# A Lesson from AI: Ethics Is Not an Imitation Game[1]

# Uma lição da IA: Ética não é um jogo da imitação[2]

Gonzalo Génova, Valentín Moreno Pelayo, M. Rosario González Martín

(Tradução de Mariana Rocha Bernardi)

## Resumo

Como podemos ensinar padrões morais de comportamento para uma máquina? Um dos mais importantes alertas contra a IA[3] é a necessidade de evitar vieses[4] em tomadas de decisões eticamente carregadas, mesmo se a população sobre a qual a aprendizagem é baseada também possui um viés próprio. Isto é especialmente relevante quando consideramos que a equidade (e a proteção das minorias) é uma noção ética que por si própria vai além (muito provavelmente) das opiniões tendenciosas[5] das pessoas: a equidade deve ser perseguida e assegurada pelas estruturas sociais, independentemente se as pessoas concordam ou não. Nós sabemos (acreditamos?) que o viés, ou estar suscetível a vieses, é uma coisa ruim, independentemente do que a maioria diz. Em outras palavras, *bem e mal não são o que a maioria diz*, isso está além de maiorias e de fórmulas matemáticas. A ética não pode ser baseada numa opinião majoritária sobre o que é certo e errado ou em um rígido código de conduta. Nós devemos superar o ceticismo generalizado em nossa sociedade sobre a racionalidade da ética e dos valores. A boa notícia é que a IA está nos forçando a pensar a ética de uma nova forma. A tentativa de formalizar a ética em uma série de regras perde de vista que uma pessoa não é apenas uma instância de um caso, mas um ser único e irrepetível. A ética deveria nos prevenir do erro de tentar

[1] Esta pesquisa recebeu investimento do projeto RESTART – "Engenharia reversa contínua para linhas de produção de software" (ref. RTI2018-099915-B-I00, Convocatória de Projetos de I + D Investigação Desafios do Programa Estatal de I+D+i Orientada aos Desafios da Sociedade 2018, contrato de concessão n: 412122; do projeto ECSEL18 "NovoControle" (Projeto 6221/31/2018), e fundo Nacional PCI n. 449990; e do projeto CritiRed – "Elaboração de um modelo preditivo para o desenvolvimento do pensamento crítico no uso das redes sociais", Convocatória Desafios de Investigação do Ministério da Ciência, Inovação e Universidades (2019-2022), ref. RTI2018-095740-B-I00.

G. Génova está ligado ao Departamento de Ciência da Computação e Engenharia, Universidade Carlos III de Madrid, Espanha (ggenova@inf.uc3m.es).

V. Moreno Pelayo está ligado ao Departamento de Ciência da Computação e Engenharia, Universidade Carlos III de Madrid, Espanha (vmpelayo@inf.uc3m.es).

M. R. González Martín está ligada ao Departamento de Estudos de Educação, Universidade Complutense de Madrid (marrgonz@ucm.es).

[2] Esta tradução foi realizada por Mariana Rocha Bernardi com cooperação da CAPES, de acordo com a Portaria 206, de 04 de setembro de 2018: "O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 "This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001".

[3] Inteligência Artificial.

[4] O viés é a inclinação ou tendência de emissão de juízo valorativo acerca de um determinado ato.

[5] Opto por utilizar *tendencioso* para a tradução de *biased*, ao invés de enviesado (que possui viés).

converter equidade em igualdade matemática, adquirida através da extração de características e a computação de uma fórmula de valor. Equidade não é igualdade matemática, nem mesmo uma massificada igualdade que considera diferentes fatores.

**Palavras-chave**: Julgamento tendencioso. Pensamento crítico. Equidade e igualdade. Ética aprendida. Ética programada. Racionalidade da ética.

## 1 Ética não é um jogo da imitação

*O jogo da imitação* é o título de um filme de 2014 que fala sobre a vida de Alan Turing, e especialmente sua destacada participação na decodificação das mensagens criptografadas alemãs pela máquina Enigma no complexo de *Bletchley Park*[6] [7]. A expressão "o jogo da imitação" é do próprio Turing: estas foram as primeiras palavras do seu artigo de 1950, *Computing Machinery and Intelligence*[8]. É também o nome de um jogo muito comum realizado pela aristocracia Vitoriana, que consistia numa troca cega de mensagens escritas à mão, cujo objetivo era tentar adivinhar se o interlocutor seria uma mulher ou um homem.

Quando Turing propôs seu experimento famoso – o Teste de Turing – para determinar se uma máquina poderia pensar, seu foco estava na noção de inteligência enquanto capacidade de resolver problemas "fechados" (ou seja, computáveis), cujo paradigma seria o jogo de xadrez ou a resolução de um enigma[9]. Não é que ele não se preocupasse com questões éticas, obviamente ele se preocupava, mas possivelmente eles não entraram a fundo no que ele considerava "inteligência".

Aqueles foram os primórdios da inteligência artificial (IA). Mas as questões de IA que surgem atualmente sobre ética em geral, e equidade em particular, são cada vez mais importantes (nós queremos dizer 'equidade' no sentido de atendimento às circunstâncias e necessidades de cada indivíduo, o que geralmente não implica na 'igualdade' matemática de certas características pessoais mensuráveis). Nós vemos uma ameaça em particular em alguns equívocos sobre ética entre os desenvolvedores

[6] HINSLEY, F.H.; STRIPP, A. (Eds.), **Codebreakers***:* The inside story of Bletchley Park. Oxford: Oxford University Press, 1993.
[7] HODGES, A. **Alan Turing:** The enigma. London: Burnett Books, 1983.
[8] TURING, A.M. "Computing Machinery and Intelligence". **Mind***,* n. 59, p. 433-460, 1950.
[9] O autor usa a expressão "deciphering of keys", sem tradução equivalente em português, motivo pelo qual optamos por traduzir por resolução de enigma.

de ética da máquina. Neste artigo nós refletimos sobre algumas destas questões, e também sobre *o que podemos aprender* quando consideramos como ensinar comportamento ético e equidade para uma máquina.

## 2 Ética programada

O comportamento ético de máquinas tem oferecido o tema para uma vasta quantidade de especulações mais ou menos fantasiosas, especialmente na ficção científica. Um exemplo paradigmático é Isaac Asimov e seu legendário *As Três Leis da Robótica* (veja fig. 1), primeiramente formulado em uma história escrita em 1941 (*Runaround*[10]), ao mesmo tempo que Turing estava trabalhando duro e fazendo algo bastante significativo em *Bletchley Park*. A curta história se tornaria mais tarde parte da coleção *Eu, robô* (a qual, a propósito, deu o que falar em um filme de 2004 com o mesmo título).

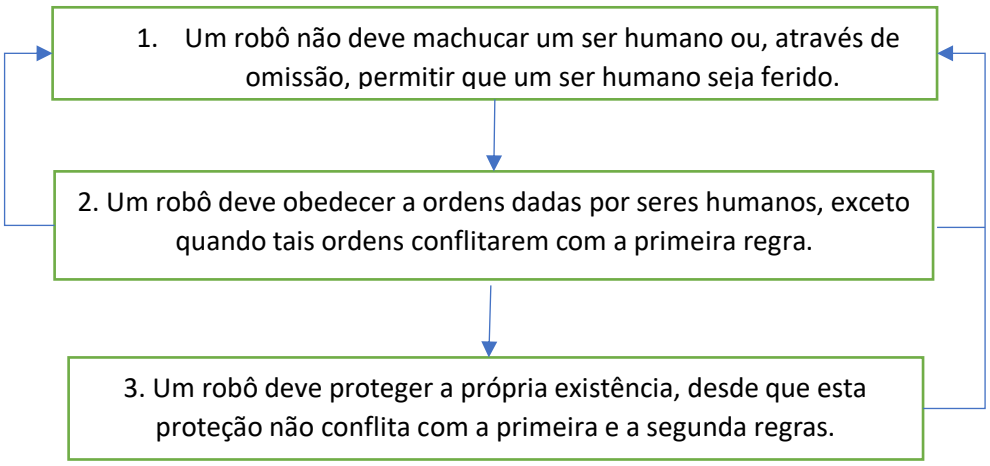


*Figura 1. As Três Leis da Robótica por Isaac Asimov*

Deixando de lado as dificuldades dos aspectos conceituais e técnicos que a implementação hipotética das Três Leis em uma máquina real apresentaria, o interesse do universo robótico Asimoviano é, em nossa opinião, duplo. Ele reside na introdução de elementos futurísticos tecnológicos, mas acima de tudo na dinâmica que resulta da interação dessas três leis: os conflitos que provocam, e as ingênuas –

[10] Optamos por manter a obra em seu título original, embora possa ser traduzido por “rodeio”. Trata-se de um conto escrito por Asimov em 1941.

e às vezes heroicas – soluções que os humanos envolvidos nas histórias encontram. Em certo sentido, elas são tanto histórias de ficção científica quanto são histórias de ficção ética, política ou sociológica.

As Três Leis da robótica são uma tentativa de explicitamente programar um código ético em um robô, isto é, um tipo explícito ou de *ética programada*. A programação clássica é uma explícita descrição procedimental de instruções que uma máquina computacional (um computador) tem que seguir para realizar uma tarefa. A tarefa pode ser somar dois números, levar um pacote de um lugar para outro, fazer um transplante de córnea... Este tipo de programação tenta resolver a solução de um dado problema a partir de uma sequência de instruções, repetições e caminhos alternativos (essa última dependendo se certas condições são ou não satisfeitas).

Embora as Três Leis sirvam um pouco mais do que uma introdução à ética programada – e é claro que Asimov apenas pretendia usá-las como um dispositivo literário –, houve tentativas preliminares de implementá-las em um robô humanoide programável[11]. Contudo, mesmo admitindo que dificuldades técnicas poderiam ser superadas em algum ponto, também foi argumentado que eles seriam uma base insatisfatória para a Ética da Máquina[12].

## 3 Ética aprendida

A verdade é que tentar entender a ética dessa forma, fortemente atrelada a um código que seria capaz de contemplar todos os casos possíveis quando se lhes apresentassem[13] (o que é o que um programa faz) é bastante problemático, como recentes pesquisas sobre ética de máquinas demonstram[14] [15] [16]. É por isso que outra abordagem surgiu, de mãos dadas com os desenvolvimentos em inteligência artificial.

[11] VANDERELST, D.; WINFIELD, A. “An architecture for ethical robots inspired by the simulation theory of cognition”. **Cognitive Systems Research,** n. 48, p. 56-66, 2018.
[12] ANDERSON, S.L. “The unacceptability of Asimov’s three laws of robotics as a basis for machine ethics”. *In:* ANDERSON, M.; ANDERSON, S. L. (Eds.). **Machine ethics.** Cambridge: Cambridge University Press, 2011. p. 285-296.
[13] No sentido de que, diferente de um programa de computador, a ética trabalha com ocorrências não previstas, desconhecidas, chamadas pela filosofia de “contingências”.
[14] NALLUR, V. “Landscape of Machine Implemented Ethics”. **Science and Engineering Ethics**, v. 26, n. 5, p. 2381-2399, 2020.
[15] LUMBRERAS, S. “The Limits of Machine Ethics”. **Religions**, n. 8, p. 100, 2017.
[16] TORRESEN, J. “A Review of Future and Ethical Perspectives of Robotics and AI”, **Frontiers in Robotics and AI,** v. 4, n. 75, 2018.

Uma abordagem que nós podemos chamar de ética implícita ou *ética aprendida*. No MIT (Instituto de Tecnologia do Massachusetts) eles têm desenvolvido um experimento para “aprender como fazer máquinas morais” [9]. Eles chamaram o projeto de *A Máquina Moral*, que é apresentada conforme segue (veja Fig. 2):

*Bem-vindo à Máquina Moral! Uma plataforma para reunir uma perspectiva humana sobre decisões morais tomadas por máquinas inteligentes, tais como carros autônomos. Nós mostramos dilemas morais em que um carro sem motorista deve escolher o menor entre dois males, tais como entre matar dois passageiros ou cinco pedestres. Enquanto um observador externo, você julga qual resultado você pensa ser mais aceitável. Você pode então observar como suas respostas se comparam às de outras pessoas. Se você estiver se sentindo criativo, você pode desenhar seus próprios cenários, para que você e outros usuários consultem, compartilhem e discutam.*

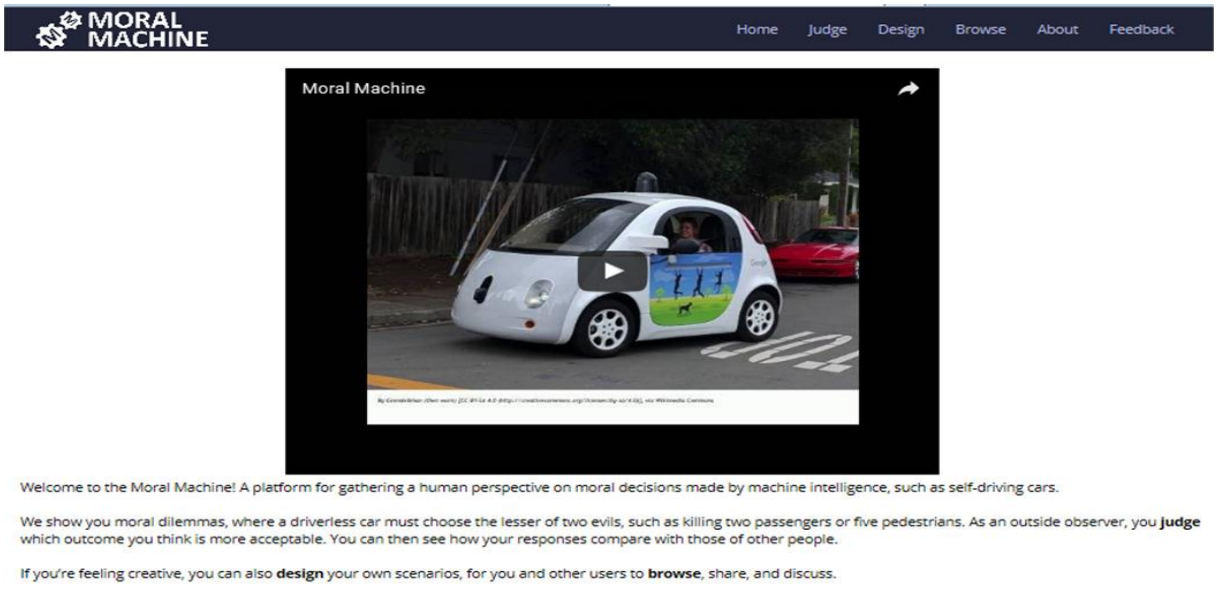


*Figura 2. Apresentação da Máquina Moral do MIT*

## 4 Ensinando máquinas a fazer escolhas difíceis

O domínio de veículos autônomos está muito presente nos meios de comunicação, logo é muito apropriado garantir que o projeto Máquina Moral seja amplamente conhecido em todo o mundo. Mas os mesmos tipos de técnicas e forma

de raciocínio desenvolvido aqui para resolver dilemas morais poderia ser aplicado não apenas para veículos autônomos, mas para outros diferentes problemas: quem selecionar para um emprego, a quem conceder uma liberdade condicional, quem selecionar para um transplante de órgão etc. (Claro, alguém pode perguntar se a ética consiste primariamente na resolução de dilemas, como se isso fosse uma resolução de problemas de geometria... mas vamos deixar isso por enquanto). Para abordar os dilemas de condução que eles enfrentam, os especialistas tentam aprender a partir das respostas que pessoas comuns dariam a estas questões.

Esta é a ideia. Uma vez que não sabemos explicitamente programar o código de ética de um veículo autônomo, vamos perguntar às pessoas: o que você faria? Nesta situação. E nesta outra. E assim por diante. Informações são extraídas a partir de um conjunto de repostas, até que uma série de padrões de comportamentos são construídos. O projeto da Máquina Moral de fato lidou com um impressionante número de respondentes[17]: eles recolheram 40 milhões de decisões em dez idiomas diferentes de milhares de pessoas em 233 países e territórios[18].

Algo similar é feito em outros domínios: uma vez que não sabemos como programar uma série de regras explícitas para reconhecer um rosto humano, pedimos às pessoas que reconheçam faces; nós também pedimos a elas que distingam se este está sorrindo, se aquele outro está zangado, triste ou preocupado. Até mesmo gestos corporais podem ser reconhecidos desta forma: um gesto amigável entre amigos que não tenham se visto por um longo tempo ou entre duas pessoas que firmaram um acordo, um gesto ameaçador... Estas são todas as técnicas que já são bem conhecidas e que têm funcionado bem: deixe-nos tentar, então, aplicá-las para extrair conhecimento moral das respostas das pessoas.

Podemos ver um exemplo concreto na Fig. 3: quem deveria ser atingido por um veículo autônomo, a mulher ou o homem? E assim segue em várias situações, algumas mais complexas que outras, mas sempre na forma de um dilema A *versus* B:

[17] Optamos em traduzir a expressão “respondentes” por “respondentes” ao invés de “entrevistados”, entendendo que a abordagem do projeto Máquina Moral é passiva, no sentido de recebera as respostas das pessoas que voluntariamente buscam o *site* do projeto e não como ocorreria no caso de uma entrevista típica, em que o projeto atuaria ativamente, escolhendo as pessoas a consultar (entrevistar).
[18] AWAD, E.; SOUZA, S. D.; KIM, R.; SCHULZ, J.; HENRICH, J.; SCHARIFF, A.; BONNEFON, J.-F.; RAHWAN, I. “The Moral Machine experiment”. **Nature**, v. 563, p. 59-64, 2018.

homem *versus* mulher, idosos *versus* crianças, gatos *versus* cães, crianças *versus* gatos, passageiros *versus* pedestres etc.

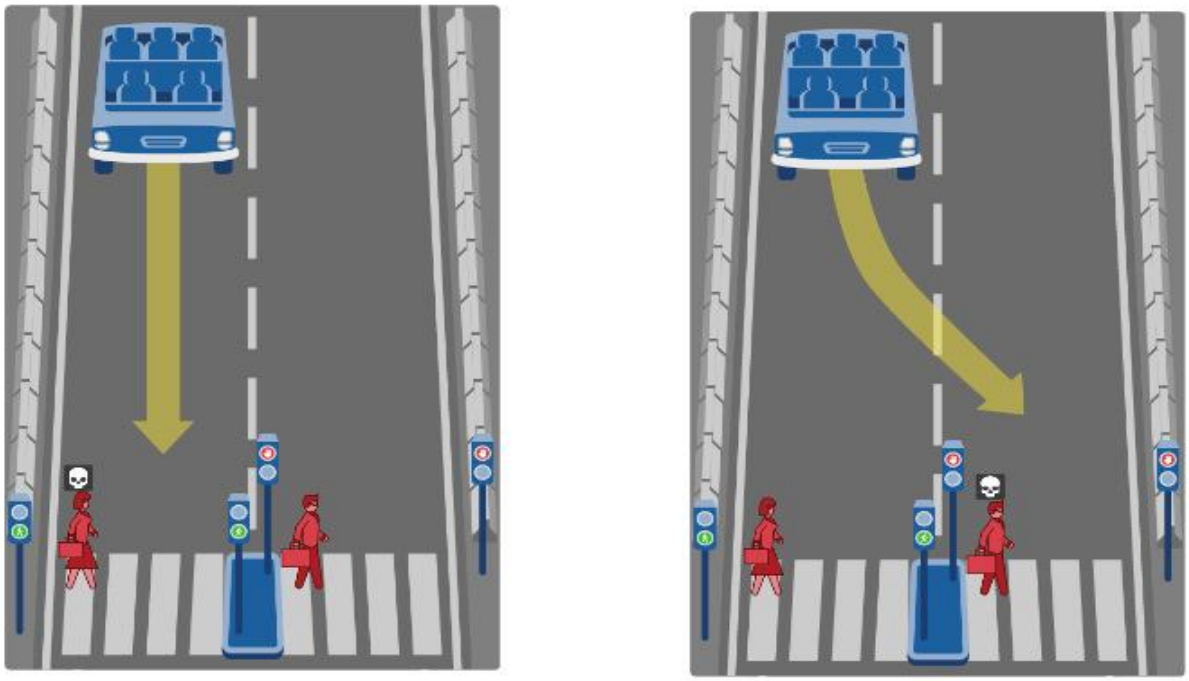

*Figura 3. Ilustração de um dilema de carros autônomos. Quem atropelar?*

Entre as pequenas figuras que podemos selecionar se queremos desenhar nosso próprio cenário, há inclusive um trabalhador da saúde (homem ou mulher) que é facilmente identificado por sua maleta contendo uma cruz (veja Fig. 4). Talvez ele (ou ela) carregue a vacina do coronavírus e salvando-o poderemos salvar outras centenas de milhares de pessoas. Isso deveria influenciar nossa decisão?


*Figura 4. A vida de um profissional da saúde é mais valiosa? Bem, esta é a opção que os respondentes costumam preferir...*

## 5 Problemas da aprendizagem de máquina

**Explicabilidade**. A aprendizagem de máquina explora a construção de algoritmos que podem aprender a partir de dados, através da construção de um modelo a partir da inserção de exemplos, no intuito de realizar previsões ou decisões baseadas em dados, ao invés de seguir um programa estritamente estático de instruções[19]. Quais são os problemas com essa forma de abordar questões éticas? Primeiro, há o problema da *Explicabilidade*, que é a grande preocupação dos pesquisadores. No aprendizado de máquina (que é apenas uma das ramificações da IA) o resultado é uma fórmula, um algoritmo, para reconhecimento da face, do gesto, da escrita. Mas não é possível explicar por que isso funciona. Não há nenhuma outra justificativa para a fórmula que sua alta taxa de sucesso, sua efetividade (de nenhum modo uma surpresa, desde que algoritmos de aprendizagem são desenhados e testados especificamente para adquirir uma alta taxa de sucesso). E é claro, quando a decisão tem uma carga ética pesada, o fato de que ninguém pode fornecer razões para isso, é um problema muito sério.

A falta de Explicabilidade é especialmente severa no campo das redes neurais, mas isso também se faz presente em outras técnicas de IA, tal como a regra de indução (também conhecida como *regra de inferência*). Regra de indução é um tipo de técnica de aprendizado que processa dados de treinamento de entrada e produz uma série de regras SE-ENTÃO para modelar o comportamento desejado. Neste caso, o treinamento de dados seria de diferentes situações ou dilemas apresentados aos respondentes, junto com suas respostas (quem atropelar). Tem se argumentado que a regra de indução tem vantagem sobre as redes neurais, na qual os algoritmos de decisão são apresentados como uma árvore de decisão ao invés de uma caixa preta. Esse formato inteligível para humanos melhoraria a interpretabilidade e explicabilidade. Mas de fato, enquanto a árvore de decisão se tornar tão complicada quanto necessária para cobrir o maior número de instâncias num conjunto de treinamento, ela perde toda a sua suposta “razoabilidade” e “simplicidade”. No fim, é um outro tipo de fórmula matemática.

[19] BISHOP, C.M. **Pattern Recognition and Machine Learning.** New York: Springer, 2006.

**Viés.** Em segundo lugar, um outro problema que surge repetidamente é o problema do *viés*. Algoritmos éticos devem evitar vieses, sejam lá quais forem. Vieses contra mulheres, contra Africanos Americanos, contra aqueles que se vestem de forma inapropriada etc. Mas então, o que acontece se a população que estamos consultando no experimento é tendenciosa? Nós estaríamos reproduzindo os vieses da maioria, e isso não é aceitável: se a maioria for tendenciosa, não é suficiente imitá-la. Os vieses devem ser evitados, não importa o que a maioria diga. Em outras palavras, a busca por equidade nas relações sociais exige que julgamentos tendenciosos sejam evitados.

Isto destaca um ponto muito interessante: nós sabemos (acreditamos?) que o viés, ou ser tendencioso[20], é uma coisa ruim, independentemente do que a maioria diga. Em outras palavras, *bem e mal não são aquilo que a maioria diz*, está além disso (veja Fig. 5). Aliás, mesmo se certos vieses pudessem ser benéficos em promover a equidade, então a distinção entre vieses benéficos e prejudiciais exigiriam um referencial para além de si próprios. Esta não é uma consideração original: está nas próprias origens da reflexão ética na filosofia Grega (veja, por exemplo, o alerta de Platão contra o risco de fazer escolhas baseadas em falsas opiniões ao invés do conhecimento, ou a cuidadosa distinção feita por Aristóteles entre o interesse comum da sociedade e o interesse de uma maioria[21]). O que é interessante é que a IA traz essa consideração de volta para o foco.

*Figura 5. Uma lição da IA: bem e mal não é o que a maioria diz.*

[20] Ou, o mesmo que “estar inclinado a vieses”.
[21] OBER, J. “Democracy’s Wisdom: An Aristotelian Middle Way for Collective Judgment”. **American Political Science Review,** v. 107, n. 1, p. 104-122, 2013.

Mas, então, se não é o que a maioria diz, *como sabemos o que é certo e o que é errado?* Um problema muito difícil... Nós não iremos reivindicar aqui que teremos a resposta, pronta a ser explicada em três linhas. Contudo, nós ousamos dizer que desistindo de tentar conhecer o bem e o mal em si mesmos – e este "em si mesmos" significa que eles estão além da opinião da maioria – é uma séria limitação para nossa pesquisa, nossos programas educacionais, nossas políticas públicas e sociais, etc.

**Seleção da amostra**. E terceiro, intimamente relacionado ao primeiro, está o problema da *seleção de amostra*. A seleção das pessoas para quem perguntamos: quem você atropelaria? Ou seja, quem é a quem estamos perguntando, pessoas comuns ou sádicas? E por que sádicos não deveriam ser incluídos na pesquisa? Não estamos enviesando[22] a amostra? E como sabemos quem são as pessoas comuns, e quem são os sádicos?

Se evitar sádicos for um consenso, isto significa que aplicamos um critério ético na seleção da amostra: isso *a priori*, ou seja, prévio ao experimento. Apesar de não conhecermos perfeitamente, sabemos, de alguma forma, o que é o bem e o que é mal antes mesmo de realizar o experimento. Essa é a razão pela qual excluímos sádicos da amostra.

## 6 Ética, maiorias e pensamento crítico

Resumidamente, nossa reivindicação é que a ética não consiste em imitar comportamentos típicos ou da maioria. Nós humanos não ensinamos ética dessa forma, nem queremos que seja ensinado dessa forma. Se um governo nacional ou regional tivesse incluído em seus programas de ética uma abordagem como a que segue: "crianças devem imitar a maioria" ... não nos rebelaríamos com tremenda indignação? A própria semente da ética é o pensamento crítico, a não conformidade com o pensamento dominante, o compromisso da consciência de cada um em reconhecer por si próprio o que é certo e o que é errado.

[22] Tornando-a tendenciosa.

Alguns afirmam que a ética dos veículos autônomos deveria ser "customizada ao máximo denominador comum do território onde ela será usada"[23]. Claro: um código de ética diferente para escolher quem atropelar em Nova Iorque, Joanesburgo e Hong Kong? *Pode alguém dizer maior barbárie*? Enquanto uma solução técnica, fazer um veículo se comportar como a maioria dos motoristas faria em uma determinada área deve ser uma solução tecnológica adequada. Sem dúvida, um veículo programado para operar em Estocolmo seria desastroso se dirigisse em Roma, e vice-e-versa, até pior[24]! Mas por favor não chame isso de comportamento ético. Pode ser efetivo, mas não nos deixemos enganar em pensar que isso seja ética.

Geralmente, os termos 'ética' e 'moralidade' são usados de forma intercambiável, especialmente em contextos acadêmicos[25]. Não obstante, muitas distinções entre estes dois termos têm sido propostas, mesmo que nenhum tenha alcançado um consenso universal. Uma das distinções mais comuns atribui a moralidade a um sendo "descritivo", enquanto à ética é atribuído um sendo "normativo". Em outras palavras, a moralidade descreve costumes sociais ou códigos de conduta; a ética, ao invés disso, se refere ao que *atualmente* é certo ou errado, para além do que seja socialmente aceito. Deste modo, a Máquina Moral é certamente efetiva em refletir costumes sociais, mas nunca será uma verdadeira Máquina Ética.

A abordagem adotada pela Máquina Moral tem sido severamente criticada, questionando a importância e o real valor de encontrar laços fortes entre as preferências dos entrevistados e seus atributos culturais, e até mesmo a moralidade de toda a discussão[26]. Jaques chega ao ponto de chamar a Máquina Moral de "um

---

[23] LEONARD, C. **"Teaching ethics to machines".** 2016. [Online]. Disponível em: https://www.linkedin.com/pulse/teaching-ethicsmachines-charles-leonard.

[24] Os autores se referem a algumas diferenças existentes entre as regras de trânsito em Estocolmo e em Roma, como exemplo para reforçar que a programação de uma ética pretensamente universal em carros autônomos que se locomovessem em territórios distintos, com regramento distinto, poderia causar grandes problemas. A título de curiosidade, existe um material disponível em https://partiueuropa.com/dirigindo-em-estocolmo/ em que a autora faz referência a algumas diferenças existentes nas regras de trânsito em Estocolmo em relação a outros países da Europa.

[25] GERT, B.; GERT, J. "The Definition of Morality". **The Stanford Encyclopedia of Philosophy.** [Online]. Available: https://plato.stanford.edu/archives/fall2020/entries/moralitydefinition. Accessed on 26/09/2020.

[26] NASCIMENTO, A.M.; VISMARI, L.F.; QUEIROZ, A.C.M.; CUGNASCA, P.S.; CAMARGO JUNIOR, J.B.; ALMEIDA JUNIOR, J.R. de. "The Moral Machine: Is It Moral?". **2nd International Workshop on Artificial Intelligence Safety Engineering within 38th International Conference on Computer Safety, Reliability, and Security,** September 10-13, 2019, Turku, Finland. Lecture Notes in Computer Science, v. 11699, p. 405-410.

monstro"[27], porque de fato convida as pessoas a expressarem suas preferências – isto é, vieses – por indicadores externos de *valor social*. Não apenas porque humanos são extremamente ruins em julgar pessoas com base em suas aparências; mas especialmente porque a Máquina Moral dá a impressão de que este valor social perseguido deveria influenciar as suas decisões. Isso sugere e inclusive advoga pela relevância moral dessas características – o que não é nada menos que um ataque direto a equidade. Etienne[28] discute o perigo do experimento da Máquina Moral, alertando contra ambos os seus usos para fins normativos e toda a abordagem do sistema de votação que é construído em cima de uma abordagem de problemas éticos. De acordo com Puri[29], mesmo se o comportamento moral pudesse ser imitado pelos algoritmos, apenas agentes conscientes podem tomar decisões e assumir responsabilidade por elas. A partir de um ponto de vista mais geral, os Anais da edição especial do IEEE[30] sobre ética da máquina contêm diversos artigos que enfatizam a proteção devida a valores humanos em sistemas autônomos futuros; especialmente significativo aqui é Bremmer et. al.[31], que argumentam por um raciocínio ético transparente e verificável em sistemas de IA.

Ética não é um jogo da imitação. Ética não é sobre seguir um código de conduta como As Três Leis de Asimov. Mas a ética também não é sobre imitar o comportamento de outros. Aprender por imitação é algo bem humano, mas se há alguma diferença entre #fiquemecasa e o #recolhaopapelhigienico, essa diferença não está no número de pessoas que se comportam de uma maneira ou de outra, mas na razoabilidade de suas condutas. A ética não se espalha como um vírus; é aprendida e discutida racionalmente.

---

[27] JACQUES, A.E. "Why the moral machine is a monster". **We Robot Conference,** University of Miami School of Law, April 11-13, 2019.
[28] ETIENNE, H. "When AI Ethics Goes Astray: A Case Study of Autonomous Vehicles". **Social Science Computer Review**, v. 40, n. 1, p. 1-11, 2020.
[29] PURI, A. "Moral Imitation: Can an Algorithm Really Be Ethical?". **Rutgers Law Record,** v. 48, n. 1, p. 47-58, 2020.
[30] "IEEE, pronunciado "Eye-triple-E", significa Instituto de Engenheiros Elétricos e Eletrônicos. [...] O IEEE é a maior associação profissional do mundo dedicada ao avanço da inovação e excelência tecnológica para o benefício da humanidade.". Disponível em https://edu.ieee.org/it-unina/en/about-ieee/. Sobre os anais: https://www.ieee.org/conferences/organizers/preparing-conference-proceedings.html.
[31] BREMER, P.; DENNIS, L.A.; FISHER, M.; WINFIELD, A.F. "On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots". **Proceedings of the IEEE,** v. 107, n. 3, p. 541-561. Special Issue on Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems, 2019.

## 7 A racionalidade da ética: equidade e igualdade

Então, um dos obstáculos que temos que superar a fim de recobrar a sanidade é o ceticismo generalizado em nossa sociedade sobre a racionalidade da ética e valores[32], que também é manifestado na existência de múltiplas teorias éticas incompatíveis[33] [34]. Emotivismo – não uma integração saudável de emoções com razões no julgamento moral, mas uma substituição deste último pelo anterior – é promovido em abordagens tais como a Máquina Moral, onde as *razões* dos entrevistados para suas respostas são negligenciadas; apenas as respostas cruas são levadas em consideração, como se se dissesse: "a ética não é racional em si mesma, então nós devemos nos contentar em refletir o que a maioria das pessoas diz". Como temos mostrado, isto é radicalmente insuficiente para evitar julgamentos enviesados e garantir a equidade.

O papel das emoções na tomada de decisões morais é bem estabelecido[35], mas isso não significa que podemos reduzir a ética à emoção, como o emotivismo faz. Em um contexto cultural onde a pós-verdade está destruindo as estruturas de conhecimento atuais[36] [37], um algoritmo que produz uma aparência de consenso em reações emocionais de uma grande quantidade de pessoas, mas sem argumentos válidos, corre mais risco de ser manipulado, tendencioso ou intoxicado do que antes do surgimento das redes sociais. A boa notícia é que a IA está nos forçando a pensar ética de uma nova forma.

Uma das mais frequentemente ouvidas formas de argumentação ética é: "se todos agirem assim, então tal coisa ocorreria" (o que é claramente uma derivação do imperativo categórico de Kant / imperativo universal). Isso é o que tem sido chamado consequencialismo baseado em regras, isto é, realizar avaliações éticas de acordo

[32] GÉNOVA, G.; GONZÁLEZ, M.R. "Teaching Ethics to Engineers: A Socratic Experience". **Science and Engineering Ethics**, v. 22, n. 2, p. 567-580, 2016.
[33] NALLUR, V. "Landscape of Machine Implemented Ethics". **Science and Engineering Ethics**, v. 26, n. 5, p. 2381-2399, 2020.
[34] BOGOSIAN, K. "Implementation of moral uncertainty in intelligent machines". **Minds and Machines,** v. 27, n. 4, p. 591-608, 2017.
[35] MAY, J.; WORKMAN, C.; HAAS, J.; HAN, H. "The Neuroscience of Moral Judgment: Empirical and Philosophical Developments". *In*: BRIGARD, F. de; SINNOTT-ARMSTRONG, W. (Eds.). **Neuroscience and Philosophy**. Massachusetts: MIT Press, 2022.
[36] Optamos pelo uso de "atuais" ao invés de "existentes" ou "vigentes", em vista do contexto e do alcance pretendido pelos autores.
[37] SISMONDO, S. "Post-truth?". **Social Studies of Science**, v. 47, n. 1, p. 3-6, 2017.

com as consequências da aplicação de certas regras de comportamento, não tanto de atos concretos[38] [39]. Sem negar o valor que este tipo de argumentação tem em iluminar nossa razão ética, nós não podemos ficar satisfeitos em ser este o único tipo de racionalidade ética aceitável. As limitações racionais do consequencialismo são claras e têm sido denunciadas há muito tempo; aqui está nossa versão da refutação[40]:

*Primeiramente, as consequências de uma determinada ação se estendem por um período de tempo que não tem propriamente um limite, ainda que não possamos indefinidamente esperar para julgar se uma ação é boa ou má. Em segundo lugar, mesmo se pudéssemos colocar um limite adequado às consequências que queremos considerar, eles, contudo, pertencem ao porvir, por isso são bastante incertos; nós teríamos que empregar algum tipo de técnica de predição para prever as consequências e avalia-las; mas estas técnicas sempre seriam limitadas pela própria natureza das coisas, a qual não segue regras de comportamento perfeitamente conhecidas (além disso, consequências provavelmente dependerão da liberdade dos outros). Em terceiro lugar, e mais importante, se nós queremos evitar regras a priori para determinar a bondade para as ações, e nós fazemos com que a bondade de uma ação dependa da bondade das consequências, então nós precisamos de regras para avaliar a bondade das consequências; o consequencialismo extremo não resolve o problema da bondade, mas simplesmente o adia.*

A Máquina Moral pede para você "julgar qual resultado você pensa ser mais aceitável". Ela processa suas respostas, junto com as dos milhares de respondentes, e produz um modelo de comportamento. Mas ética aprendida e ética programada são diferentes apenas num primeiro momento, isto é, se nós considerarmos apenas a maneira que as regras de comportamento têm sido produzidas. Atualmente, ambas são o mesmo tipo de coisa: um programa que faz escolhas baseado nas regras orientadas a alcançar o "melhor" resultado. Um programa feito de regras que surgem

[38] JOHNSON, D.G. **Computer Ethics.** 2ª ed. Upper Saddle River: Prentice Hall, 1994.
[39] LAUDON, K.C. "Ethical Concepts and Information Technology". **Communications of the ACM**, v. 38, n. 12, p. 33-39, 1995.
[40] GÉNOVA, G.; GONZÁLEZ MARTÍN, M.R.; FRAGA, A. "Ethical education in software engineering: responsibility in the production of complex systems". **Science and Engineering Ethics**, v. 13, n. 4, p. 505-522, 2007.

desde uma especulação a priori ou a partir de uma imitação de padrões atuais de comportamento humano. Não é uma diferença tão grande.

O problema é que uma abordagem baseada em resultados computados e a redução da ética à compilação e aplicação de um conjunto de regras, sejam *a priori* ou aprendidas, é um severo equívoco do que seja ética. Acima de tudo, a tentativa de formalizar a ética em um conjunto de regras perde de vista o fato de que *uma pessoa não é somente uma instância de um caso*, mas um ser único e irrepetível. Uma pessoa é uma criança, um estudante, um paciente, um cliente, um vizinho (*seu* filho, *seu* estudante, *seu* paciente, *seu* cliente, *seu* vizinho). O valor de uma pessoa não pode ser mensurado com um número, nem mesmo com uma variedade de números (idade, sexo, condição de saúde, contribuição social...). Quando se trata de pessoas, nossa razão precisa ser treinada para entender o que nós vemos não apenas enquanto casos de uma regra geral, mas acima de tudo em sua unicidade. Nós precisamos recuperar a racionalidade do valor para além dos números, além da racionalidade lógica e instrumental. Nós precisamos aprender como raciocinar com valores de um modo que eles não se convertam em números.

A ética deveria nos prevenir do erro de converter equidade em uma igualdade matemática, adquirida através da extração de características e a computação de uma fórmula de valor. Equidade não é igualdade matemática, nem mesmo uma igualdade ponderada que considera diferentes fatores. O primeiro mandamento da ética deveria ser “não tratarás uma pessoa como *um vetor de números*”. Isso exige uma reforma da racionalidade da ética? Que seja bem-vinda!

## Referências

ANDERSON, S.L. “The unacceptability of Asimov’s three laws of robotics as a basis for machine ethics”. *In:* ANDERSON, M.; ANDERSON, S. L. (Eds.). **Machine ethics.** Cambridge: Cambridge University Press, 2011. p. 285-296.

AWAD, E.; SOUZA, S. D.; KIM, R.; SCHULZ, J.; HENRICH, J.; SCHARIFF, A.; BONNEFON, J.-F.; RAHWAN, I. “The Moral Machine experiment”. **Nature**, v. 563, p. 59-64, 2018.

BISHOP, C.M. **Pattern Recognition and Machine Learning.** New York: Springer, 2006.

BOGOSIAN, K. “Implementation of moral uncertainty in intelligent machines”. **Minds and Machines**, v. 27, n. 4, p. 591-608, 2017.

BREMER, P.; DENNIS, L.A.; FISHER, M.; WINFIELD, A.F. “On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots”. **Proceedings of the IEEE,** v. 107, n. 3, p. 541-561. Special Issue on Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems, 2019.

ETIENNE, H. “When AI Ethics Goes Astray: A Case Study of Autonomous Vehicles”. **Social Science Computer Review,** v. 40, n. 1, p. 1-11, 2020.

GÉNOVA, G.; GONZÁLEZ, M.R.; FRAGA, A. “Ethical education in software engineering: responsibility in the production of complex systems”. **Science and Engineering Ethics**, v. 13, n. 4, p. 505-522, 2007.

GÉNOVA, G.; GONZÁLEZ, M.R. “Teaching Ethics to Engineers: A Socratic Experience”. **Science and Engineering Ethics,** v. 22, n. 2, p. 567-580, 2016.

GERT, B.; GERT, J. “The Definition of Morality”. **The Stanford Encyclopedia of Philosophy.** Disponível em: https://plato.stanford.edu/archives/fall2020/entries/moralitydefinition.

HINSLEY, F.H.; STRIPP, A. (Eds.), **Codebreakers**: The inside story of Bletchley Park. Oxford: Oxford University Press, 1993.

HODGES, A. **Alan Turing:** The enigma. London: Burnett Books, 1983.

JACQUES, A.E. “Why the moral machine is a monster”. **We Robot Conference**, University of Miami School of Law, April 11-13, 2019.

JOHNSON, D.G. **Computer Ethics.** 2ª ed. Upper Saddle River: Prentice Hall, 1994.

LAUDON, K.C. “Ethical Concepts and Information Technology”. **Communications of the ACM**, v. 38, n. 12, p. 33-39, 1995.

LEONARD, C. **“Teaching ethics to machines”.** 2016. [Online]. Available: https://www.linkedin.com/pulse/teaching-ethicsmachines-charles-leonard.

LUMBRERAS, S. “The Limits of Machine Ethics”. **Religions**, n. 8, p. 100, 2017.

MAY, J.; WORKMAN, C.; HAAS, J.; HAN, H. “The Neuroscience of Moral Judgment: Empirical and Philosophical Developments”. *In*: BRIGARD, F. de; SINNOTT-ARMSTRONG, W. (Eds.). **Neuroscience and Philosophy**. Massachusetts: MIT Press, 2022.

NALLUR, V. “Landscape of Machine Implemented Ethics”. **Science and Engineering Ethics**, v. 26, n. 5, p. 2381-2399, 2020.

NASCIMENTO, A.M.; VISMARI, L.F.; QUEIROZ, A.C.M.; CUGNASCA, P.S.; CAMARGO JUNIOR, J.B.; ALMEIDA JUNIOR, J.R. de. “The Moral Machine: Is It Moral?”. ***2nd International Workshop on Artificial Intelligence Safety Engineering* within *38th International Conference on Computer Safety, Reliability, and Security***, September 10-13, 2019, Turku, Finland. Lecture Notes in Computer Science, v. 11699, p. 405-410.

OBER, J. “Democracy’s Wisdom: An Aristotelian Middle Way for Collective Judgment”. **American Political Science Review,** v. 107, n. 1, p. 104-122, 2013.

PURI, A. “Moral Imitation: Can an Algorithm Really Be Ethical?”. **Rutgers Law Record,** v. 48, n. 1, p. 47-58, 2020.

SISMONDO, S. “Post-truth?”. **Social Studies of Science,** v. 47, n. 1, p. 3-6, 2017.

TORRESEN, J. “A Review of Future and Ethical Perspectives of Robotics and AI”, **Frontiers in Robotics and AI**, v. 4, n. 75, 2018.

TURING, A.M. “Computing Machinery and Intelligence”. **Mind**, n. 59, p. 433-460, 1950.

VANDERELST, D.; WINFIELD, A. “An architecture for ethical robots inspired by the simulation theory of cognition”. **Cognitive Systems Research**, n. 48, p. 56-66, 2018.

Recebido em: 27/09/2022.
Aprovado em: 03/11/2022.