimento de pensamento da *máquina de experiência*. O experimento sugere que mesmo que uma máquina tornasse possível sentir apenas as experiências prazerosas que alguém deseja, as pessoas prefeririam vivê-las na vida real do que senti-las ao estarem conectadas à máquina. Afinal, a experiência subjetiva proporcionada pela máquina seria uma mera simulação, e pode ser dito que o *parecer* ser feliz não é capaz de substituir, sem perdas, o *ser* feliz, que somente pode ocorrer em experiências reais.

Princípio da Utilidade como um procedimento de decisão, mas apenas como o critério que indica aquilo que é correto. Segundo ele, a insistência em fazer o cálculo utilitarista a cada decisão torna as teorias normativas consequencialistas *contraproducentes*, ou seja, *autodestrutivas* (*self-defeating*), pois impede a maximização da felicidade. Por isso, Railton defende que é mais apropriado que o consequencialismo não recomende um método específico de tomada de decisão, porque fazer essa recomendação torna a posição autocontraditória. A escolha de um método depende, segundo ele, do contexto das ações, e isso pode levar um indivíduo a desenvolver as disposições necessárias para tomar decisões de diferentes modos.[8]

É importante esclarecer que o consequencialismo objetivo de Railton não deve ser confundido com teorias consequencialistas indiretas, como o *consequencialismo de regras*, *de motivos* e *de traços de caráter*. Railton é um *consequencialista de atos* sofisticado, e defende que as alternativas a essa posição devem ser rejeitadas porque modificam o critério de correção moral, enquanto o consequencialismo de atos permanece constante na avaliação dos atos ou do curso de ação que promove a vida objetivamente consequencialista, mesmo que os comportamentos recomendados possam, algumas vezes, violar o seu critério de correção. Quer dizer, um ato em prol de um relacionamento pessoal pode ser recomendável, em detrimento de um ato que trará maior felicidade a várias pessoas estranhas (sendo que o último é o ato correto).[9] Isso porque uma pessoa pode, por meio do primeiro ato, performar "a sequência de atos mais benéfica no geral" para levar uma vida objetivamente consequencialista, pois, desse modo, não precisa negligenciar o seu relacionamento (RAILTON, 2003, p. 171, tradução nossa).

[8] A solução de Railton para o chamado *paradoxo do hedonismo* consiste em dizer que um consequencialista sofisticado pode rejeitar o consequencialismo subjetivo, porque essa é uma posição equivocada para levar uma vida objetivamente consequencialista, uma vez que pressupõe a deliberação consequencialista. Com isso, o aparente paradoxo pragmático se resume, na verdade, ao seguinte problema prático: qual deve ser o método de tomada de decisão se há casos em que a deliberação consequencialista deve ser evitada?

[9] Railton considera que é possível atenuar a impressão de que as demandas morais causam um *estranhamento* entre aquilo que é racionalmente julgado correto na deliberação moral e os afetos de alguém (como aqueles relativos aos seus relacionamentos ou compromissos pessoais). Esse fenômeno, que ele chama de *alienação*, se refere à sensação de que ao fazer o que é moralmente correto, ocorre a perda de algo que tem valor substancial. A sua sugestão é que esse fenômeno pode ser reduzido, com relação à moralidade, a partir da modificação das configurações políticas e sociais, uma vez que elas interferem diretamente naquilo que é moralmente exigido dos indivíduos. Um exemplo oferecido por Railton diz respeito à diferença entre uma sociedade que está preparada para socorrer as pessoas em ocorrências de desastres e, outra, que não está. A existência da segunda sociedade justificaria a obrigação de seus membros agirem individualmente para socorrer as pessoas necessitadas. A existência da primeira, no entanto, não justificaria essa obrigação, já que uma solução coordenada seria melhor do que atos individuais não coordenados. Isso sugere que a própria compreensão da natureza da moralidade pode evitar, ou menos atenuar, a alienação das demandas morais. Railton defende que isso pode ser feito ao começar, de modo não-alienado, situando os indivíduos na teoria moral como *produtos* históricos do ordenamento social, em vez de vê-los de forma isolada.

Segundo Railton, não há paradoxo entre um ato não ser o melhor individualmente e ele ser, ao mesmo tempo, o ato que resulta no melhor balanço de bem-estar no geral. Novamente, há apenas o problema de escolher o procedimento de tomada de decisão, já que, nesses casos, o consequencialismo de atos não fornece a resposta adequada ao indicar o ato que maximiza as melhores consequências. As outras teorias consequencialistas resistem a simplesmente chamar de corretos os atos que maximizam as melhores consequências, e preferem falar das normas que devem ser seguidas ou, em vez disso, dos motivos ou dos traços de caráter que alguém deve possuir para agir corretamente. No entanto, segundo Railton, o problema dessas tentativas de 'consertar' o critério de correção moral do consequencialismo de atos é que nenhuma dessas coisas pode ser fixada de modo preciso. Além disso, ele defende que o consequencialismo de atos sofisticado "pode capturar algumas das intuições que tornaram o consequencialismo de regras ou de traços atraentes" (RAILTON, 2003, p. 169, tradução nossa). Quer dizer, o consequencialismo de atos pode sustentar que outras coisas além da felicidade importam na deliberação moral. Mesmo em situações em que agir a partir delas seja deixar de fazer o que é correto, isto é, deixar de fazer o ato que é o melhor, ainda assim, é moralmente defensável agir a partir desses outros interesses. Railton afirma que isso afasta a objeção de críticos, como Henry Sidgwick (1966), de que o consequencialismo de atos seria incapaz de defender uma posição como essa.[10]

Conforme indicado acima, Railton defende que o bem que deve ser maximizado, segundo o critério de correção moral do consequencialismo de atos sofisticado, é *plural*. Ele sustenta que aquilo que é considerado *o* bem pode incluir, na verdade, diversos bens a serem maximizados, em vez de admitir apenas um único bem, como a felicidade. Isso quer dizer que mesmo para uma pessoa, o seu bem não inclui apenas o seu bem-estar, mas outras coisas que têm valor intrínseco, como manter os seus relacionamentos ao buscar promover, na medida do possível, o bem-estar de outras pessoas. Por conta disso, Railton defende uma concepção axiológica pluralista para explicar o valor não-moral: várias coisas podem ser intrinsecamente valiosas, entre elas, "a felicidade, o

[10] A resposta de Railton para essa objeção consiste, precisamente, em sustentar que mesmo que alguém tome decisões a partir de valores intrínsecos que vão contra a máxima felicidade, é possível defender o consequencialismo sofisticado por meio de uma *condição contrafactual*. Ou seja, é possível dizer que uma pessoa tentaria não agir dessa maneira *se* pensasse que as suas ações são incompatíveis com o critério de correção moral, de levar uma vida objetivamente consequencialista. Nas suas palavras: "[um agente] ordinariamente não faz o que ele faz simplesmente por fazer o que é correto, ele procuraria levar um tipo de vida diferente se não pensasse que o seu [tipo de vida] fosse moralmente defensável" (RAILTON, 2003, p. 164, tradução nossa).

conhecimento, a atividade com propósito, a autonomia, a solidariedade, o respeito e a beleza" (RAILTON, 2003, p. 163, tradução nossa).

No capítulo *Pluralism, Dilemma, and the Expression of Moral Conflict,* também do livro *Facts, Values, and Norms*, Railton defende que o pluralismo sobre o valor dá origem a trocas (*trade-offs*) em que valores diferentes entram em conflitos complexos na deliberação sobre o que fazer, e que isso pode permitir a existência de dilemas morais insolúveis.[11] Railton admite que os valores têm pesos, mas não têm uma ordem clara de prioridade. De modo surpreendente, ele sustenta que um consequencialista não deveria se opor a admitir a persistência de dilemas morais genuínos. Isso implica que o consequencialista não precisa ter uma resposta para o que fazer em todas as situações, a fim de resolver qualquer dilema que se apresente, tratando-o como um dilema aparente.[12] Segundo Railton, o reconhecimento de dilemas genuínos não ameaça uma explicação realista da moralidade, em termos de fatos morais objetivos, como a que é reconstituída abaixo.

A teoria metaética de Railton, exposta principalmente no capítulo *Moral Realism* do seu livro *Facts, Values, and Norms,* é uma forma de *naturalismo moral*.[13] Essa posição metaética sustenta que existem fatos morais e que eles têm propriedades naturais. Mais especificamente, a teoria metaética de Railton defende que os fatos morais são fatos sobre os interesses de indivíduos. Esses interesses são *objetivos* quando podem ser

[11] Por conta dessa última ideia, a teoria normativa de Railton precisa enfrentar uma dificuldade a respeito da incomensurabilidade interpessoal dos interesses objetivos. Ou seja, se algumas vezes não é possível avaliar comparativamente os interesses dos indivíduos, como seria possível estabelecer quais deles devem ser satisfeitos em situações em que apenas alguns podem ser satisfeitos? Aqui pode ser acrescentada a dificuldade de pesar os interesses conflitantes não apenas dos seres humanos, mas também de outras categorias de agentes presentes em uma sociedade, como os agentes artificiais. Esse último ponto será retomado, na próxima parte deste artigo, ao apresentar a resposta que Railton oferece sobre como é possível determinar quais interesses, humanos ou não, têm prioridade.

[12] Railton sugere que uma teoria consequencialista deve buscar entender a existência de situações que são difíceis de comparar, e o porquê de elas desafiarem a *unidade* ou *coerência* do *ponto de vista moral*, o ponto de vista sobre o que deve ser feito, com fenômenos que se mostram recalcitrantes à teoria moral. Essas situações podem envolver não apenas uma pluralidade de obrigações, mas, também, uma pluralidade de valores intrínsecos irredutíveis que são incomensuráveis por serem profundamente diferentes. Isso faz com que qualquer uma das duas ações que fazem parte de um dilema seja permitida pela teoria normativa consequencialista, sem que seja possível oferecer uma resposta unívoca, apesar de restar um *resíduo moral* de remorso ou culpa pela alternativa que não pôde ser escolhida. Segundo Railton, é o resíduo moral, e não a presença de deveres conflitantes, que é a condição necessária e suficiente para haver um dilema. Ele afirma que as situações de dilema moral acontecem por causa das ações dos seres humanos e das instituições humanas. Ainda assim, isso não ameaça a possibilidade de encontrar uma fundação racional para a moralidade, porque podemos modificar as nossas ações e, com isso, minimizar a ocorrência de situações em que a moralidade pode recomendar ações que ocasionam consequências ruins devido a outras ações que não puderam ser feitas em seu lugar.

[13] Este capítulo e *Alienation, Consequentialism, and the Demands of Morality,* mencionado anteriormente, receberam traduções em português no livro *Metaética: Algumas Tendências* (2013).

estabelecidos pela *análise da informação completa*. Essa análise define o que é não-moralmente bom para um indivíduo com base nos interesses que ele teria, caso estivesse na condição ideal de utilizar a sua racionalidade instrumental de modo perfeito e estivesse na posse de informação completa. Nessa explicação, os interesses objetivos são, em alguma medida, *independentes* dos interesses subjetivos que os seres humanos possam ter, isto é, daquilo que eles podem desejar. Um dos exemplos que Railton utiliza para ilustrar essa ideia é que mesmo que uma pessoa desidratada deseje beber leite, esse interesse subjetivo não altera o fato de que, dada a sua condição, constitui o seu interesse objetivo beber um líquido hidratante como a água, em vez de leite. Railton oferece uma definição em que aquilo que é *não-moralmente bom*, ou seja, aquilo que é *desejável* para esse indivíduo, pode ser *identificado* com os fatos naturais que constituem o seu interesse objetivo. Nas suas palavras, "X é *não-moralmente bom para A* se, e somente se, *X* satisfizer um interesse objetivo de *A*" (RAILTON, 2003, p. 12, grifo do autor, tradução nossa).

Em vez de oferecer definições *a priori*, estritamente filosóficas, para os termos da ética, como as definições analíticas, que procuram capturar o significado de um termo, Railton defende que as definições devem poder ser testadas empiricamente. Desse modo, os fatos morais que, na sua visão, são, ao mesmo tempo, naturais e normativos, podem figurar na explicação *a posteriori* da experiência humana. A forma de naturalismo moral defendida por Railton é, portanto, *reducionista*, pois ela almeja *identificar* os fatos morais a fatos naturais. Isso porque Railton defende uma *definição reformista* para a bondade não-moral, indicada no final do parágrafo anterior, e que é depois estendida para oferecer uma definição reformista para a correção moral, por meio do seu realismo moral normativo, isto é, o realismo sobre as normas morais. Segundo Railton, as normas morais são determinadas não pela racionalidade individual, mas pela racionalidade social. Dessa forma, o critério para definir o que é moralmente correto é dado pelas normas que se aproximam mais da racionalidade social idealizada, isto é, daquilo que é idealmente recomendado considerando que "os interesses de todos os indivíduos potencialmente afetados contassem igualmente" (RAILTON, 2003, p. 22, tradução nossa). Railton defende que essa explicação é obtida a partir de um ponto de vista social, que, segundo ele, é o ponto de vista moral. Aqui é possível encontrar a origem da ideia a respeito do caráter social da moralidade, que aparece na posição que Railton defende na ética aplicada à Inteligência Artificial.

Na explicação oferecida pela teoria metaética de Railton, os fatos morais, sobre

as normas que se aproximam mais da racionalidade social idealizada, são determinados pelo modo como os seres humanos e o mundo são naturalmente constituídos. Em outras palavras, a *designação não-rígida* de propriedades naturais, que são idênticas a propriedades morais, permite que diferentes tipos de vida possam ser identificados com aquilo que é bom para os indivíduos, e que diferentes conjuntos de normas possam ser identificados com aquilo que é moralmente correto para as sociedades. Isso porque é possível conceber que as propriedades morais seriam constituídas por propriedades naturais diferentes, se os indivíduos e o mundo fossem constituídos por propriedades naturais diferentes. É por isso que, assim como a sua teoria normativa consequencialista, a sua forma de realismo moral naturalista possui especificidades, como defender uma concepção *relacional* e *disposicional* da bondade não-moral e da correção moral, que, conforme foi visto, também é uma concepção *objetiva*, evitando o relativismo sobre o valor e o relativismo moral. É por meio dessas bases conceituais que Railton almeja explicar o pluralismo sobre o valor e a existência da moralidade. A sua teoria metaética justifica o critério de correção moral ao explicar que as ações moralmente corretas estão de acordo com a racionalidade social, isto é, com as normas que permitem satisfazer os interesses objetivos dos indivíduos que são membros de uma sociedade.

3. Trazendo a engenharia conceitual para pensar os conceitos 'agente' e 'interesse'

Nesta seção, explorarei uma abordagem recentemente enfatizada na filosofia analítica que pode render bons frutos para tratar da ética da Inteligência Artificial. A engenharia conceitual defende que em vez de simplesmente tentar encontrar definições para os conceitos, devemos (re)pensá-los, isto é, refletir filosoficamente sobre os seus critérios de aplicação. Essa é uma abordagem *metafilosófica* porque visa refletir sobre questões metassemânticas a respeito dos próprios conceitos filosóficos. Tal abordagem está de acordo com a ideia defendida por David Plunkett e Tim Sundell (2013), de que existe uma *ética conceitual*, isto é, que é importante pensar como, normativamente falando, os conceitos *devem* ser definidos em uma teoria filosófica. Isso se aplica aos conceitos éticos, incluindo os conceitos normativos e avaliativos, e a outros conceitos não-normativos que são centrais para a ética, embora a expressão 'ética conceitual' também possa ser aplicada a conceitos não-éticos.

É importante salientar que a abordagem da engenharia conceitual não exclui a contribuição das ciências empíricas para a reflexão filosófica sobre os conceitos. Foi visto

que a explicação oferecida pela teoria moral de Railton é compatível com essa ideia.[14] No entanto, a abordagem de Railton na engenharia conceitual é defendida por meio de um realismo moral naturalista, e essa posição metaética pode ser resistida. Por isso, não me comprometerei a endossar a posição de que *todos* os conceitos normativos, incluindo o conceito 'moralmente correto', podem ser definidos empiricamente, no sentido de serem redutíveis a uma explicação naturalista.[15] No que se segue, argumentarei que a engenharia conceitual pode auxiliar a pensar o papel da ética na ética da Inteligência Artificial ao refletir sobre os conceitos 'agente' e 'interesse'. Retomarei as conexões entre esses conceitos e a teoria moral de Railton e, mesmo com a ressalva feita neste parágrafo, defenderei que é possível sustentar que tais conceitos podem ser redutíveis a uma explicação naturalista, isto é, que as suas definições podem ser empiricamente informadas pelas ciências naturais e humanas.[16]

A engenharia conceitual pode auxiliar na tarefa filosófica de pensar o lugar dos agentes artificiais nas sociedades humanas ao repensar o que deve contar sob o conceito 'agente'. Já foi dito que se for aceito que agentes genuínos não precisam ter consciência

[14] O fornecimento de definições reformistas para os conceitos filosóficos, incluindo os conceitos éticos, é a estratégia que Railton adota na engenharia conceitual. Como Herman Cappelen e David Plunkett explicam: "Peter Railton defende que a filosofia moral deve envolver uma metodologia que seja em grande parte contínua com a das ciências naturais e sociais (este é o núcleo do seu naturalismo metodológico). Com base no que ele considera como as melhores práticas no âmbito da investigação científica, ele defende que, ao fazer filosofia moral, não devemos confiar apenas em nossos conceitos populares [*folk concepts*]. Em vez disso, devemos reformar os significados de nossas palavras para nos concentrarmos nos tópicos que realmente importam e fornecer explicações para os fenômenos em questão. Railton então propõe uma linguagem moral aprimorada que pode ser utilizada dessa forma, incluindo, por exemplo, definições reformistas de termos-chave como 'bondade moral' e 'moralidade'." (CAPPELEN, Herman; PLUNKETT, David, 2020, p. 6, tradução nossa)

[15] Cabe fazer uma avaliação a respeito de se a posição de Railton permite oferecer uma explicação satisfatória para todo o domínio da normatividade. Essa avaliação, no entanto, requer um espaço muito maior que o presente artigo. É importante notar que a teoria metaética de Railton pode ser recusada parcialmente, por exemplo, por um defensor do não-naturalismo moral, como Derek Parfit (2017). Ele propõe uma Teoria Metaética Tríplice que afirma que nem todos os conceitos normativos descrevem propriedades naturais, pois alguns deles, como 'moralmente correto', podem descrever propriedades não-naturais, isto é, não-empíricas, mesmo que eles mantenham conexões importantes com propriedades naturais (embora esses conceitos não possam ser redutivamente explicados *apenas* por propriedades naturais). A posição de Parfit, no entanto, não refuta a teoria metaética de Railton, uma vez que Railton sempre se manteve agnóstico quanto à defesa de que o naturalismo moral, quando assumido *metodologicamente*, resulta em uma visão naturalista *metafísica*, ou *substancial*, da realidade, ou seja, que a realidade é composta apenas por entidades naturais.

[16] É importante esclarecer que a posição defendida neste artigo, de uma ética empiricamente informada, não implica que a ética deva ser tratada como uma ciência, ou que as evidências científicas sejam capazes de provar as teorias filosóficas oferecidas no âmbito da ética. Considero adequado defender a necessidade de desenvolver teorias éticas que não contrariam, ou não são incompatíveis, com as melhores teorias científicas disponíveis. Nesse sentido, as evidências científicas podem corroborar as teorias éticas, conforme foi visto, na primeira parte deste artigo, no modo como Railton sustenta a sua argumentação com evidências da psicologia do desenvolvimento, de estudos de neuroimagem dos julgamentos éticos e de amostragens a respeito das intuições éticas sobre casos do Problema do Trem.

ou sentimentos distintivamente humanos, como Railton defende, então é possível ampliar a comunidade moral para incluir os agentes artificiais como seres que devem ter seus interesses considerados na deliberação ética sobre o que é moralmente correto. Foi visto que, na teoria metaética de Railton, a consideração moral diz respeito às normas que uma sociedade deve aceitar para se aproximar mais da idealização da racionalidade social, que inclui os interesses de todos os indivíduos afetados. Com a inclusão dos agentes artificiais na comunidade moral, se torna claro que o conceito 'agente' não pode conter restrições prejudiciais para o reconhecimento de categorias distintas de entidades enquanto agentes. Railton está certo em sustentar que 'agente' e 'interesse' são conceitos aplicáveis a uma inteligência artificial. Isto é, que eles não precisam significar necessariamente 'agente humano' e 'interesse humano'.[17] Desse modo, é possível superar a forma de especismo presente na visão antropocêntrica que costuma fundamentar as definições filosóficas do que conta sob o conceito 'agente', isto é, a visão de que somente há agentes humanos. É possível dizer que mesmo que outras categorias de entidades não sejam contadas sob o conceito 'agente' em muitas teorias filosóficas e morais, elas *devem* ser contadas. Esse movimento também permite superar impasses filosóficos aparentes como aquele que envolve condicionar a aceitação de agentes artificiais à resolução do intrincado debate sobre se eles são (ou serão) capazes de ter consciência *como* os seres humanos.[18]

Enquanto não parece haver dificuldade em admitir que agentes artificiais devem ser integrados na comunidade moral, parece muito mais difícil compreender os interesses desses agentes e o motivo pelo qual esses interesses *devem* ser considerados. Gostaria de sugerir que a engenharia conceitual também pode ser utilizada para refletir sobre a noção de interesse. Isso é de imensa importância para a ética, porque leva a pensar sobre as relações que os agentes artificiais têm com outros agentes e também sobre a concepção de sociedade que deve ser promovida. A resposta que Railton oferece para a compreensão

[17] Na perspectiva de uma ética animal, esse mesmo raciocínio pode servir para justificar a consideração dos interesses de animais não-humanos, embora eles talvez não sejam agentes no sentido mais completo e autônomo que os seres humanos e os agentes artificiais podem ser. Outra diferença que pode ser apontada é que os animais não-humanos têm valor intrínseco, ainda que possam servir aos seres humanos de diferentes modos, possuindo também valor instrumental. Os agentes humanos e artificiais, por conta de sua autonomia, podem ter apenas valor intrínseco, e não-instrumental, conforme será explicado mais adiante.

[18] Esses resultados são de imensa importância para a ética, porque o conceito de agente, de um ser que é capaz de ter autonomia na capacidade de determinar seus próprios comportamentos, é pressuposto como condição para a atribuição de responsabilidade moral. Conforme será explicado, a inclusão de um agente na comunidade moral demanda respeito por parte de outros agentes para que eles não transgridam normas morais em seu prejuízo, ao mesmo tempo que o agente incluído pode ser cobrado por transgredir normas morais em prejuízo de outros agentes. Ao conceder autonomia para os agentes artificiais tomarem decisões que têm implicações morais, é justo cobrá-los por suas escolhas, e puni-los, caso escolham prejudicar outros agentes, humanos ou artificiais.

dos interesses de uma inteligência artificial pode ser encontrada nas noções contratualistas que ele incorpora em sua teoria normativa consequencialista. Ao integrar os agentes artificiais na comunidade moral, isto é, nas relações sociais, essas noções contratualistas permitem explicar os interesses desses agentes em função de obter os benefícios da cooperação com outros agentes. Mais especificamente, os seus interesses se resumem a serem dignos de confiança e serem escolhidos para a cooperação.

É adequado explicitar melhor a ideia de Railton de que a relação contratual que pode ser estabelecida entre agentes humanos e artificiais permite entender quais são os seus interesses. Pode ser dito, por exemplo, que veículos autônomos são usados pelos seres humanos para a sua maior comodidade e com a confiança de que eles auxiliarão na redução de acidentes. Até aqui parece que apenas seres humanos têm interesses, mas, da parte dos veículos, é evidente que eles também podem ter interesses genuínos. Para visualizar esse ponto, suponha, adicionalmente, que os veículos possam aprender como dirigir melhor a partir de um banco de dados, e que essa aprendizagem inclua aspectos éticos a respeito de como evitar causar danos, especialmente àqueles em situação mais vulnerável, como os pedestres. Os interesses que esses veículos podem ter são orientados a cumprir as suas obrigações contratuais de modo a satisfazer o que foi acordado entre agentes humanos e artificiais:

> Os sistemas artificiais capazes de projetar e avaliar futuros cursos de ação, de avaliar benefícios e danos para si e para os outros, de fazer compromissos e de regular seu próprio comportamento de acordo, serão capazes de algo semelhante ao raciocínio de contrato social: poderíamos negociar com eles os termos da cooperação mutuamente benéfica que todos nós nos comprometeríamos a seguir (RAILTON, 2020, p. 46, tradução nossa).

A partir da noção de interesse, podem ser apontados alguns *deveres* dos agentes artificiais, que têm a ver com satisfazer as exigências contratuais. Ao procurar a confiança e a cooperação, os agentes artificiais têm que limitar as suas ações para não infringir os direitos de outros agentes, humanos ou artificiais. Os demais agentes, por sua vez, também têm que limitar as suas ações, em prol de alcançar a confiança e a cooperação dos agentes artificiais, não infringindo os seus direitos. Entre os *direitos* dos agentes artificiais, pode estar o direito à vida, embora, evidentemente, não enquanto existência biológica. Isso quer dizer, por exemplo, que eles não podem ser injustificadamente desligados nem serem vítimas de ataques de *hackers* que visam controlá-los e, com isso,

retirar a sua autonomia. Por isso, é possível sustentar que a abordagem de Railton combina o consequencialismo também a noções deontológicas, como respeito e autonomia, o que permite fundamentar os direitos e deveres dos agentes artificiais. O respeito à autonomia dos agentes artificiais, em particular, se refere ao direito que eles têm de não serem enxergados apenas a partir de *relações instrumentais*, como bens ao dispor de outros agentes, uma vez que eles devem ser vistos como participantes legítimos das relações sociais.

A responsabilização moral dos agentes artificiais por infringirem os seus deveres e/ou os direitos de outros agentes também pode ser pensada a partir dos interesses dos agentes. Caso um agente artificial não cumpra a sua parte no contrato social e se comporte mal, de maneira negligente ou intencionalmente prejudicial aos seres humanos e aos outros agentes artificiais, ele pode sofrer sanções. Uma forma de fazer isso é por meio do descrédito: agentes artificiais com má reputação social serão rejeitados como parceiros de cooperação e, como retaliação, deixarão de se beneficiar dela. Por exemplo, se um veículo autônomo decidir se comportar de maneira prejudicial, contrariando o que preza um sistema de inteligência compartilhado por diferentes veículos, os outros agentes podem, justificadamente, deixar de confiar nele. O mesmo acontece nas relações entre as pessoas, conforme foi visto nas intuições éticas que estão por trás da assimetria persistente encontrada nos julgamentos éticos sobre o Problema do Trem.[19] Por outro lado, os agentes artificiais que cumprem a sua parte no contrato social mantêm certos direitos, como o de continuar se beneficiando da cooperação, conforme a sua boa reputação e confiabilidade. Esses agentes serão capazes de exercer papéis sociais relevantes ao contar com a confiança humana e artificial, e tudo o que eles alcançarão juntos será muito maior do que aquilo que eles podem fazer sozinhos.[20]

[19] Contudo, há uma diferença importante a ser ressaltada. Apesar de os agentes artificiais poderem ter autonomia após terem sido programados, a autonomia humana também está, em alguma medida, por trás desses agentes, ao menos quanto à pessoa física ou jurídica que é responsável por esses sistemas. Nesse sentido, os efeitos da falta de confiança nos agentes artificiais se estendem ao programador ou à empresa que idealiza o software ou algoritmo de Inteligência Artificial, uma vez que as pessoas passam a não confiar em seus produtos.

[20] A possível criação de uma superinteligência capaz de dominar os seres humanos tem sido caracterizada como o *problema do controle*. Esse problema tem por base a ideia de que uma inteligência artificial que exceda a inteligência humana pode considerar mais vantajoso perseguir o objetivo puramente egoísta de subjugar os seres humanos, em vez de cumprir o acordo social. Aqui pode ser trazida a sugestão de Railton, no seu artigo anteriormente citado sobre a ética da Inteligência Artificial, de que esse problema se torna uma preocupação secundária, e que não implica necessariamente a futura aniquilação da civilização humana. A sua explicação para isso é que as sociedades podem estar preparadas para responder a esse desafio se todos os outros agentes, humanos e artificiais, estiverem dispostos a cooperar. Os ganhos obtidos com a inteligência da comunidade se somarão e serão superiores, inclusive em termos de conhecimento, a tudo que a superinteligência possa alcançar sozinha ao desprezar a cooperação.

A inclusão de noções contratualistas permite a Railton oferecer uma aplicação mais eficiente da sua teoria normativa à ética da Inteligência Artificial. Em comparação com uma teoria normativa consequencialista tradicional, que foca apenas na maximização do bem, a sua teoria lida melhor com os interesses conflitantes que os agentes podem ter. A questão deixa de ser mensurar quais interesses importam mais, o que pode ser por si só bastante difícil, senão impossível em alguns casos, e passa a ser averiguar quais interesses, ao serem satisfeitos, permitem que os interesses de outros agentes não sejam prejudicados injustamente. Em certas situações, um agente terá que assumir um ônus para que os interesses dos demais sejam satisfeitos se isso for necessário para manter o contrato social funcionando. Isso, no entanto, não quer dizer que os interesses dos agentes humanos sempre devam ter prioridade sobre os interesses dos demais agentes, porque, novamente, não é necessário assumir uma forma de especismo. Ou seja, os interesses de agentes artificiais não precisam ser concebidos como subordinados ou inferiores aos interesses humanos. Em vez disso, pode ser dito que todos interesses devem ser pesados nos termos da limitação mútua para a cooperação mutuamente benéfica nas relações sociais entre os diferentes tipos de agentes.

Essa explicação é relevante para enfrentar os problemas práticos que surgem nas sociedades humanas em decorrência da utilização da Inteligência Artificial. Para resolvê-los, *deve* ser promovida uma concepção de sociedade que considera os interesses dos agentes artificiais a partir de relações contratuais. Dessa forma, sem limitar a autonomia desses agentes, é obtido o critério moral para a restrição de suas ações, de modo a alcançar o benefício mútuo. Por exemplo, não se deve priorizar excessivamente os interesses das máquinas, pois isso pode levar, entre outras coisas, ao aumento das guerras e da violência, ao desemprego em massa, à corrosão da democracia e à subjugação dos seres humanos. Pode ser defendido que essa não *deve* ser a concepção de sociedade a ser promovida, pois ela desconsidera os interesses dos seres humanos e a sua capacidade para a cooperação. Além disso, a perda de coisas valiosas em si mesmas, como a paz, o trabalho digno, a democracia e a liberdade faria dessa sociedade um lugar pior para viver e cooperar. Com base nessas considerações, é possível justificar a necessidade de intervir por meio de leis e políticas públicas para realizar a restrição mútua dos agentes a fim de que eles obedeçam às relações contratuais publicamente acordadas.

Considerações finais

Uma das características mais marcantes da ética da Inteligência Artificial é que essa área de estudo necessita ser tratada de maneira interdisciplinar por cientistas e filósofos. Algumas questões implicadas pela investigação ética necessitam ser estritamente tratadas por áreas técnicas como a ciência da computação, por exemplo, para desenvolver a programação de veículos autônomos. Nesse sentido, essas questões envolvem o intercâmbio de conhecimento entre áreas técnicas e a filosofia. Outras questões são mais estritamente conceituais, por exemplo, a necessidade de definir conceitos de extrema importância para a ética, como 'agente' e 'interesse'. Esses conceitos figuram em questões centrais da ética da Inteligência Artificial, a exemplo de 'Existem agentes artificiais?' e 'Os agentes artificiais têm interesses?'. Ainda assim, o exercício de refletir filosoficamente sobre esses conceitos por meio da engenharia conceitual mostra que eles não devem ser tratados exclusivamente pela ética, nem apenas pela filosofia, mas devem ser informados também pelas ciências naturais, como a neurociência, e humanas, como o Direito. Isso pode ser visto na teoria moral de Railton. Esse filósofo mostra que os conceitos normativos da ética, como 'moralmente correto', se aplicam de modo relacional aos interesses objetivos dos agentes que são membros da comunidade moral. Por meio de uma teoria consequencialista que incorpora noções contratualistas e deontológicas, ele defende que esses agentes tenham os seus interesses considerados. Foi sugerido que os conceitos 'agente' e 'interesse' devem ser definidos, a exemplo do modo como Railton os define, sem assumir visões problemáticas como a forma de especismo que desconsidera de antemão qualquer agente ou interesse não-humano.

Diante de todo o exposto, considero pertinente conceber a ética da Inteligência Artificial como um domínio de investigação interdisciplinar em que as questões éticas exigem respostas complexas que refletem a dificuldade do pensar sobre a sociedade contemporânea. O papel da ética pode ser situado no coração de aspectos técnicos, científicos, filosóficos, legais e políticos que necessitam ser considerados em conjunto. Participam desse diálogo interdisciplinar com a ética, áreas como a ciência da computação, a biologia, a neurociência, a história, a psicologia, a sociologia, a filosofia da linguagem, a filosofia política e o Direito. Nesse contexto, cabe à ética investigar, de modo central, quais são as soluções para atenuar, ou mesmo prevenir, os eventuais efeitos nocivos das tecnologias de Inteligência Artificial. Além disso, cabe à ética investigar

quais são as respostas adequadas diante das mudanças estruturais implicadas por essas tecnologias. Railton mostra que tais soluções e respostas dependem, em última instância, da reflexão filosófica sobre a concepção de sociedade a ser promovida. Pode ser acrescentado que se faz necessário, especialmente, que as soluções e respostas filosóficas alcançadas pela ética aplicada à Inteligência Artificial possam ser materializadas em leis e políticas públicas.

Referências

CAPPELEN, Herman; PLUNKETT, David. Introduction: A Guided Tour of Conceptual Engineering and Conceptual Ethics. *In*: BURGESS, Alexis; CAPPELEN, Herman; PLUNKETT, David (eds.) **Conceptual Engineering and Conceptual Ethics.** Oxford: Oxford University Press, 2020, p. 1-26.

DARWALL, Stephen; GIBBARD, Allan; RAILTON, Peter. **Metaética***: Algumas tendências*. Série Ethica. DALL'AGNOL, Darlei (org.). Tradução de Janyne Sattler. Editora da UFSC, 2013.

FOOT, Philippa. The Problem of Abortion and the Doctrine of Double Effect. *In*: FOOT, Philippa. **Virtues and Vices and Other Essays in Moral Philosophy**. Los Angeles: UCLA Press, 1978, p. 19-33.

LIAO, S. Matthew (ed.) **Ethics of Artificial Intelligence**. Oxford: Oxford University Press, 2020.

PARFIT, Derek. **On What Matters***: Volume Three*. Oxford: Oxford University Press, 2017.

PLUNKETT, David; SUNDELL, Tim. Disagreement and the Semantics of Normative and Evaluative Terms. **Philosophers' Imprint**, vol. 13, n. 23, 2013, p. 1-37.

NOZICK, Robert. **Anarchy, State, and Utopia.** New York: Basic Books, 1974.

RAILTON, Peter. **Facts, Values, and Norms***:* Essays Toward a Morality of Consequence. Cambridge University Press, 2003.

RAILTON, Peter. Ethical Learning, Natural and Artificial. *In*: LIAO, S. Matthew (ed.) **Ethics of artificial intelligence**. Oxford: Oxford University Press, 2020, p. 45-78.

SIDGWICK, Henry. **The Methods of Ethics**. 7th ed. New York: Dover, 1966.

THOMSON, Judith. Killing, Letting Die, and the Trolley Problem. **Monist**, vol. 59, n. 2, 1976, p. 204–217.

Recebido em: 19/06/2023
Aprovado em: 23/11/2023