OPEN ACCESS

PUCRS | ESCOLA DE HUMANIDADES

VERITAS (PORTO ALEGRE)
Revista de Filosofia da PUCRS
Veritas, Porto Alegre, v. 68, n. 1, p. 1-17, jan.-dez. 2023
e-ISSN: 1984-6746 | ISSN-L: 0042-3955

http://dx.doi.org/10.15448/1984-6746.2023.1.44388

SEÇÃO: ÉTICA E FILOSOFIA POLÍTICA

Veículos Autônomos e Equilíbrio Reflexivo Amplo e Coletivo

Autonomous Vehicles and Collective and Wide Reflective Equilibrium

Vehículos Autónomos y Equilibrio Reflexivo Amplio y Colectivo

Denis Coitinho[1]
orcid.org/0000-0002-2592-5590
deniscs@unisinos.br

Recebido em: 13 fev. 2023.
Aprovado em: 21 ago. 2023.
Publicado em: 13 nov. 2023.

Resumo: O objetivo deste artigo é refletir sobre a necessidade de contarmos com padrões morais para orientar os veículos autônomos (VAs) e propor o procedimento do equilíbrio reflexivo (ER) para tal fim. Com isso em mente, inicio com uma investigação sobre o desacordo moral, para saber como devemos decidir, em casos de incerteza, argumentando que devemos fazer uso de um procedimento que congregue diferentes critérios normativos. Após, apresento uma rota interessante de investigação, que é o método de equilíbrio reflexivo coletivo na prática (CREP) como proposto por Savulescu, Gyngell e Kahane (2021), que corrige os resultados do experimento *Moral Machine* e propõe princípios de uma política pública para regular os VAs. O próximo passo é analisar o procedimento do ER, identificando suas características básicas de consistência, reflexividade, holismo e progressividade. Com isso, será possível, na sequência, apontar os limites do CREP, em razão de ele deixar de fora o critério normativo das virtudes e não formar um sistema coerente de crenças amplo o suficiente. Por fim, apresento a sugestão do equilíbrio reflexivo amplo e coletivo (ERAC), de forma a dar conta da pluralidade normativa que é base de nossa sociedade e propor uma metodologia para identificar o padrão moral para os VAs.

Palavras-Chave: Veículos autônomos, inteligência artificial, incerteza moral, normatividade, equilíbrio reflexivo.

Abstract: The aim of this paper is to reflect on the need to have moral standards to guide autonomous vehicles (AVs) and to propose a procedure of reflective equilibrium (RE) for this purpose. Bearing this in mind, I begin with an investigation into moral disagreement to find out how we should decide in cases of uncertainty, arguing that we should use a procedure that brings together different normative criteria. Afterwards, I present an interesting investigation route, which is the method of collective reflective equilibrium in practice (CREP) as proposed by Savulescu, Gyngell and Kahane (2021), which corrects the results of the Moral Machine experiment and proposes principles of public policy to regulate VAs. The next step is to analyze the RE procedure, identifying its basic characteristics of consistency, reflexivity, holism and progressiveness. Next, I point out the limits of CREP, because it leaves out the normative criterion of virtues and does not form a coherent system of beliefs that is wide enough. Finally, I present the suggestion of collective and wide reflective equilibrium (CWRE) in order to consider the normative plurality that is the basis of our society and propose a methodology to identify the moral standard for VAs.

Keywords: Autonomous vehicles, artificial intelligence, moral uncertainty, normativity, reflective equilibrium.

Resumen: El objetivo de este artículo es reflexionar sobre la necesidad de contar con estándares morales para guiar a los vehículos autónomos (VAs) y proponer un procedimiento de equilibrio reflexivo (ER) para tal fin. Teniendo esto en cuenta, comienzo con una investigación sobre el desacuerdo moral para saber cómo debemos decidir en casos de incertidumbre, argumentando que debemos utilizar un procedimiento que reúna diferentes criterios normativos. Posteriormente, presento una interesante ruta de investigación, que es el método de equilibrio reflexivo colectivo en la práctica (CREP) propuesto por Savulescu, Gyngell y Kahane (2021), que corrige los resultados del experimento de la Máquina Moral

[1] Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil.

y propone principios de política pública para regular los VAs. El siguiente paso es analizar el procedimiento del RE, identificando sus características básicas de consistencia, reflexividad, holismo y progresividad. A continuación señalo los límites del CREP, porque deja de lado el criterio normativo de las virtudes y no conforma un sistema coherente de creencias que sea lo suficientemente amplio. Finalmente, presento la sugerencia de equilibrio reflexivo amplio y colectivo (ERAC) para considerar la pluralidad normativa que es la base de nuestra sociedad y propongo una metodología para identificar el estándar moral para los VAs.

Palabras clave: Vehículos autónomos, inteligencia artificial, incertidumbre moral, normatividad, equilibrio reflexivo.

I

A tecnologia orientada por inteligência artificial (IA)[2] já é uma constante em nossas vidas – seja pelas recorrentes recomendações de músicas e filmes que recebemos pelos serviços de *streaming*, tal como Spotify e Netflix, respectivamente, seja pelos filtros de *e-mails* que usamos ou pela utilização de algum aplicativo de navegação por GPS, nas cidades e/ou nas estradas, como o Waze –, e ela não parece problemática em muitas áreas. Ao contrário, ela parece facilitar a nossa vida. Entretanto, essa tecnologia está se estendendo progressivamente para certos domínios, nos quais, possivelmente, terá um impacto maior, como decidir em circunstâncias de risco, estabelecer prioridades entre pessoas e fazer julgamentos complexos e, portanto, ter que tomar decisões que já se enquadram no domínio moral. Por isso, parece importante pensar sobre os algoritmos que alimentam esses produtos. Os carros autônomos, por exemplo, precisarão tomar decisões sobre como distribuir o risco entre os passageiros, pedestres e ciclistas, isto é, entre os que utilizam as vias públicas. As armas autônomas letais terão que identificar e selecionar os alvos humanos que serão eliminados. Por sua vez, algoritmos que já estão em uso, nos sistemas de saúde e judiciário, em alguns países, estabelecem a prioridade de quem receberá um transplante de órgão, bem como aconselham os juízes sobre quem deve obter liberdade condicional ou uma sentença maior de prisão (BONNEFON; SHARIF; RAHWAN, 2020).

A promessa da IA[3] é que ela simplifique e qualifique a nossa vida, a saber, melhorando as decisões humanas, evitando acidentes de trânsito, salvando vidas em uma guerra, otimizando o processo de doação e transplante de órgãos, até mesmo prevenindo crimes violentos etc., pois esses algoritmos podem, teoricamente, decidir sem vieses cognitivos ou ruídos, de forma lógica e racional, evitando qualquer parcialidade e discriminação. Entretanto, pode ser que isso não ocorra tal como esperado, uma vez que todas essas decisões referidas anteriormente, inevitavelmente, envolvem padrões éticos e avaliações morais complexas. Por exemplo, os carros autônomos devem sempre se esforçar para minimizar as baixas, mesmo que, às vezes, isso signifique sacrificar seus próprios passageiros para um bem maior? As armas autônomas letais devem sempre objetivar a vitória, mesmo ao custo de eliminar um alvo civil ou um soldado ferido, o que colocaria em risco a dignidade humana? As crianças devem sempre ter prioridade para transplantes de órgãos, mesmo quando um paciente mais velho é uma combinação genética melhor para um órgão disponível? Os algoritmos usados em tribunais devem sempre procurar reduzir a reincidência, mesmo que essa redução resulte em uma discriminação injusta para os réus negros, como tem acontecido atualmente com o programa COMPAS usado no sistema judiciário norte-americano?[4].

---

[2] IA se refere, essencialmente, ao uso de máquinas e *softwares* para realizar tarefas que caracteristicamente requerem a inteligência quando realizadas por humanos. É a capacidade de um sistema, como um *software* ou incorporado em um aparelho, de executar tarefas comumente associadas a seres inteligentes. Ver Frankenfield (2022). Ver, também, Copeland (2001).

[3] Existem dois tipos de IA. A IA forte envolve *software* que busca raciocinar e formar decisões cognitivas da maneira como as pessoas fzem; procura reproduzir, no mundo digital, os processos nos quais os cérebros humanos se envolvem quando deliberam e tomam decisões. A IA fraca, por sua vez, visa apenas fornecer assistência inteligente a atores humanos, exigindo somente que as máquinas sejam melhores em tomar decisões em alguns assuntos do que os humanos e que o façam de maneira eficaz dentro dos parâmetros definidos pelos humanos. Ver Rozenfield (2016) e Etzioni e Etzioni (2017, p. 410-411).

[4] O Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) é um programa usado em 46 estados do sistema judiciário norte-americano que tem o papel de prever quem reincidirá no crime, influenciando as decisões de liberdade condicional, fiança, condenações etc. O teste realizado pelo ProPublica demonstrou que o COMPAS está frequentemente errado, além de ser tendencioso contra os negros. A comparação mostrou que o programa tende a apontar erroneamente réus negros como futuros criminosos,

Veja que, no caso dos veículos autônomos (VAs)[5], em que pese a expectativa de menor estresse e maior segurança no trânsito, alguns problemas já começaram a aparecer, e se parece urgente discutir se eles devem ou não estar equipados com certos padrões morais e, se sim, quais seriam esses padrões. Em 2018, um carro da Uber – que está em teste – atropelou Elaine Herzberg, de 49 anos, enquanto ela conduzia uma bicicleta pela estrada em Tempa, Arizona. Os investigadores do caso disseram que o motorista de segurança do carro, Rafael Vasquez, estava assistindo a um programa de televisão e foi acusado de homicídio culposo. O problema é que o sistema automático do carro falhou em identificar a senhora Herzberg e sua bicicleta como um perigo eminente de colisão da forma que era esperado. Em 2019, um carro da Tesla, com sistema de piloto-automático, bateu em outro carro e matou duas pessoas no Texas. Segundo as autoridades policiais, o veículo estava em alta velocidade, e não tinha nenhum motorista de segurança no interior do carro. E, para dramatizar a urgência da discussão, é importante destacar que alguns dos maiores fabricantes de automóveis anunciaram planos de lançar carros inteiramente autônomos até 2024 e que já há VAs em circulação, sem a presença de nenhuma intervenção humana, atualmente. É o caso dos táxi-robôs que já estão circulando em Wuhan, na China, desde dezembro de 2022, os quais estão sendo operacionalizados pela empresa Baidu, bem como o caso dos táxi-robôs oferecidos pela Waymo, empresa ligada à Google, e pela Cruise, que já estão ofertando serviços em Phoenix e San Francisco[6].

Dado que esses veículos podem causar danos, alguns eticistas argumentam que os VAs precisam ser capazes de diferenciar as decisões certas das erradas. Em outras palavras, eles deveriam ter a capacidade de fazer um raciocínio moral[7]. De forma similar, muitos pesquisadores de IA defendem que, se essas máquinas podem tomar milhares de decisões cognitivas baseadas em informações por conta própria, como quando diminuir a velocidade, quando parar, quando ceder e assim por diante, elas também devem ser capazes de tomar decisões éticas. Como dito por Anderson e Anderson, "Idealmente, gostaríamos de confiar em máquinas autônomas para tomar decisões éticas corretas por conta própria, e isso requer que criemos uma ética para as máquinas" (ANDERSON; ANDERSON, 2011, p. 1). Assim, os VAs com IA parecem precisar de orientação ética, e isso porque eles tomam decisões sem intervenção humana, as quais podem ser danosas aos agentes[8].

E essa demanda ética parece se dar porque podem acontecer situações em que o VA terá que escolher entre dois males, isto é, entre desviar ou não de um pedestre para não feri-lo ou escolher se deve sempre salvar o número maior de vidas, da mesma forma que pode ser o caso em que terá que decidir sobre privilegiar as crianças em relação aos idosos ou as mulheres em relação aos homens, se seria correto atropelar um animal não humano para evitar um prejuízo material com a colisão do carro em uma barreira ou, ainda, se seria obrigatório atropelar um animal para salvar a vida das pessoas que estivessem no interior do veículo. Note-se que parece obrigatório, moralmente, que um VA sem passageiros, tal como

colocando-os na categoria de possíveis reincidentes quase duas vezes mais do que os réus brancos. Ver Larson *et al.* (2016). Ver, também, Mesa (2021).

5 VAs são sistemas complexos que utilizam o *deep learning* (aprendizagem profunda) e podem ser categorizados como de IA forte, a exemplo dos carros da Tesla e Waymo. Eles são programados para coletar informações, processá-las, tirar conclusões e mudar a maneira como se comportam, sem intervenção ou orientação humana. Ver Kaur e Rampersad (2018).

6 Para mais informações desses casos, ver as seguintes matérias: *Uber's self-driving operator charged our fatal crash* (BBC NEWS, 2020); *Two killed in driverless Tesla car crash, officials say* (PIETSCH, 2021); *China's Baidu operates taxi night in Wuhan* (ENGLISH NEWS, 2022); *Waymo expand service area in 2 cities* (WESSLING, 2022).

7 Wallach e Allen (2009) argumentam que esse raciocínio moral, para os agentes morais artificiais (AMAS), deve estar baseado em como nós tomamos decisões morais, e isso nos força a pensar profundamente sobre o funcionamento humano, especialmente no caso da deliberação moral, propondo um modelo de decisão ética que congregue os modelos principialísticos e intuitivos. Ver Wallach e Allen (2009, p. 3-12).

8 Anderson e Anderson (2011) defendem, corretamente, que devemos estabelecer um diálogo entre eticistas e especialistas de IA para encontrar um modelo apropriado e exequível de estabelecer padrões morais para as máquinas autônomas que usam IA. Ver Anderson e Anderson (2011).

um táxi-robô, desviasse de uma pessoa que seria atropelada ou mesmo que desviasse de um animal não humano, ao preço de o carro colidir com um poste de iluminação – desviar o carro implicaria danificar um bem móvel, é claro, mas salvaria uma vida. Ou, ainda, pode-se imaginar um cenário em que haveria a alternativa de desviar o carro para não atropelar e matar um pedestre, ao preço de bater em uma barreira, apenas machucando levemente o passageiro. Veja-se que essas são questões morais relevantes, porque refletem sobre valores muito importantes para nós, como a dignidade humana, o valor da vida, a igualdade, entre outros. Mas quem decidirá sobre isso?

A partir dos exemplos dados, pode-se perceber, facilmente, que o uso da IA em VAs levanta questões muito significativas e que exigem reflexão sobre a capacidade ou natureza da agência humana e artificial, responsabilidade moral e legal, censura e punição. Mesmo sendo essas questões muito importantes, o foco neste artigo será a respeito do padrão normativo que pode ser utilizado para orientar a decisão dos sistemas de IA nos VAs. E o problema específico é que, por mais que haja uma concordância geral de que esses carros devem dispor de padrões morais para tomar decisões, discordamos abertamente sobre o que faz uma ação ser certa ou errada, uma vez que o desacordo moral é um fenômeno social facilmente identificado, bem como discordamos sobre qual é a melhor teoria moral que deveria ser acessada para fundamentar esse tipo de decisão. Por exemplo, deveríamos usar um princípio de maximização do bem-estar de alguma teoria utilitarista, deveríamos utilizar o princípio deontológico de universalizabilidade ou, alternativamente, deveríamos consultar a ética das virtudes? Além do mais, discordamos, inclusive, se o melhor caminho seria mesmo fazer uso de uma teoria moral em casos como esses.

Muitos defendem que os VAs devem apenas seguir a legislação existente, tais como leis de trânsito, código penal, civil etc., sem a necessidade de inclusão de leis morais no sistema de IA[9]. Dada essa grande divergência, penso ser relevante compreender as características essenciais do ER, e, para tal, será necessário, anteriormente, refletir sobre a complexidade da normatividade, investigando os limites dos padrões morais tradicionais. De posse disso, minha estratégia será propor um tipo específico de ER como uma forma de raciocínio moral exequível para identificarmos os padrões normativos que podem orientar os VAs.

Dito isso, apresento a sequência que percorrerei no texto. Inicio com uma investigação sobre o desacordo moral, para saber como devemos decidir em casos de incerteza, argumentando que devemos fazer uso de um procedimento que congregue diferentes critérios normativos. Após, apresento uma rota interessante de investigação, que é o procedimento de equilíbrio reflexivo coletivo na prática (CREP), como proposto por Savulescu, Gyngell e Kahane (SAVULESCU; GYNGELL; KAHANE, 2021), que corrige os resultados do experimento *Moral Machine* (AWAD *et al*., 2018) e propõe princípios normativos para uma política pública, a fim de regular os VAs. O próximo passo será refletir sobre o ER, identificando suas características básicas de consistência, reflexividade, holismo e progressividade. Com isso, será possível apontar os limites do CREP, em razão de ele deixar de fora o critério normativo das virtudes e não formar um sistema coerente de crenças amplo o suficiente, o que parece trazer certos problemas. Por fim, apresento a sugestão do equilíbrio reflexivo amplo e coletivo (ERAC), de forma a dar conta da pluralidade normativa que é base de nossa sociedade e propor uma metodologia para lidar com a questão.

II

Um dos principais problemas para o desenvolvimento e a posterior aplicação de padrões morais nos VAs é que temos uma grande discordância moral, isto é, discordamos abertamente sobre qual ação é correta e qual ação é errada, bem como discordamos sobre o que seria uma

[9] Por exemplo, Etzioni e Etzioni (2017, p. 415).

vida boa e, inclusive, sobre qual padrão normativo-moral é o mais adequado para fundamentar as nossas decisões éticas. Questões sobre aborto, eutanásia, melhoramento humano, direitos dos animais, entre outras, são bons exemplos para reconhecermos o fenômeno do desacordo moral. A ação correta seria a que maximiza o bem-estar dos envolvidos ou, alternativamente, seria a ação que se deseja universal e que não instrumentaliza os agentes? Devemos sempre cumprir a promessa, mesmo ao custo de algum resultado indesejado, ou seria correto mentir para salvar a vida de alguém inocente que está sendo injustamente perseguido? Discordamos, também, sobre o que significa mesmo justiça. Seria a liberdade ou a igualdade? Ou seria uma combinação desses dois valores? Esses são pequenos exemplos de nosso desacordo moral. E, para além disso, discordamos sobre qual teoria moral deve ser acessada para orientar nossas decisões éticas. Seria o utilitarismo (de ato) melhor do que um modelo deontológico, como o kantismo, porque pode levar em conta os melhores resultados de uma ação, ou o kantismo teria superioridade ao princípio da maximização do bem-estar, porque isso poderia implicar casos de violações de direitos individuais? Há, por outro lado, quem defenda que nem o utilitarismo nem o kantismo seriam a melhor teoria moral, pois são modelos principialísticos que não estariam conectados com a estrutura emocional e disposicional do agente, e, por isso, deveríamos eleger a ética das virtudes, que toma a ação correta como a que seria aprovada por um agente virtuoso, o que coloca a questão de como devemos viver como anterior à pergunta de como devemos agir.

A respeito dessa discordância normativa, é importante reconhecer que, inclusive, temos disputas teóricas significativas sobre se devemos decidir a partir de princípios morais ou a partir das nossas próprias intuições sobre o certo e o errado, sobretudo quando se discute a respeito da necessidade de máquinas autônomas estarem equipadas com padrões éticos, de forma a tomá-las como agentes morais artificiais capazes de raciocínio moral. Muitos defendem que fazer um raciocínio moral seria aplicar certos princípios éticos no caso específico. Por exemplo, para saber se devemos aprovar a eutanásia voluntária, bastaria aplicar o princípio da maximização do bem-estar para ver se essa ação traz mais bem-estar aos envolvidos ou não. Ou bastaria que se aplicasse o princípio da universalizabilidade e não instrumentalização, para ver se essa eutanásia seria aprovada por uma regra que se deseja universal e que não toma o agente apenas como um meio. Por outro lado, muitos eticistas objetam que o principialismo é limitado por sua abstração e rigorismo, e, por isso, deveríamos adotar uma perspectiva que parta das intuições e decisões dos agentes e que busque um aperfeiçoamento, a partir dessa dimensão, tal como a ética das virtudes, de modo a levar em conta a dimensão das disposições e experiências dos agentes. Essa disputa é conhecida, na literatura, como o confronto dos modelos morais *top-down* (de cima para baixo) e *bottom-up* (de baixo para cima)[10].

Uma estratégia que parece mais frutífera para lidar com os VAs, a partir da incerteza moral, é adotar uma perspectiva pluralista. Em vez de termos que escolher entre uma das teorias tradicionais do debate, como a teoria utilitarista, a teoria deontológica e a ética das virtudes, podemos tentar uni-las em uma nova teoria normativa de primeira ordem. Assim, teríamos à disposição uma teoria moral híbrida, que utiliza vários critérios normativos distintos, mas coerentes entre si. Até porque a normatividade moral excludente parece insatisfatória para resolver problemas complexos, tais como os ocasionados pela IA. Problemas práticos complexos exigem, muitas vezes, critérios normativos diversos, para a identificação de soluções. Uma teoria moral tradicional usa, em geral, apenas um critério normativo, como a maximização do bem-estar,

[10] Essa discussão é muito comum no campo da ética da IA, de forma que eticistas e cientistas de IA se dividem a respeito de qual o melhor modelo para se adotar, com a finalidade de desenvolver agentes morais artificiais que podem decidir moralmente ou, no limite, estabelecer um raciocínio moral. Wallach e Allen, por exemplo, defendem que devemos usar um modelo que misture as abordagens *top-down* e *bottom-up,* e, para eles, a ética das virtudes é o modelo ético apropriado para tal tarefa. Ver Wallach e Allen (2009, p. 117-124).

considerando o utilitarismo, a universalizabilidade e a não instrumentalização, bem como considerando o kantismo ou mesmo algum agente virtuoso, como o prudente, e tendo em mente a ética das virtudes. O ponto é que, em muitas situações, são exigidos outros critérios, sendo necessário usar outro padrão normativo para resolver a questão, o que revelaria um problema de coerência interna da teoria. Como uma teoria utilitarista usaria um critério deontológico de liberdade sem se mostrar incoerente? Ou como uma teoria deontológica poderia usar um critério de maximização do bem-estar sem recair em contradição interna? Como a ética das virtudes poderia fazer uso de um princípio da dignidade humana, para assegurar os direitos humanos, sem comprometer a sua própria estrutura? Mesmo o utilitarismo de regra, que busca congregar o princípio da maximização do bem-estar, com uma regra universal que leva ao bem maior, enfrenta esse problema de coerência interna[11].

Para deixar mais clara essa limitação, deixem-me fazer referência aos problemas do *trolley*[12]. De forma geral, são apresentados dilemas morais para ver que intuições os agentes apresentam, com o objetivo de identificar o princípio moral correspondente. Um dos dilemas mais conhecidos questiona a correção de se sacrificar alguém para salvar mais vidas. Quando se pergunta às pessoas o que elas devem fazer ao ver um trem desgovernado, a 100 km/h, que irá atropelar cinco operários no trilho à frente, mas há uma possibilidade de acionar uma alavanca para desviar o trem, havendo apenas um operário nesse trilho alternativo, a resposta geralmente dada é: sim, deve-se desviar o trem. Por sua vez, se a pergunta é se devemos empurrar uma pessoa robusta de uma passarela para salvar os cinco operários nos trilhos do trem, a resposta geralmente dada é: não. A interessante questão é saber por que o princípio que parece certo, no primeiro caso, é tido como errado no segundo exemplo? O princípio no primeiro caso, claramente, aponta para a ideia de que é correto sacrificar uma pessoa para salvar cinco, isto é, de que devemos salvar o maior número possível de pessoas – e isso porque a vida humana importa, o que nos remete ao princípio da maximização do bem-estar. Por sua vez, o princípio, no segundo caso, mostra-nos que é errado matar um ser humano inocente, não importando as consequências, princípio esse que se choca com o primeiro. Isso seria semelhante à situação de se ver como errado matar uma pessoa saudável para transplantar seus órgãos para cinco doentes graves que precisam desse tipo de tratamento para sobreviver. Note que há um conflito dos princípios aqui. Por um lado, sabemos que é correto o esforço para salvar mais vidas. Ao mesmo tempo, sabemos que é errado matar um ser humano inocente.

Nesse contexto dos problemas do *trolley*, o utilitarismo, por exemplo, diria que é correto salvar o número maior de vidas, uma vez que seu princípio moral básico é o da maximização do bem-estar; assim, seria correto acionar a alavanca para desviar o trem e salvar os cinco operários, mesmo com o ônus da morte de um agente. Aqui, temos a aceitação do princípio do sacrifício. E, por esse mesmo princípio, também seria correto empurrar o homem robusto para salvar os cinco operários, bem como seria correto, teoricamente, fazer o transplante para salvar os cinco doentes. O problema é que, de acordo com nossos juízos morais comuns, é errado tirar a vida de um agente inocente, mesmo considerando as boas consequências. Por outro lado, o kantismo diria que é errado acionar a alavanca, para desviar o trem, visando salvar os cinco operários, visto que o princípio da não instrumentalização proíbe qualquer ação que considera uma pessoa apenas um meio para o bem das outras, trazendo, por consequência, a morte de cinco agentes. Isso pa-

---

[11] A principal objeção ao utilitarismo de regra é que ele seria incoerente, porque a teoria permite ações que não maximizam o bem-estar, embora a teoria esteja comprometida, tacitamente, com a maximização, e, assim, essa teoria não seria mais o utilitarismo. Ver Rajczi (2016, p. 857-860).

[12] Os *trolley problems* são uma série de experimentos mentais, em ética e psicologia, envolvendo diversos dilemas éticos, na forma de se é correto ou não sacrificar uma pessoa para salvar um número maior de agentes. O ponto central é fazer uso das intuições morais sobre certos casos particulares para se chegar a certos princípios éticos. Originalmente, o dilema colocado por Foot foi na figura de um condutor de um trem, e não a de um espectador. Sobre o tema, ver Foot (1978) e Thomson (1976).

rece problemático porque também concordamos que devemos salvar o número maior de pessoas, talvez porque pensamos que a dignidade humana tenha valor. E, mesmo que o utilitarismo e o kantismo quisessem valorizar esses juízos morais comuns, eles teriam um problema de coerência interna na própria teoria.

Por essa razão, seria desejável poder contar com uma teoria moral híbrida que dispusesse de vários critérios normativos, como o critério utilitarista das melhores consequências, o critério deontológico da não instrumentalização, que garante o respeito universal à dignidade humana, e, até mesmo, um conjunto de virtudes que são padrões normativos não da ação correta, mas da vida boa, o que parece contribuir no caso de precisarmos contar com uma compreensão mais abrangente do certo e errado, alinhando, de forma coerente, esse conjunto plurinormativo. E, se isso ainda não está disponível para nós, penso que podemos tentar ao menos usar uma estratégia ou metodologia para conectar esses vários critérios normativos coerentemente.

A título de exemplo, Kyle Bogosian (2017), no artigo *Implementations of moral uncertainty in intelligente machines*, defende uma estratégia plurinormativa, de modo que, para resolver o problema do desacordo moral, devemos levar em consideração todas as teorias éticas, no momento da decisão, e buscar por ações que mostrem o maior valor entre elas, em vez de buscar selecionar uma teoria moral e ignorar as outras – até porque, dado que estamos tão incertos quanto aos padrões morais, é altamente improvável que um sistema moral particular esteja inteiramente correto. O argumento é o seguinte: enquanto discordamos uns dos outros sobre ética, ainda devemos concordar em construir máquinas morais baseadas na incerteza, mesmo que rejeitemos a ideia como um guia para o comportamento humano. Para tal, ele usa o modelo defendido por MacAskill (2016), de "incerteza normativa", que argumenta que devemos fazer juízos, para guiar a ação, baseados em todas as teorias em que o agente tem algum nível de confiança, decidindo a partir da maximização do valor da escolha-valiosa esperada e permitindo a agregação e comparação interteórica de diferentes estruturas normativas[13]. O aspecto positivo dessa abordagem é que um agente pode tomar decisões prudentes que objetivem subsidiar a ação, de acordo com os vários valores atribuídos pelas pessoas, podendo evitar implicações contraintuitivas (BOGOSIAN, 2017).

Mesmo julgando interessante essa estratégia proposta por Bogosian, neste artigo quero testar o alcance da aplicação do método do ER para identificar padrões morais para os VAs, tendo em mente a importância da plurinormatividade, uma vez que se trata de um procedimento utilizado de forma usual para justificar princípios morais, através de um ajuste mútuo com os juízos morais (ponderados), em que se tem grande confiança, e, também, com os juízos factuais, apoiados por certas teorias que sejam relevantes no caso – por exemplo, justificando princípios de justiça, a partir de sua coerência com um sistema coerente de crenças, o que conduz a um equilíbrio reflexivo amplo (ERA)[14]. E esse procedimento parece promissor, nesse contexto de incerteza normativo-moral, exatamente porque ele pode ser interpretado como um método que nos auxilia a saber o que devemos fazer quando estamos em dúvida, sendo o fim de um processo deliberativo, em que se pesam razões e se escolhe um certo curso de ação a respeito de um problema particular (KUSHNER; BELLIOTTI; BUCKNER, 1991), podendo ser usado, também, como um método para subsidiar discussões e justificar tomadas de decisões coletivas e não exclusivamente pessoais

---

[13] Bogasian (2017) define a abordagem de MacAskill como uma teoria metanormativa, caracterizada da seguinte forma: o agente investiga uma situação que exige uma decisão, a partir de <S, t, A, T, C>, em que S é uma decisão tomada, t é o tempo, e A é o conjunto de possíveis ações a serem tomadas. T é o conjunto de teorias normativas a serem consideradas, no qual a teoria T_i é uma função da situação de decisão, que produz um escore cardinal e ordinal de escolhas-valiosas para a ação CW_i (A) para todas as ações a ∈ A. $C(T_i)$ é uma função de credibilidade ou aceitação (*credence*), atribuindo valores em [0, 1] para cada $T_i \in T$. Uma teoria metanormativa é uma função de situações de decisão que produz um ordenamento das ações em A em termos de sua adequação (*appropriateness*). Ver Bogosian (2017, p. 598). Ver, também, MacAskill (2016, p. 969).

[14] Ver Rawls (1971), Daniels (1979) e Scanlon (2003).

(BRANDSTEDT; BRÄNNMARCK, 2020)[15].

Com isso em mente, no restante do artigo, analiso as características do ER, iniciando com o uso específico do método que é feito por Savulescu, Gyngell e Kahane (2021) (CREP), na discussão sobre quais princípios deveriam ser escolhidos como padrão moral para os VAs. Após, investigo o procedimento como um tipo específico de raciocínio moral que se caracteriza pela consistência, pela reflexividade, pelo holismo e pela progressividade. Por fim, sugiro uma reformulação no CREP, de modo, sobretudo, a incorporar o padrão normativo das virtudes e a formar um amplo sistema coerente de crenças.

III

Em recente artigo, Savulescu, Gyngell e Kahane (2021) defenderam o método do ER, de forma a usá-lo para acomodar a preferência pública sobre como os VAs devem decidir, em situações de emergência, em tensionamento com os princípios de certas teorias éticas, como o utilitarismo, o deontologismo kantiano e o contratualismo rawlsiano. Eles chamaram esse método de "equilíbrio reflexivo coletivo na prática" – em inglês, *Collective Reflective Equilibrium in Practice* (CREP). Em CREP, os dados coletados e tabulados sobre as preferências públicas, ao redor do mundo, com a utilização da plataforma *on-line Moral Machine*, devem servir apenas como *input* em um processo deliberativo público que busca pela coerência entre as atitudes, os comportamentos e os princípios éticos, e não como a última palavra sobre a questão, uma vez que as intuições morais podem expressar preconceitos, vieses e interesses pessoais e/ou corporativos (SAVULESCU; GYNGELL; KAHANE, 2021). A ideia é ver se os juízos ponderados sobrevivem se confrontados aos princípios éticos do utilitarismo, kantismo e contratualismo. E, em caso de inconsistência, devemos revisar nossas crenças iniciais, nas quais temos grande confiança, para alcançar uma situação de reflexão adequada. De posse disso, eles propõem uma política pública para normatizar essa situação (SAVULESCU; GYNGELL; KAHANE, 2021). Mas, antes de detalhar o procedimento de CREP, deixem-me circunscrever o estudo *Moral Machine*.

O experimento *Moral Machine* – conduzido por pesquisadores do Massachusetts Institute of Technology (MIT) – consistiu em uma pesquisa, via plataforma *on-line*, quanto às preferências das pessoas, ao redor do mundo, sobre como os VAs deveriam decidir em situações de emergência, em que a morte de alguém é eminente. Por exemplo, um carro sem freios deveria desviar de sua rota para não atropelar cinco pessoas ao custo de atropelar uma? Ele deveria estabelecer preferência entre seres humanos e animais? Ou, ainda, ele deveria privilegiar os que cumprem as leis de trânsito ou privilegiar os mais saudáveis, em contraposição aos que têm sobrepeso? Assim, foram coletadas 40 milhões de decisões em 10 línguas e em 233 países. A partir da tabulação dos dados, foram identificadas nove preferências públicas, em ordem decrescente de confiança: (1) salvar vidas humanas; (2) salvar mais vidas; (3) salvar os mais jovens; (4) salvar os que cumprem as leis de trânsito; (5) salvar as pessoas de *status* social mais elevado; (6) salvar os mais magros, em contraposição aos que têm sobrepeso; (7) salvar as mulheres, em contraposição aos homens; (8) salvar os pedestres, em contraposição aos passageiros; (9) optar que o veículo continue seu movimento. Dessas nove, três preferências públicas mais fortes foram identificadas, a saber: (i) salvar vidas humanas; (ii) salvar mais vidas; e (iii) salvar os mais jovens (AWAD *et al.*, 2018).

Nas palavras dos responsáveis pelo experimento *Moral Machine:*

> Como mostrado na Fig. 2a, as preferências mais fortes são observadas para poupar humanos em relação aos animais, poupar mais

[15] Kushner, Belliotti e Buckner (1991) se referem ao ER como uma abordagem sistemática para tratar de um dilema ético. Nessa compreensão, ele é um método de tomada de decisão que pode ser usado para ajustar os princípios e juízos existentes a respeito de um problema particular, em vista de obter uma decisão justificada (KUSHNER; BELLIOTTI; BUCKNER, 1991). Brandsted e Brännmark (2020, p. 357), por sua vez, sugerem uma versão do ER que pode ser tomado como "uma ferramenta para um raciocínio público sobre problemas práticos com o objetivo de encontrar soluções compartilhadas". Para mais detalhes do método do ER e de seus diferentes usos na filosofia moral e política e na epistemologia, ver Rechnitzer (2022, p. 1-9).

> vidas e poupar vidas jovens. Assim, essas três preferências podem ser consideradas blocos de construção essenciais para a ética das máquinas ou, pelo menos, tópicos essenciais a serem considerados por legisladores. De fato, essas três preferências diferem fortemente do nível de controvérsia que comumente suscita entre eticistas (AWAD *et al*., 2018, p. 60).

Para Awad *et al*. (2018), assim, essas preferências públicas, demonstradas pelo experimento *Moral Machine*, podem servir como um modelo para o desenvolvimento de uma ética para máquinas com IA ou como um padrão normativo relevante, que pode ser levado em conta pelos legisladores quando estiverem trabalhando em uma agenda para regular os VAs[16].

Para Savulescu, Gyngell e Kahane (2021), é muito relevante podermos contar com esses dados da preferência coletiva, para se estabelecer uma política pública, a fim de normatizar os VAs por duas razões centrais. A primeira é que não contamos com verdades morais, disponíveis ao nosso conhecimento, que nos mostre o que seria o certo e o errado em tais circunstâncias, considerando, inclusive, que as teorias morais mais conhecidas nos dão respostas diferentes a esses dilemas. Por exemplo, para o utilitarismo, seria correto salvar mais vidas, uma vez que seu critério normativo é o da maximização do bem-estar; entretanto, para o kantismo, isso seria errado, pois feriria o princípio da dignidade humana. A segunda razão é que, em democracias liberais, a legitimidade das leis depende do apoio público, sobretudo quando se trata de casos complexos em que há forte desacordo. Para os pesquisadores, "a questão é, então, como integrar tal evidência sobre intuições tão abrangentes na reflexão ética, e especialmente nas decisões sobre a regulação de novas tecnologias" (SAVULESCU; GYNGELL; KAHANE, 2021, p. 5).

Os autores em tela observam corretamente que, embora muito relevante, esses dados podem expressar preconceitos, viés de tribalismo e interesses pessoais e/ou corporativos. Por exemplo, eles ressaltam que essas preferências não são igualmente fortes em todas as três regiões pesquisadas (ocidente, oriente e sul). Nos países da região oriental, por exemplo, a preferência por jovens é mais fraca do que a preferência pelos que cumprem as regras de trânsito. Na região sul, o que inclui países da América Central e América do Sul, por outro lado, a preferência por pessoas de maior *status* social é quase tão forte quanto a preferência pelos jovens. A crítica é que, se utilizarmos essas preferências, precisamos decidir se vamos escolher as preferências mais fortes ou apenas as preferências mais fortes em uma região específica do mundo. Mais importante, algumas preferências não devem ser usadas como base de uma política pública, pois muitas delas são tribalistas, podendo estabelecer prioridade aos compatriotas em relação aos estrangeiros ou aos parentes em relação aos estranhos; ainda, muitas delas discriminam com base na etnia, na classe social ou no gênero. Então, o que fazer? Como muitas preferências podem expressar distorções, elas devem ser justificadas. Assim, o ER é usado para tal fim[17].

Savulescu, Gyngell e Kahane (2021), portanto, adotam o método de ER tal como formulado por Rawls, adaptando-o para um ER coletivo na prática (CREP), entendido como um processo deliberativo em que tanto as intuições públicas como as teorias éticas serão igualmente importantes para se poder identificar o consenso através de uma reflexão adequada[18]. Dessa forma, inicia-se com o *input* fornecido pelas intuições públicas, que são as preferências mostradas pelo estudo *Moral*

[16] Para uma crítica a essa proposta, ver o artigo de John Harris (2020), *The Immoral Machine*, em que ele critica, sobretudo, a tentativa de identificar uma moralidade pública para orientar os VAs.

[17] Os autores ponderam que os participantes do estudo conduzido por Awad *et al*. (2018) não foram perguntados se estabeleceriam preferências em salvar membros de seu próprio grupo étnico ou em salvar compatriotas, ao invés de estrangeiros, da mesma forma que não foram perguntados se estabeleceriam alguma preferência em salvar a sua própria vida ou a de sua família. O problema é que outros estudos realizados em menor escala relevam essas preferências tribalistas e autointeressadas, priorizando os compatriotas e parentes, bem como expressando discriminação étnica, social e de gênero. Ver Savulescu, Gyngell e Kahane (2021, p. 3-4).

[18] Em CREP: (i) os deliberadores são políticos, eticistas, juristas, cientistas, médicos e cidadãos em geral; (ii) as intuições iniciais (*input*) são preferências do público; (iii) a referência teórica é dada pelos valores e princípios contidos nas teorias éticas tradicionais; (iv) o resultado (*output*) é alcançado em uma política pública justificada em um processo democrático; e (v) a iteração ou repetição (*iteration*) é limitada. O ponto central é entender o procedimento como um processo deliberativo público realizado por um agente coletivo, e não como uma deliberação individual realizada por uma pessoa privada. Ver Savulescu, Gyngell e Kahane (2021, p. 6).

Machine, e, em seguida, filtram-se essas intuições para selecionar os juízos ponderados, retirando as intuições que são frutos de preconceitos e distorções. Em CREP, os juízes competentes não estão usando suas próprias intuições e devem aplicar determinados critérios para verificar a razoabilidade dessas preferências. Eles selecionam três preferências que poderiam ser tomadas como juízos ponderados: (i) salvar mais vidas; (ii) salvar os jovens; e (iii) salvar as mulheres. Em uma segunda etapa, busca-se a coerência entre esses juízos ponderados e os princípios morais de três teorias éticas que têm uma grande aceitação das pessoas, a saber: utilitarismo, kantismo e contratualismo. Levando isso em consideração, "(...) precisamos verificar se estas intuições estão realmente respondendo a razões eticamente plausíveis, o que aumentaria nossa confiança em sua validade. Afinal, mesmo os juízos ponderados abrangentes podem estar errados" (SAVULESCU; GYNGELL; KAHANE, 2021, p. 8).

Por exemplo, o kantismo, com o princípio da dignidade humana, aprovaria a preferência de salvar vidas humanas, mas proibiria a preferência de salvar mais vidas, uma vez que toda vida humana tem a mesma dignidade, e, pela mesma razão, proibiria a preferência aos jovens e às mulheres. O utilitarismo, por sua vez, a partir do princípio da maximização do bem-estar, aprovaria a preferência de salvar vidas humanas, em razão do critério preferencial da vida que seria mais útil, bem como aprovaria as preferências de salvar mais vidas e salvar os jovens, no entanto rejeitaria a preferência às mulheres, visto que isso não parece maximizar o bem-estar. De forma similar, o contratualismo, com o critério da imparcialidade fornecido pelo véu da ignorância, aprovaria as preferências de salvar vidas humanas, salvar mais vidas e salvar os jovens, mas rejeitaria, também, a atitude preferencial às mulheres, pois, recoberto pelo véu da ignorância, não se saberia a qual gênero se pertence.

Com este trabalho realizado, de confrontar as intuições públicas com os princípios das teorias éticas mais influentes no debate, a terceira etapa seria propor uma política pública justificada, que teria por critério normativo salvar, preferencialmente, os seres humanos e salvar o maior número de pessoas. A preferência em salvar os jovens poderia ser objeto de posterior deliberação, pois demandaria pesquisa empírica. A preferência pelas mulheres seria rejeitada, o que significaria realizar uma revisão nas convicções morais iniciais. Com isso, identifica-se, claramente, que: "A contribuição do CREP é esclarecer o papel dos dados sobre as preferências públicas e das restrições impostas pelas teorias éticas nesse processo" (SAVULESCU; GYNGELL; KAHANE, 2021, p. 9).

Após essa breve apresentação do CREP, é importante esclarecer por que eu acredito que essa seja uma rota interessante de investigação. Em primeiro lugar, porque não persegue a ideia de querer encontrar um princípio moral universal para ser usado em qualquer caso, de forma a tomar esse princípio como verdadeiro, apostando em uma metodologia de baixo para cima, que parte das intuições compartilhadas das pessoas, ao redor do mundo, para testá-las por sua coerência com um conjunto normativo plural. Como não está disponível para nós identificarmos os princípios morais verdadeiros, no sentido representacionista, um bom ponto de partida é tomarmos os juízos morais comuns, em que temos grande confiança, e testá-los por sua coerência com princípios éticos que possuem uma grande aceitação na comunidade. Em segundo lugar, essa estratégia parece interessante porque interpreta o ER como uma decisão coletiva, e não como uma deliberação particular de um agente privado, como é usual. Como estamos tentando identificar padrões normativos para os VAs que estarão em uso no mundo inteiro, é relevante que as convicções de todos sejam levadas em conta. E, assim, o final do processo deliberativo é uma decisão pública, que pode servir de modelo para uma futura legislação. Isso também é importante do ponto de vista da legitimidade política, pois, em democracias liberais, as políticas públicas dependem da aprovação dos cidadãos, ao menos em um certo nível, evitando que essa importante decisão seja tomada apenas por *ex-*

perts ou tecnocratas ou mesmo seja orientada exclusivamente pela lógica do mercado. Por fim, a terceira razão é que ela usa diferentes padrões éticos para testar as intuições morais compartilhadas, utilizando as teorias morais que são relevantes no debate. Isso é importante porque, em um contexto de incerteza normativa-moral, é mais prudente apostar na plurinormatividade.

Mesmo considerando que essa rota representa um grande avanço, penso que ela tem algumas limitações. Por exemplo, o CREP não usa o critério normativo das virtudes, que é muito relevante tanto na linguagem cotidiana de censura e elogio como na história da filosofia moral. Por isso, acredito que a ética das virtudes deveria substituir o contratualismo como um terceiro padrão normativo-moral que servirá de teste às intuições compartilhadas[19]. Também, ele não esclarece, satisfatoriamente, como se dá a filtragem das intuições compartilhadas para transformá-las em juízos ponderados. Sabemos apenas que os deliberadores são políticos, eticistas, juristas, cientistas, médicos e cidadãos em geral, mas não há um detalhamento de como essa filtragem ocorre, e isso pode introduzir uma arbitrariedade certamente indesejada. Por exemplo, no experimento *Moral Machine*, das nove preferências públicas levantadas, foram selecionadas as três que tiveram maior aprovação, a saber: salvar vidas humanas, salvar mais vidas e salvar os mais jovens. Porém, com o uso do CREP, foram selecionadas outras preferências, que são: salvar mais vidas, salvar os jovens e salvar as mulheres – e em nenhum momento os autores explicaram como se deu essa escolha. Uma terceira limitação, penso, é que esse ER parece ser estreito (*narrow*) e não amplo (*wide*), estabelecendo a coerência apenas entre os juízos ponderados e os princípios éticos. Com um ERA, por sua vez, pode-se testar o *input* inicial e os princípios éticos com os fatos do mundo, isto é, com os juízos factuais de certas teorias científicas, formando um sistema coerente de crenças mais amplo.

Com isso em mente, vou procurar esclarecer, brevemente, os elementos centrais do ER, como um tipo específico de raciocínio moral, na próxima seção, para, por fim, apresentar o ERAC como uma alternativa ao CREP.

IV

O ER segue as linhas gerais do raciocínio moral por consistência. Esse raciocínio se caracteriza por partir das intuições morais dos agentes e buscar estabelecer uma rede mais ampla de conexões, exigindo que se trate igualmente os casos iguais, o que implica expor as inconsistências entre os juízos morais e os casos concretos, contrapondo-se ao modelo dedutivista, que aplica um princípio diretamente a um caso concreto. A semelhança se dá porque o ER é um procedimento que exige a coerência ou consistência lógica no raciocínio, exigindo que as crenças morais ponderadas sejam coerentes com certos princípios morais e, também, com certas teorias morais e não morais relevantes ao caso investigado[20]. Claro que ele é um procedimento formal, servindo para justificar os princípios ou mesmo uma tomada de decisão, mas segue a mesma linha do raciocínio por consistência, objetivando encaixar e ajustar os juízos morais ponderados, em ordem de alcançar a coerência deles com as premissas de nossos maiores comprometimentos morais em geral. Suas características centrais são a coerência ou a consistência, a reflexividade, o holismo e a progressividade[21].

A característica central do ER é a coerência, de forma similar ao raciocínio por consistência, uma vez que ambos os aspectos lidam com a

---

[19] A exclusão do contratualismo rawlsiano e a substituição pela ética das virtudes se dão porque o contratualismo também é um modelo deontológico, tal como o kantiano, e o padrão normativo das virtudes não está sendo considerado em CREP. Como a ideia é que o procedimento de ER seja plurinormativo, o tensionamento dos juízos ponderados com as teorias éticas deve se dar pelos critérios oferecidos pelas três teorias morais mais tradicionais no debate, a saber: a maximização do bem-estar; a universalizabilidade e não instrumentalização; e o padrão do agente virtuoso.

[20] Harman, Manson e Sinnott-Armstrong (2010) classificam os dois tipos de raciocínio moral como modelo dedutivo e equilíbrio reflexivo. Eles explicam o equilíbrio reflexivo como um tipo de raciocínio alternativo ao modelo dedutivo, que parte das intuições morais dos agentes e busca estabelecer uma rede mais ampla de conexões. Já Campbell e Kumar (2012), no artigo *Moral reasoning on the ground*, classificam o modelo alternativo como de raciocinio moral por consistência (*consistency moral reasoning*).

[21] Para uma aproximação entre o ER e o raciocínio moral por consistência, ver Campbell (2014).

inconsistência moral. O ER tem relação com a inconsistência entre um juízo moral particular e um princípio ético geral, enquanto o raciocínio por consistência tem relação com a inconsistência entre dois ou mais juízos morais particulares – mas, ainda assim, em ambos os casos, a inconsistência é uma razão justificável para resolver o conflito. A justificação para resolver cada conflito, rejeitando um juízo moral ou princípio em favor de outro, é relativa às normas de fundo implícitas que são justificadas no procedimento. Com isso, pode-se perceber que a função básica do ER é resolver o conflito entre nossos juízos morais intuitivos, e isso será possível quando juízos e princípios formarem um amplo sistema coerente de crenças, o que incluirá as normas e os juízos factuais de teorias científicas relevantes, formando um *ERA*. Em ERA, a justificação é alcançada quando a inconsistência no sistema de crenças é eliminada (CAMPBELL; KUMAR, 2012).

Uma segunda característica muito importante é a reflexividade. Para não ser o caso de se justificar uma dada intuição moral apenas por sua coerência com um princípio ético, o que poderia implicar conservadorismo, é importante que o procedimento obtenha informação factual, e, assim, os juízos ponderados e princípios são testados por sua coerência com os juízos factuais. Mas não é suficiente que os fatos sejam reconhecidos, sendo necessário refletir sobre eles para entender corretamente seu significado. Isso deve servir para apontar nossos preconceitos, de forma que nossas suposições estão sempre abertas ao exame e à reflexão, e nada é visto como não revisável. Aliás, pode-se dizer que a revisibilidade é a característica básica do ER. Lembrando o nosso exemplo anterior de uma discussão sobre os direitos dos homossexuais, refletir cuidadosamente sobre alguns fatos – tal como o fato médico de a orientação sexual dos agentes não ser opcional ou do fato social de que o casamento tem mais a ver com o cuidado do que com a procriação – pode servir de razão para uma revisão da crença. Isso já explica a característica de holismo do ER, uma vez que seu objetivo básico é conectar valores e fatos, de maneira ampla, bem como considerar a pluralidade de valores morais e não morais de uma sociedade democrática, pensando a justificação de princípios ou tomada de decisão interpessoalmente[22].

Uma última característica que gostaria de destacar, associando o ER ao raciocínio por consistência, é a sua progressividade, isto é, que ele ocorre de maneira lenta e envolvendo toda a sociedade, o que possibilita o progresso moral. Esse progresso ou essa mudança moral nada mais é do que a correção das distorções de nossos juízos éticos, as quais impedem que se avalie de forma equitativa o estatuto moral dos agentes, significando uma maior inclusividade ética. De um ponto de vista histórico de longa duração, isso permitiu que se incluíssem, no círculo moral do cuidado e da equidade, os outros povos, as mulheres, os escravos e, até mesmo, os animais não humanos, por exemplo. A condenação da escravidão, da intolerância religiosa e do sexismo é um exemplo desse tipo de progresso, bem como a afirmação universal dos direitos humanos e a preocupação com o bem-estar dos animais. Assim, podemos entender o ER como um procedimento que ajuda a superar o tribalismo, pois conduz a uma correção das inconsistências dos juízos morais que direcionamos aos diferentes agentes na sociedade[23].

V

A partir das características destacadas na seção anterior, passo, agora, a apresentar um procedimento alternativo que corrige algumas limitações do CREP, incluindo o critério norma-

[22] Kushner, Belliotti e Buckner (1991) apontam quatro características centrais da metodologia do ER: (i) deve obter informação factual para testar os juízos ponderados e os princípios éticos; (ii) deve refletir sobre esses fatos para entender seu significado; (iii) deve proceder como na ciência, em que as suposições estão sempre abertas ao exame e à reflexão; e (iv) deve evitar erros intelectuais, como argumentos não sólidos, analogias erradas ou generalizações. Ver Kushner, Belliotti e Buckner (1991, p. 289-292).

[23] Pode-se entender o progresso moral como uma melhora deliberativa em certas práticas sociais, com um impacto positivo no mundo, que se relaciona de diferentes maneiras com as capacidades teóricas e práticas dos agentes. Como um fenômeno social, esse progresso implica a observação de uma melhora normativa-ética geral – por exemplo, com o fim da instituição da escravidão ou mesmo com a criação da Declaração Universal dos Direitos Humanos. Ver Albersmeir (2022, p. 229-234). Ver, também, Buchanan e Powell (2018).

tivo das virtudes, contando com a coerência a certos fatos e estabelecendo um ERAC. Veja que a força do ERAC, nesse caso específico dos VAs, é que, mesmo havendo forte desacordo moral, assim como havendo um desacordo sobre qual teoria ética deveria ser privilegiada como padrão normativo, chegar em um sistema coerente de crenças – entre nossos juízos morais ponderados, com certos princípios éticos aceitos pela tradição e, ainda, com certas crenças de teorias científicas válidas em uma reflexão adequada – é já apresentar uma justificação pelo respaldo que se obtém do "apoio mútuo de muitas considerações, de tudo se encaixando em uma visão coerente" (RAWLS, 1971, p. 21).

A ideia básica é pensar o ERAC como um método de tomada de decisão em nível global, que é coletivo e público por partir das intuições morais compartilhadas, que foram identificadas pelo experimento *Moral Machine*; em seguida, deve-se filtrá-las, para que sejam consideradas juízos ponderados, e, depois, testar esses juízos ponderados, por sua coerência com os princípios éticos das principais teorias que mais influenciam o debate moral (maximização do bem-estar, imperativo categórico e agente virtuoso) e, ainda, por sua coerência com juízos factuais das teorias de fundo importantes para o caso. Assim, ele pode ser tomado como uma destacada ferramenta para o raciocínio público, ao lidar com o problema de qual padrão moral deve orientar os VAs, e isso muito em razão de sua plurinormatividade.

Para além da coerência entre os juízos ponderados, os princípios éticos e as crenças factuais relevantes, o ERAC contará com um critério para tomar os juízos ponderados com uma credibilidade inicial, sendo independente da coerência, bem como o sistema coerente de crenças que é formado deverá possuir certas características normativas, tais como as virtudes teóricas de consistência, simplicidade, precisão, fertilidade, entre outras. Portanto, o objetivo é que o resultado possa orientar uma política pública justificada em um processo democrático de decisão[24].

Inicio, então, com a primeira etapa, identificando o *input* e filtrando os juízos ponderados. Da mesma forma que em CREP, parte-se das preferências do público que foram identificadas pelo experimento *Moral Machine*, contando-as como as intuições morais iniciais. Recordando, o experimento identificou nove preferências públicas globais, sendo que três dessas preferências tiveram maior adesão, a saber: salvar vidas humanas, salvar mais vidas e salvar os jovens. Ainda, similarmente ao CREP, os deliberadores são políticos, eticistas, cientistas de AI, juristas, médicos e cidadãos em geral. Assim, a ideia é que os deliberadores sejam especialistas no assunto em tela, podendo contar com sua perícia e experiência para tratar desse tema complexo, bem como que eles tenham certas virtudes, como razoabilidade, imparcialidade e mente aberta.

Então, o próximo passo dos deliberadores é analisar se as preferências globais, elencadas pelo experimento *Moral Machine*, podem contar como juízos ponderados, isto é, como juízos morais confiáveis. No entanto, diferentemente de CREP, os juízos ponderados não são assegurados só pela confiança dos agentes, mas sim pelas próprias qualidades epistêmicas e morais dos deliberadores, de forma a contar com uma credibilidade inicial, que será independente da coerência. Seguindo o modelo sugerido por Beauchamp e Childress (2013), devemos contar com "juízes morais" que possuem virtudes epistêmicas e morais relevantes, tais como ser imparcial ou ter simpatia e compaixão pelo bem-estar dos outros. Isso parece importante, porque uma das principais críticas feitas ao ER é sobre a falta de credibilidade inicial das crenças. A objeção usual é que o método teria uma fraqueza epistemológica, porque ele quer justificar a crença moral por sua coerência com os princípios éticos e outras crenças, mas as crenças iniciais são asseguradas apenas por algo vago, como a confiança do

[24] Em recente livro, Tanja Rechnitzer (2022) apresenta, detalhadamente, um caso de estudo para aplicação do método do ER, a fim de justificar o princípio da precaução, em casos de incerteza do dano, embora, em sua proposta, a deliberação seja pessoal e não coletiva. Sobre a característica de raciocínio público do ER, ver Brandstedt e Brännmark (2020, p. 368-371) e Baserin (2017, p. 2-14).

agente, o que poderia implicar o conservadorismo, sendo o mesmo que contar com crenças do repertório do agente que podem ser moralmente insatisfatórias[25].

Então, para resolver esse problema, iniciamos com a filtragem dos deliberadores nas preferências de salvar vidas humanas, salvar mais vidas e salvar os jovens. Eles podem analisar, também, se alguma das outras seis preferências elencadas no experimento poderia servir como juízo ponderado. Dado que os deliberadores são especialistas, em suas áreas de *expertise*, e conhecem o tema dos VAs, além de possuírem algumas virtudes como a razoabilidade, a mente aberta e a imparcialidade, creio não ser controverso que eles aprovariam as três preferências globais identificadas pelo *Moral Machine* para serem testadas posteriormente. Qualquer acréscimo para utilizar alguma das outras preferências listadas – como salvar os que seguem as leis de trânsito ou salvar os mais ricos, os mais magros, as mulheres ou os pedestres ou, ainda, optar que o veículo continue seu movimento – deveria ser justificado. Veja que, em CREP, a preferência por salvar as mulheres foi identificada como uma crença inicial que foi testada pela coerência com os princípios éticos, sendo reprovada pelos três princípios. Mas não houve nenhuma justificação de por que uma preferência listada em sétimo lugar e que não foi indicada pelo experimento deveria contar como um juízo ponderado, embora os autores tenham dito, corretamente, que devemos acreditar inicialmente nas preferências públicas tomadas como robustas, o que excluiria dados que não sejam confiáveis ou representativos (SAVULESCU; GYNGELL; KAHANE, 2021, p. 7).

A próxima etapa é justificar os juízos ponderados por sua coerência com um sistema coerente de crenças, que é formado por princípios éticos e crenças cientificas. Começo com os princípios éticos. Vamos considerar três princípios morais formulados pelas três principais teorias éticas que possuem grande aceitação no debate contemporâneo e que são tradicionais, considerando a história da ética, a saber: o princípio da maximização do bem-estar (utilitarismo), o princípio da universalizabilidade e não instrumentalização (kantismo) e o princípio do agente virtuoso (ética das virtudes). Além disso, vamos ver se as preferências em salvar vidas humanas, salvar mais vidas e salvar os jovens seriam aprovadas por esses princípios. Esse passo é importante, pois a referência teórica é dada pelos valores e princípios contidos nas teorias morais tradicionais, possibilitando o contraste dos juízos ponderados com os critérios normativo-morais das teorias éticas mais influentes e em que a comunidade confia. Exatamente por essa razão, diferentemente do CREP (SAVULESCU; GYNGELL; KAHANE, 2021, p. 10), testaremos as intuições compartilhadas, por sua coerência com as teorias éticas que, de fato, têm a confiança das pessoas, selecionando a ética das virtudes no lugar do contratualismo rawlsiano.

No modelo utilitarista de ato, a ação correta é a que maximiza a felicidade ou o bem-estar dos envolvidos. Claramente, ele apela para as melhores consequências do ato e para os melhores resultados. Assim, o princípio da maximização do bem-estar aprovaria as três preferências. A preferência em salvar vidas humanas seria aprovada em razão do critério preferencial da vida que seria mais útil, bem como aconteceria com as preferências de salvar mais vidas e de salvar os jovens em detrimento dos mais velhos, considerando que os jovens teriam mais utilidade no futuro. Já no modelo kantiano, a ação correta é a que seria aprovada por uma regra, a qual se deseja que seja universalizada e que não instrumentalize as pessoas. Portanto, o princípio da dignidade humana ou da não instrumentalização aprovaria a preferência de salvar vidas humanas, mas proibiria a preferência de salvar mais vidas, uma vez que toda vida humana tem a mesma

[25] Beauchamp e Childress (2013, p. 400) defendem que devemos iniciar com aquelas crenças morais que são parte da "moralidade comum" e feita por "juízes morais" que possuem virtudes relevantes, tais como a imparcialidade e simpatia. Embora concordem que a justificação é uma questão de coerência, ponderam que apenas a coerência não é suficiente. Assim, devemos contar com juízos ponderados que sejam aceitáveis, inicialmente, sem referência à coerência. São juízos que são confiáveis, isto é, em que temos confiança, a partir da própria história da experiência moral, não sendo o caso de uma intuição individual.

dignidade, e, pela mesma razão, proibiria a preferência aos jovens[26].

Por fim, na ética das virtudes, a ação correta é a que seria aprovada/realizada por um agente virtuoso, sendo esse aquele que delibera sobre os meios adequados para um bom fim e identifica a mediedade. Nesse modelo, a virtude intelectual da prudência é chave, pois é o que garante que se identifiquem os meios adequados para a realização de um fim bom. Porém, além da prudência, o agente deverá contar com virtudes morais que lhe possibilitem encontrar o meio-termo entre os extremos em um caso específico. Nesse caso em tela, além da prudência, penso que o agente deveria contar com a benevolência, de forma a considerar o bem-estar de todos os envolvidos e, sobretudo, a justiça, para dar a cada um o que é devido imparcialmente. Assim, o agente virtuoso aprovaria a preferência por salvar vidas humanas e salvar mais vidas, em razão da ideia de que a virtude é uma disposição de caráter que possibilita uma vida boa ou um florescimento, o qual estará fortemente associado à condição de racionalidade. Sobre a preferência em salvar os mais jovens, é incerta a maneira como o agente virtuoso julgaria. Certamente, há razões para salvar os jovens – e isso ligado à ideia de cuidado que se deve aos mais vulneráveis. Por outro lado, um agente prudente poderia querer investigar mais o caso, olhando o que a lei diz, que consequências negativas poderiam ter na comunidade, como a dos idosos serem tratados de forma desigual etc.[27]

A última etapa consiste em contrastar a coerência dos juízos ponderados e princípios éticos com a dos juízos factuais assegurados por teorias aceitas pelos pares e relevantes no caso em tela e, também, estabelecer um contraste com as leis. A ideia é formar um sistema coerente de crenças mais abrangente. Diferentemente de CREP, que propõe um ER estreito (ERE), proponho, aqui, um ERA, de modo a conectar os valores com os fatos. Com isso em mente, poderia se analisar como as leis, tanto nacionais como internacionais, avaliariam essas três preferências, bem como poderia se levar em conta os relatórios e regulamentos de comissões de ética que tratam dos problemas envolvidos com a IA[28]. Um exemplo disso é o relatório da comissão de ética do Ministério dos Transportes e Infraestrutura Digital da Alemanha, que trata especificamente desse problema dos VAs e proíbe, expressamente, qualquer distinção baseada em características pessoais, tais como idade, gênero, constituição física ou mental[29]. Também, poderia se observar os parâmetros que são usados na medicina, ao redor do mundo, em casos de distribuição de bens escassos. Por exemplo, durante a pandemia de COVID-19, os profissionais de saúde utilizaram o critério de maximização de bem-estar para privilegiar os mais jovens e mais saudáveis (com menos comorbidades), em relação ao uso dos respiradores e leitos de UTI. Isso parece que, de fato, salvou mais vidas, mesmo estando em desacordo com as intuições morais que privilegiam os mais doentes (vulneráveis)[30]. Também, os deliberadores poderiam considerar fatores econômicos, pensando no bem comum.

O resultado do ERAC será a proposta de certos parâmetros que servirão de orientação para o funcionamento dos VAs. Com certeza, o parâmetro de salvar vidas humanas em contraposição a um bem móvel ou a um animal, por exemplo, será um deles. Talvez, os parâmetros de salvar mais vidas e salvar os mais jovens, preferencialmente, também façam parte dessa proposta. Entretanto, isso irá depender do trabalho realizado pelos deliberadores. A ideia básica é que esses parâmetros sejam justificados, tanto em seu contexto de construção como em uma justificação pública,

---

[26] Sobre a caracterização do utilitarismo e kantismo, ver Brink (2006), e Hill (2006).

[27] Aqui, estou seguindo a sugestão de Hursthouse (1991 *apud* CRISP; SLOTE, 1997, p. 219), que apresenta o modelo da ética das virtudes de forma similar aos modelos consequencialista e deontológico, explicando-o da seguinte forma: "P.1. Uma ação é correta *se* ela for o que um agente virtuoso faria em determinadas circunstâncias. P.1a. Um agente virtuoso é aquele que age virtuosamente, isto é, é aquele que tem e exercita as virtudes. P.2. Uma virtude é um traço de caráter que um ser humano precisa para florescer ou viver bem".

[28] Sobre o conteúdo normativo dos diversos regulamentos existentes para IA, ver Floridi (2021).

[29] Ver o relatório da Ethics Commission (2017).

[30] Sobre o tema, ver Grover, McClelland e Furnham (2020).

de forma a contar como uma decisão coletiva e democrática, pois, como esse tema envolve a segurança de todos, é importante a participação de toda a sociedade (ao menos por representação), nessa tomada de decisão, o que já serviria de antídoto a um modelo em que a decisão seja feita tecnocraticamente e orientada apenas pelas regras do mercado econômico[31].

Importante considerar, por fim, que os juízos ponderados, os princípios éticos e os juízos factuais formam um sistema de crenças que parece ser coerente, além de possuir outras virtudes teóricas, tais como simplicidade, precisão e, até mesmo, fertilidade – e isso em razão de seus atributos básicos. Uma das características do sistema formado é sua plurinormatividade, incluindo o princípio da maximização do bem-estar, o princípio da dignidade humana e as virtudes de prudência, benevolência e justiça coerentemente. Outro aspecto é a sua conexão entre os valores e os fatos, com uso das crenças científicas e da legislação como forma de testar a razoabilidade dos juízos ponderados e princípios éticos. Uma última peculiaridade para a qual gostaria de chamar atenção é a sua dimensão democrática, uma vez que a proposta será uma deliberação pública que contará com a aceitação dos envolvidos. E esse sistema, de fato, apresenta uma solução ao problema.

Claro que teríamos que investigar mais detalhadamente as propriedades do sistema formado e, até mesmo, compará-lo com os outros sistemas alternativos. Porém, dada a nossa incerteza normativo-moral, parece que adotar a política que é mais coerente com os juízos ponderados, com os modelos éticos que confiamos e com relevantes dados factuais confere uma alta ordem de justificação em uma sociedade democrática.

Referências

ALBERSMEIR, Frauke. *The Concept of Moral Progress*. Berlin: De Gruyter, 2022.

ANDERSON, Michel; ANDERSON, Susan Leigh. General Introduction. *In:* ANDERSON, Michel; ANDERSON, Susan Leigh (eds.). *Machine Ethics*. New York: Cambridge University Press, 2011. p. 1-4.

AWAD, Edmond *et al*. The Moral Machine Experiment. *Nature*, New York, v. 563, p. 59-64, 2018.

BBC NEWS. Uber's self-driving operator charged our fatal crash. *BBC News*, London, 2020. Disponível em: https://www.bbc.co.uk/news/technology-54175359. Acesso em: 26 jan. 2023.

BEAUCHAMP, Tom L.; CHILDRESS, James F. *Principles of Bioethical Ethics*. Oxford: Oxford University Press, 2013.

BOGOSIAN, Kyle. Implementations of Moral Uncertainty in Intelligent Machines. *Minds & Machines,* Norwell, v. 27, p. 591-608, 2017.

BONNEFON, Jean- François; SHARIF, Azim; RAHWAN, Iyad. The Moral Psychology of AI and Ethical Opt-Out Problem. *In:* LIAO, Matthew (ed.). *Ethics and Artificial Intelligence*. New York: Oxford University Press, 2020. p. 109-126.

BRANDSTEDT, Eric; BRÄNNMARCK, Johan. Rawlsian Constructivism: A Practical Guide to Reflective Equilibrium. *The Journal of Ethics*, Hanover, v. 24, p. 355-373, 2020.

BRINK, David. Some Forms and Limits of Consequentialism. *In:* COPP, David (ed.). *The Oxford Handbook of Ethical Theory.* New York: Oxford University Press, 2006. p. 380-423.

BUCHANAN, Allen; POWELL, Russell. *The Evolution of Moral Progress: A Biocultural Theory*. New York: Oxford University Press, 2018.

CAMPBELL, Richmond. Reflective Equilibrium and the Moral Consistency Reasoning. *Australasian Journal of Philosophy*, Adelaide, v. 92, n. 3, p. 433-451, 2014.

CAMPBELL, Richmond; KUMAR, Victor. Moral Reasoning on the Ground. *Ethics*, Chicago, v. 122, n. 2, p. 273-312, 2012.

COPELAND, Jack. *Artificial Intelligence*: A Philosophical Introduction. Oxford: Blackwell, 2001.

CRISP, Roger; SLOTE, Michael (ed.). *Virtue Ethics*. Oxford: Oxford University Press, 1997. p. 217-238.

DANIELS, Norman. Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy*, New York, v. 76, n. 5, p. 256-282, 1979.

ENGLISH NEWS. China's Baidu operates taxi night in Wuhan. *English News*, Beijing, 2022. Disponível em: https://english.news.cn/20221227/c06149e517884fabb79d1b0cad7950d1/c.html. Acesso em: 26 jan. 2023.

ETHICS COMMISSION. Automated and connected driving. *Federal Ministry of Transport and Digital Infrastructure*, Berlin, 2017. Disponível em: www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf. Acesso em: 5 fev. 2023.

[31] Tasioulas (2022) elabora uma interessante abordagem ética para a IA que transcende as limitações e distorções do modelo utilitarista de maximização do bem-estar, propondo o paradigma da ética humanista, que defende o pluralismo de valores e a justiça dos procedimentos, e não apenas o foco nos resultados e a participação individual e coletiva na construção das regras que orientarão a IA.

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

FLORIDI, Luciano; COWLS, Josh. A Unified Framework of Five Principles for AI in Society. *In:* FLORIDI, Luciano (ed.). *Ethics, Governance and Policies in Artificial Intelligence*. Berlin: Springer, 2021. p. 5-17.

FOOT, Philippa. *Virtues and Vices.* Oxford: Blackwell, 1978.

FRANKENFIELD, Jake. Artificial Intelligence: What It Is and How It is Used. *Investopedia*, New York, 2022. Disponível em: https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp. Acesso em: 26 jan. 2023.

GROVER, Simmy; MCCLELLAND, Alastair; FURNHAM, Adrian. Preferences for Scarce Medical Resource Allocation: Differences between Experts and the General Public and Implications for the Covid-19 Pandemic. *British Journal of Health Psychology*, Toronto, v. 25, p. 889-901, 2020.

HARMAN, Gilbert; MANSON, Kelby; SINNOTT-AMSTRONG, Walter. Moral Reasoning. *In:* DORIS, John Michael (ed.). *The Moral Psychology Handbook.* Oxford: Oxford University Press, 2010. p. 206-245.

HARRIS, John. The Immoral Machine. *Cambridge Quarterly of Healthcare Ethics*, Cambridge, v. 29, p. 71-79, 2020.

HILL, Thomas. Kantian Normative Ethics. *In:* COPP, David (ed.). *The Oxford Handbook of Ethical Theory.* New York: Oxford University Press, 2006. p. 480-514.

KAUR, Kanwaldeep; RAMPERSAD, Giselle. Trust in Driverless Cars: Investigating Key Factors Influencing the Adoption of Driverless Cars. *Journal of Engineering and Technology Management*, Amsterdam, v. 48, p. 87-96, 2018.

KUSHNER, Thomasine; BELLIOTTI, Raymond A.; BUCKNER, Donald. Toward a Methodology for Moral Decision Making in Medicine. *Theoretical Medicine and Bioethics*, Hanover, v. 12, n. 4, p. 281-293, 1991.

LARSON, Jeff *et al*. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*, New York, 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Acesso em: 26 jan. 2023.

MACASKILL, William. Normative Uncertainty as a Voting Problem. *Mind*, East Sussex, v. 125, n. 500, p. 967-1004, 2016.

MESA, Natalia. Can the Criminal Justice System's Artificial Intelligence ever be Truly Fair? *Massive Science*, New York, 2021. Disponível em: https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/. Acesso em: 26 jan. 2023.

PIETSCH, Bryan. Two killed in driverless Tesla car crash, officials say. *The New York Times*, New York, 2021. Disponível em: https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html. Acesso em: 26 jan. 2023.

RAJCZI, Alex. On the Incoherence Objection to Rule-Utilitarianism. *Ethical Theory and Moral Practice*, Hanover, v. 19, p. 857-876, 2016.

RAWLS, John. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.

RECHNITZER, Tanja. *Applying Reflective Equilibrium*: Towards the Justification of a Precautionary Principle. New York: Springer, 2022.

ROZENFIELD, Monica. The next step for artificial intelligence is machines that get smarter on their own. *The Institute*, [*S.l.*], 2016. Disponível em: http://theinstitute.ieee.org/technology-topics/artificial-intelligence/the-next-step-for-artificial-intelligence-is-machines--thatget-smarter-on-their-own. Acesso em: 23 jan. 2023.

SAVULESCU, Julien; GYNGELL, Christopher; KAHANE, Guy. Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies. *Bioethics*, Toronto, v. 35, n. 7, p. 1-12, 2021.

SCANLON, Thomas. Rawls on Justification. *In:* FREEMAN, Samuel (ed.). *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press, 2003. p. 139-167.

TASIOULAS, John. Artificial Intelligence, Humanist Ethics. *Daedalus*: The Journal of the American Academy of Arts & Sciences, Cambridge, v. 151, n. 2, p. 232-243, 2022.

THOMSON, Judith Jarvis. Killing, Letting Die, and the Trolley Problem. *Monist*, Oxford, v. 54, p. 204-217, 1976.

WALLACH, Wendel; ALLEN, Colin. *Moral Machine: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009.

WESSLING, Brianna. Waymo expand service area in 2 cities. *The Robot Report*, Santa Barbara, 2022. Disponível em: https://www.therobotreport.com/waymo-expands-service-area-in-2-cities/. Acesso em: 23 jan. 2023.

Denis Coitinho

Doutor em Filosofia pela Pontifícia Universidade Católica do Rio Grande do Sul. Professor do Programa de Pós-Graduação da Universidade do Vale do Rio dos Sinos (Unisinos). Bolsista produtividade do CNPq.

Endereço para correspondência:

DENIS COITINHO

Avenida Unisinos, 950

Cristo Rei, 93022-750

São Leopoldo, RS, Brasil

Os textos deste artigo foram revisados pela Texto Certo Assessoria Linguística e submetidos para validação dos autores antes da publicação.