A Ilusão do que conta como Agente

Luís Moniz Pereira
(NOVA Laboratory for Computer Science and Informatics: NOVA-LINCS,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa)
e
Fernando da Costa Cardoso
(Conselho Nacional de Pesquisa: CNPq Brasil)

Por vezes, mais do que um esforço analítico, o filósofo deveria recorrer ao olhar do antropólogo para compreender os assuntos que o interessam. Com isso, queremos dizer que ele deve pôr de lado por um momento as suas ferramentas críticas, a fim de desenvolver o seu interesse pela categorização. Embora a história de tentativas como esta ainda esteja por escrever, esforços como o de Latour e Woolgar (2013) indicam que vale a pena, pois desta forma liga-se o jogo de conceitos não apenas a eles mesmos, mas a outros aspectos como, para continuar com o exemplo, a dimensão agonística da vida laboratorial. É o que pretendemos neste artigo: fornecer uma simples categorização de diferentes tipos de sistemas que podem ser descritos como máquinas morais (ou agentes éticos artificiais), de modo que, se se permitir que as ferramentas críticas desenvolvidas sejam reincorporáveis, o filósofo terá como objectivo, em primeiro lugar, destacar outras conexões. Desta forma, aproximamo-nos do desenvolvimento de máquinas morais e agentes éticos artificiais, à medida que chegamos a um momento que se pode entender como uma espécie de consolidação que se alcançou nesta segunda década do século.

A nossa análise destaca três tipos de máquinas mais ou menos identificáveis:

(1) Simuladores de certos comportamentos;
(2) Emuladores de padrões eticamente sensíveis;

e, finalmente, o tipo de dispersão que identificamos como

(3) Replicadores e aprimoradores.

Ao mesmo tempo em que alcançámos essa imagem do campo, fornecemos o que identificámos como o produto do esforço de categorização do filósofo: que estes diferentes

tipos de sistemas contribuem para a compreensão da dimensão ética, numa era de expansão desta, e na qual tais seres contam já como seus membros.

Essa segunda parte deste breve artigo destaca como cada um desses três tipos releva a necessidade de reavaliação de uma concepção central da dimensão ética, respectivamente, a autonomia, o papel das aparências e da identidade e, finalmente, as nossas concepções sobre quais os ideais que nos movem e determinam o que valorizamos.

Introdução

A literatura oferece tentativas de categorização do campo conhecido como moralidade artificial ou ética das máquinas, um campo que compartilha connosco o objectivo declarado de uma dupla compreensão, embora, como veremos, valorizemos as diferenças tanto as que são devidas a pontos de partida diferentes quanto as que se devem a diferentes interesses. Por exemplo, Wallach e Allen (2009) propõem que os diferentes sistemas e modelos desenvolvidos, tanto por filósofos que tentam apoiar concepções de como a ética opera com o recurso à computação quanto pela comunidade dos que trabalham a desenvolver aplicações, podem ser classificados de acordo com o nível de onde partem. Os esforços teóricos que tentam apoiar ou demonstrar uma certa visão da nossa própria vida ética, ou de uma ou de outra teoria sobre alguns aspectos dela, assumem o que eles chamam de abordagem "de cima para baixo", onde um certo modelo é construído e governado antecipadamente pelo que poderíamos entender como as implicações dessas teorias. Deste modo, as teorias da racionalidade são aplicadas à construção de um módulo "ético" (operado, por exemplo, pelas funções *prima facie* identificadas como importantes para o comportamento ético dos seres humanos) que circunscreve posteriormente o funcionamento do sistema como um todo a diferentes objectivos (como agregados prudenciais ou de segurança, por exemplo). Formulações "de baixo para cima" de como desenvolver máquinas morais na direcção oposta, fornecem de alguma forma um caminho mais intuitivo para fundamentar a ética nos seres que consideramos capazes de participar da dimensão ética. Fazem isso de certa forma emulando uma dimensão de aprendizagem e evolução – seguindo a sugestão de Turing (1950) de não se concentrar na mente adulta, mas na da

criança –, algo semelhante ao que supomos ter acontecido connosco. Essa categorização fornece aos autores exactamente o panorama de crítica que queremos alcançar.

Mas a crítica lida com a interpretação, e a tarefa hermenêutica só pode ser enriquecida pela complementaridade de diferentes focos de análise. Aquilo que sugerimos aqui é buscado tendo em mente uma ideia eminentemente moderna: que só entendemos o que construímos, e que o esforço no campo das máquinas morais pode ser entendido como uma tentativa de nos compreendermos.

A nosso ver, podemos organizar os diferentes sistemas desenvolvidos, a fim não tanto de esgotar o domínio mas para destacar o modo como sistemas aparentemente incomensuráveis, especialmente porque projectados com objectivos diferentes, compartilham de facto um ponto comum de autodesenvolvimento.

1. Simulação – Autonomia

Em suma, propomos uma interpretação de dois esforços computacionais para dizer, a um interessado em Ética, que os fenómenos que ele tenta entender poderiam ser capturados num nível mais simples com resultados frutíferos para essa investigação. Um nível certamente ainda não rodeado pelos grandes valores que tão prontamente tenta identificar com a Ética, perdendo, no processo dessa mesma identificação, uma perspectiva que poderia ter permitido uma multiplicidade de agentes, com diferentes graus (e respectivas restrições) de autonomia, proporcionando, assim, um relato mais rico da autonomia. Ambicionamos fornecer uma concepção crítica da autonomia de maneira que evitasse esquecer os inícios mais humildes desses valores, dos nossos próprios valores. No entanto, este resultado é colateral em relação aos nossos objectivos principais, uma vez que não era nossa intenção fornecer aqui uma lição sobre a humildade do que valorizamos e sobre a sua precariedade no nosso mundo. Se assim se desejar, isso pode ser alcançado dentro de uma discussão filosófica, por exemplo, nas observações finais do ensaio de Bernard Williams sobre a sorte moral (Williams, 1981), onde o autor destaca o carácter circunstancial da ética e, portanto, o facto de que esta não pode ser o terreno mais alto que é inacessível aos agentes – a "jóia brilhante" de Kant –, mesmo que sejam agentes humanos.

Só uma melhor compreensão de como podemos avaliar estas coisas poderia alcançar este resultado.

Com isso em mente, os nossos objectivos têm sido os de fornecer essa melhor compreensão. Um objectivo tem sido contribuir para uma melhor compreensão da autonomia e, seguindo as repercussões do nosso primeiro exemplo, em Cardoso e Pereira (2015), uma melhor compreensão do modo como a autonomia pode ser interpretada como tendo evoluído, traçando a sua evolução no campo dos agentes artificiais. Certamente, os modelos são apenas modelos, mas as tarefas gémeas da nossa investigação, a de desenvolver agentes morais mais legítimos e a de uma melhor compreensão de nós mesmos, permanecem abertas. Acreditamos que o nosso trabalho é realizado no nexo dessas duas tarefas e, portanto, que ele responde de forma mais eficiente a esses dois desafios do que a comum avaliação *a priori*. De facto, a nossa posição é a de que os modelos melhores possibilitam melhores teorias, uma vez que permitem a eliminação de algumas das limitações dos exercícios de poltrona, os quais, quando modelados, revelam resultados contra-intuitivos e dificuldades inesperadas. Sustentamos que os modelos são carregados de teoria e tendenciosos, no sentido de que as limitações relativas ao que pensamos que deve ser o caso realmente regulam – reconhecendo aqui uma dimensão de heteronomia – o que eventualmente obteremos como resultados. Isso apoia a necessidade de uma forte co-evolução da construção de modelos e da teoria.

Margaret Boden (1998, 2008) estabeleceu um quadro semelhante para a compreensão da autonomia no campo da vida artificial. As suas preocupações eram as de afastar a noção de que, num mundo determinista – um mundo como os que são simulados nos nossos modelos, seja como os que ela descreve, seja, seguindo alguns resultados da ciência, um como o nosso –, o surgimento de agentes sobrenaturais serviu, principalmente, para confirmar o nosso carácter delirante, com a implicação correlata de uma negação da nossa liberdade. O nosso objectivo tem sido mais estreito, mostrando que temos ferramentas computacionais, não dependentes apenas das nossas intuições, para investigar o conceito de autonomia e de Ética de forma mais geral, oferecendo assim uma oportunidade de actualizar os nossos próprios códigos morais e o meta-raciocínio sobre esses códigos. E, deste modo, a nossa investigação é paralela à dela, na medida em que ajuda a identificar as formas que levantam a nebulosidade que cobriu o conceito de autonomia, ao mesmo tempo

que elimina os erros devidos a ilusões de intuição não-tutelada. Se este ponto for percebido sem adulterações, ficará claro que apenas com modelos mais amplos e melhores podemos alcançar este resultado. Afinal, a liberdade segue-se à compreensão e também pode ter evoluído.

2. Emulação – Aparências

A forte posição de Bringsjord et al. (2006) sobre o assunto certamente poderia ser o ponto de partida para o nosso debate sobre um segundo tipo de sistema que podemos identificar no campo das máquinas morais. A sua formulação usa um *modus tollens* simples para rejeitar a ideia de cópia e a possibilidade de um sucesso real no teste de Turing (1950). O seu argumento é o seguinte:

1. Projecto de construção de pessoas => Pessoas são autómatos
2. Pessoas são autómatos

Assim sendo, segue-se

3. Projecto de construção de pessoas

Estabelecendo "2" a noção de que nós, seres humanos – o caso clássico das pessoas –, possuímos uma série de características (ele menciona o livre-arbítrio, a capacidade de introspecção infalível, a experiência interior) que não estão disponíveis para as máquinas, e tal é apoiado pela sua concepção anti-behaviorista acerca da necessidade de um certo processo de fluxo nas mentes.

Certamente, a ideia de uma cópia do processo que acontece quando alguém escolhe ou avalia algo, claramente mais radical que a "mera" emulação, foi descartada por alguns autores como necessária para o projecto de construção de agentes. É assim que lemos Wallach (2010), por exemplo, quando ele aceita alguma validade nas observações do experimento da Sala Chinesa de Searle e, com base em algumas avaliações cépticas (as de Torrance, 2008, e Franklin, 2003), descarta essa necessidade. Para apoiar a sua posição, ele

aponta para que "pode ser possível que robôs sofisticados possam actuar de maneira moralmente aceitável sem ser fenomenalmente conscientes" (Wallach, 2010, p. 246). Indo mais além, sugere uma separação entre "fenómenos fenomenais" e "atributos funcionais da consciência". A proposta de emulação de Pollock (1995) estava claramente relacionada com essa abordagem.

3. Replicação e Aprimoramento: Reflectindo sobre quem achamos que somos ou devemos ser

Quando deixamos o reino do possível para imaginar fins, entramos no que é de alguma maneira o coração da questão da Ética. Este coração, situado nos nossos ideais, cria certamente dificuldades aos programadores mais prudentes que, justamente preocupados em separar o seu território do da ficção e das ideias nebulosas, podem sentir certo desconforto com as propostas de replicação e aprimoramento que vão sendo analisadas. A nossa insistência aqui é a de que os factos e as ficções criam para si mesmos uma imagem. Neste caso, uma imagem de nós mesmos, ou como indicámos na secção anterior, de nós como membros de uma comunidade de agentes.

Coda

Há uma ilusão neste campo de actividade que é a de que só precisamos resolver o problema uma única vez para obter um resultado seguro. A dependência deste campo em relação aos resultados de uma avaliação sempre contínua do que conta como agente, como autonomia, como acção própria e – ainda que não muito investigada nesta área –, de quais são os limites dos contextos e situações, bloqueia definitivamente essa esperança ilusória.

Algumas das nossas próprias incursões nessa *Terra Incognita* intimam o quanto há sucessivamente por descobrir, nomeadamente em Cardoso e Pereira (2015), Pereira (2016a), Pereira (2016b), Pereira e Saptawijaya (2016), Pereira e Saptawijaya (2017), Saptawijaya e Pereira (2018), Han e Pereira (2018).

Agradecimentos

L. M. Pereira agradece o apoio de FCT/MEC NOVA-LINCS PEstUID /CEC /04516/2013. F. C. Cardoso agradece o apoio do Conselho Nacional de Pesquisa (CNPq / Brasil). Os autores agradecem ao Professor Manuel Curado a tradução inicial para português do rascunho original em inglês.

Referências

Allen, Colin; Gary Varner; e Jason Zinser (2000). "Prolegomena to any future artificial moral agent", *Journal of Experimental & Theoretical Artificial Intelligence*, 12: 3, pp. 251-261.

Anderson, Michael; e Susan Leigh Anderson, eds. (2011). *Machine Ethics*. Cambridge: Cambridge University Press.

Anderson, Michael; e Susan Leigh Anderson (2008). "ETHEL: Toward a Principled Ethical Eldercare Robot", in *Proceedings AAAI Fall 2008 Symposium on AI in Eldercare: New Solutions to Old Problems*. November 7-9, 2008, Arlington, Virginia, USA. Menlo Park CA: Association for the Advancement of Artificial Intelligence **(**AAAI) Press.

Arkin, Ronald C. (2009). *Governing Lethal Behavior in Autonomous Robots*. Abingdon: CRC Press / Taylor and Francis Group.

Boden, Margaret A. (1998). "Autonomy and Artificiality", in Andy Clark e Josefa Toribio, eds., *Cognitive Architectures in Artificial Intelligence: The Evolution of Research Programs*. San Francisco CA: Garland Publishing, Inc., pp. 300-306.

Boden, Margaret A. (2008). "Autonomy: What is it?", *Biosystems*, 91: 2, pp. 305-308.

Bringsjord, Selmer; Konstantine Arkoudas; e Paul Bello (2006). "Toward a general logicist methodology for engineering ethically correct robots", *IEEE Intelligent Systems*, 21: 4, pp. 38-44.

Cardoso, Fernando; e Luís Moniz Pereira (2015). "On artificial autonomy emergence: a view from the foothills of a challenging climb", in Jeffrey White e Rick Searle, eds., *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. Hershey PA: IGI Global, pp. 51-72.

Danielson, Peter (1992). *Artificial Morality: Virtuous Robots for Virtual Games*. London: Routledge.

Danielson, Peter (2010). "Designing a machine to learn about the ethics of robotics: The N-reasons platform", *Ethics and Information Technology*, 12: 3, pp. 251-261.

Dennett, Daniel C. (2004). *Freedom Evolves*. London: Penguin.

Floridi, Luciano (2007). "A look into the future impact of ICT on our lives", *The Information Society*, 23: 1, pp. 59-64.

Haidt, Jonathan (2007). "The new synthesis in moral psychology", *Science*, 316: 5827, 18 May, pp. 998-1002.

Han, The Anh; e Luís Moniz Pereira (2018). "Evolutionary Machine Ethics", in Oliver Bendel, Hrsg., *Handbuch Maschinenethik*. Berlin: Springer. (Acessível em https://link.springer.com/referencework/10.1007/978-3-658-17484-2.)

Kowalski, Robert (2011). *Computational Logic and Human Thinking: How to be Artificially Intelligent*. New York: Cambridge University Press.

Latour, Bruno; e Steve Woolgar (2013). *Laboratory Life: The Construction of Scientific Facts*. Princeton NJ: Princeton University Press.

Lin, Patrick; Keith Abney; e George A. Bekey (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge MA: The MIT Press.

Lopes, Gonçalo Cardoso; e Luís Moniz Pereira, (2010). "Prospective storytelling agents", in Manuel Carro e Ricardo Peña, eds., Practical Aspects of Declarative Languages (12th International Symposium PADL 10). Berlin: Springer-Verlag, pp. 294-296.

Mikhail, John (2011). *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. New York: Cambridge University Press.

Pereira, Luís Moniz (2016a). "Software sans Emotions but with Ethical Discernment", in: S. Silva (ed.), *Morality and Emotion: (Un)conscious Journey into Being*. London: Routledge, pp. 83-98.

Pereira, Luís Moniz (2016b). *A Máquina Iluminada – Cognição e Computação*. Porto: Fronteira do Caos.

Pereira, Luís Moniz; e Saptawijaya, A. (2011). "Modelling Morality with Prospective Logic", in M. Anderson e S. L. Anderson, eds., *Machine Ethics*. Cambridge: Cambridge University Press, pp. 398-421.

Pereira, Luís Moniz; e Ari Saptawijaya (2015). "Bridging Two Realms of Machine Ethics", in Jeffrey White e Rick Searle, eds., *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. Hershey PA: IGI Global, pp. 197-224.

Pereira, Luís Moniz; e Ari Saptawijaya (2016). *Programming Machine Ethics*. Berlin: Springer.

Pereira, Luís Moniz; e Ari Saptawijaya (2017). "Counterfactuals, Logic Programming and Agent Morality", in Rafal Urbaniak e Gillman Payette, eds., *Applications of Formal Philosophy: The Road Less Travelled*. Berlin: Springer, pp. 25-54.

Petersen, Arthur C. (2012). *Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and their role in Climate Science and Policy Advice*. Abington: CRC Press, Taylor and Francis Group.

Pollock, John L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge MA: The MIT Press.

Saptawijaya, Ari; e Luís Moniz Pereira (2018). "From Logic Programming to Machine Ethics", in Oliver Bendel, Hrgb., *Handbuch Maschinenethik*. Berlin: Springer. (Acessível em https://link.springer.com/referencework/10.1007/978-3-658-17484-2.)

Seth, Anil K. (2010). "Measuring autonomy and emergence via Granger causality", *Artificial life,* 16: 2, pp. 179-196.

Turing, Alan M. (1950). "Computing Machinery and Intelligence". *Mind*, 59, pp. 433-460.

Wallach, Wendell; e Colin Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

White, Jeffrey; e Rick Searle, eds. (2015). *Rethinking Machine Ethics in the Age of Ubiquitous Technology*. Hershey, PA: IGI Global.