Responsabilidade Moral Distribuída: Contribuições para o Debate sobre Inteligência Artificial Ética e Responsável

Elizabeth Maria Freire de Jesus

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE)
Universidade Federal do Rio de Janeiro (UFRJ)
Cidade Universitária, Ilha do Fundão - Rio de Janeiro

beth@nce.ufrj.br

Abstract. *The construction of AI systems takes place in distributed and heterogeneous environments, involving an extensive network of human, artificial and hybrid agents, interactions and actions. The objective of this work is to contribute to the debate on ethical and responsible AI, using Luciano Floridi's analytical and conceptual framework, emphasizing his distributed moral responsibility approach as a possible and plausible way to deal with the difficulty of locating agency and attributing responsibility. moral considering the vast, diverse and distributed network of agents involved in the construction of intelligent systems.*

Resumo. *A construção de sistemas de IA se dá em ambientes distribuídos e heterogêneos, envolvendo uma extensa rede de agentes humanos, artificiais e híbridos, interações e ações. O objetivo deste trabalho é contribuir no debate sobre IA ética e responsável, recorrendo ao quadro analítico e conceitual de Luciano Floridi enfatizando a sua abordagem de responsabilidade moral distribuída como uma via possível e plausível para lidar com a dificuldade de localização da agência e atribuição de reponsabilidade moral considerando a vasta, diversa e distribuída rede de agentes envolvidos na construção de sistemas inteligentes.*

1. Introdução

A pervasividade da Inteligência artificial (IA) e seus impactos em praticamente todas as dimensões da vida humana é incontestável, embora nem sempre a sua presença ou agência sejam facilmente identificáveis e o alcance dos impactos de sua ação conhecidos ou dimensionados. Questões como até que ponto esses impactos serão negativos ou positivos, para quem, de que forma, em que lugares e em que escala de tempo estão na base dos debates atuais sobre IA ética e responsável.

As preocupações éticas em torno da sistemas inteligentes (SI) que utilizam conceitos e técnicas de inteligência artificial, integrando métodos matemáticos, lógicos e computacionais – sistemas de inteligência artificial (sistemas de IA) –, aumentaram significativamente nas últimas décadas em face da miríade de decisões e ações cada vez mais delegadas à IA e, em especial, aos seus potenciais e reais danos ou riscos ao ser humano, ao meio ambiente e aos direitos humanos fundamentais. Um caso emblemático que ganhou grande visibilidade na mídia foi o estudo do ProPublica, de 2016, que mostrou

o viés discriminatório racial em decisões automatizadas sobre liberdade condicional, ao considerar os(as) detentos(as) negros(as) com tendo maior chance de reincidência criminal [Angwin et al. 2016].

O debate controverso em torno da agência moral e da responsabilidade moral pelos possíveis efeitos benéficos ou danosos decorrentes das tecnologias de automoção de um modo geral e, em particular, de sistemas de IA não é novo. No início da década de 40, Norbert Wiener, um dos precursores da ética da informação [Bielby 2014], alertava para a conscientização das mais diversas áreas sobre os efeitos benéficos e maléficos dos avanços científicos, como a automação (Chaves and Bernardo 2000).

A construção de SI se dá com e através de arcabouços teórico-conceituais e metodológicos, ideias, discursos, pessoas, instituições, contratos, políticas e infraestrutura de software e hardware que deem conta do exigido poder de armazenamento, processamento e análise de grande volume de dados. Pode-se considerar, portanto, que a consubstanciação desses artefatos é devida às interações e ações de diversos e heterogêneos agentes.

O objetivo deste trabalho é contribuir no debate sobre IA ética e responsável, recorrendo ao quadro analítico e conceitual de Luciano Floridi e de sua abordagem de responsabilidade moral distribuída (RMD) como uma via possível e plausível para lidar com a dificuldade de localização da agência e atribuição de reponsabilidade moral envolvendo sistemas de IA, tendo-se em vista a vasta, diversa e distribuída rede de agentes envolvidos na construção desses sistemas.

2. Sistemas multiagentes e o problema da responsabilidade e imputabiidade da ação moral

O debate ético envolvendo em sua problematização a composição da agência e das novas matrizes acionais não mais exclusivamente autogeridas por um agente humano, individual, autônomo e racional [González de Gomez 2022], tem colocado a discussão em torno de agentes artificiais (AAs) autonômos, "inteligentes" e suficientemente informados que são capazes de executar ações moralmente relevantes, boas ou más, independentemente dos seres humanos que os projetaram e desenvolveram, como é o caso de sistemas de IA.

Agentes inteligentes interagem com seu ambiente e aprendem a partir do que percebem, podendo aumentar seus conhecimentos sobre o próprio ambiente e, possivelmente alterar a si próprios [Russel and Norvig 2010]. Para isso, em grande medida, este aprendizado depende de um grande volume de dados. Para os autores, "um dos ambientes mais importantes para agentes inteligentes é a Internet".

Um recente e controverso exemplo é o ChatGPT treinado a partir de vasto *corpora* produzido pela coleta massiva de dados linguísticos (texto e fala) na internet [Nunes and Soares and Ferro 2023]. *Prosumers* (produtores e consumidores de dados e informação) constituem uma miríade de possíveis fontes de dados, confiáveis e não, que "contribuem" para o aprendizado deste e de outros sistema de IA que aprendem com o ambiente – a internet. Tendo-se em conta as desigualdades de acesso, produção e consumo de dados e informações na internet, como também as idiossincrasias cultural, linguística, ideológica, étnica e etc., não surpreendentemente tem sido observado a produção de textos eventualmente carregados de valores morais negativos, como ofensas, preconceitos e discriminação, com importantes efeitos deletérios reais e potenciais para indivíduos,

grupos, para a sociedade como um todo e para a democracia. Ademais, os traços comportamentais que os usários vão deixando à medida em que navegam na internet, suas pegadas digitais, produzem um superávit comportamental que são transformados em produtos de predição comercializados em mercados de comportamentos futuros [Zuboff 2020].

São inúmeros os sistemas inteligentes com comportamentos e limitações que produzem efeitos moralmente negativos. A IA do Google Cloud Vision, em um estudo com foco em retratos de mulheres negras, constatou que as fotos apresentavam recorrentemente o rótulo "peruca" sempre que seus cabelos estavam em evidência. Não havia no banco de dados, portanto, rótulos indicativos para cabelos cacheados ou que não fossem lisos [Silva and Mintz 2020]. A rotulagem de dados imprescindível para uma variedade de sistemas de IA é realizado por uma multidão de *crowdworkers* espalhados em diversos países que trabalham remotamente realizando essa entre outras microtarefas que podem ser impactadas pela diversidade dos ambientes em que são realizadas, gerando, por exemplo, vieses discriminatórios.

A heterogeneidade e pluraridade inerente às formações composicionais contemporâneas da agência e da ação distribuída de sistemas multiagentes – que podem agregar corporações, agências governamentais, tecnologias, pessoas, normas, etc. – estão na base da dificuldade, ou mesmo da impossibilidade de localização e atribuição de responsabilidade moral. Este é um problema colocado para os sistemas de IA. Um exemplo classico é o acidente ocorrido em 2018, no Arizona, envolvendo um carro autônomo da Uber que causou a morte de um pedestre. Quem seria(m) o(s) responsável(eis)? (os desenvolvedores do software; os projetistas do hardware; a montadora que produziu o carro, a plataforma, no caso a Uber, o usuário do carro, o pedestre, a regulação local (ou a inexistência de uma). Como atribuir responsabilidade? Como responsabilizar?.

A problematização e reflexão próprias da ética colocados em torno de sistemas inteligentes que podem ser fontes legítimas (ou não) de ações morais, devem abarcar a análise do *design*, implantação, controle, comportamento e manutenção desses artefatos, incluindo uma melhor compreensão dos ambientes multiagentes de realização desses processos [Coeckelbergh 2019; Stahl 2023], bem como as diversas e diferentes interações que neles acontecem [Floridi 2016].

3. Fundamentação teórica-conceitual

Agência moral e responsabilidade são categorais centrais no debate ético. Ambos os conceitos são complexos, multifacetados, debatidos e disputados por várias correntes teóricas e práticas, não sendo diferente, quando colocados no debate sobre IA ética ou ética na IA, ou seja, sobre os desafios éticos colocados pela IA atual e do futuro próximo.

Nem toda ação humana é uma ação moral. Atos propriamente morais "são somente aqueles nos quais podemos atribuir ao agente uma responsabilidade não só pelo que se propôs realizar (ou seja, intencionalidade), mas também pelos resultados ou consequências da sua ação" [Sánchez Vásquez 2020]. O ato moral humano pressupõe, portanto, consciência, liberdade, autonomia, motivo e intenção; pressupõe a existência de possibilidades de escolha, justificação e responsabilidade pelas escolhas feitas e responsabilização.

É dada agência à IA, no sentido em que estas fazem coisas no mundo baseada em escolhas que têm consequências. Se considerada a agência moral como algo exclusivamente humano, a discussão ética em torno do comportamento, das escolhas e das consequências da IA, que, ao fim e ao cabo envolvem inúmeros e heterogêneos atores, levam à necessidade de uma maior plasticidade conceitual e analítica da agência e da responsabilidade.

Em sua análise crítica, Floridi desloca o foco de uma ética de responsabilidade baseada em ações intencionais dos indivíduos e orientada à punições e recompensas individuais para uma ética de responsabilidade baseada nas interações multiagentes orientada para o bem da humanidade, do planeta e para a promoção e preservação de direitos humanos fundamentais [Floridi 2016].

Nessa direção, Floridi argumenta que as éticas tradicionais (ocidentais) com suas premissas filosóficas antropocêntricas da agência e o foco inteiramente e apenas na natureza intencional das ações humanas, não dão conta para lidar com a alocação e atribuição de responsabilidade moral em ambientes multiagentes [Floridi 2013]. O foco passa a ser a avaliação da ação moral não do remetente (do agente da ação individual), e sim do receptor individual ou coletivo da ação (de quem ou o quê recebe o efeito de uma ação). De outra forma, Floridi considera que as ações sejam avaliadas com base em seu impacto no bem-estar do ambiente como um todo e de seus habitantantes.

3.1. Ação moral distribuída e atribuição de responsabilidade moral distribuída

No seu quadro conceitual-analítico, Floridi considera que muitas ações acabam sendo neutras, ou seja, não são nem moralmente boas nem moralmente más, seja porque, considerando a vasta rede de agentes, seus efeitos reais são muito pequenos para serem moralmente significativos, ou porque se anulam ou se compensam mutuamente ou ainda por não terem a força e a direção suficientes para superar o seu status de neutras, de forma a passarem de neutras para moralmente carregadas positiva ou negativamente [Floridi 2013].

Em ambientes distribuídos e heterogêneos – alguns humanos, alguns artificiais (por exemplo, um software) e alguns hibrídos (por exemplo, um grupo de pessoas trabalhando mediados por uma plataforma digital) – ações moralmente boas ou más podem surgir resultantes de ações que se realizam em interações locais entre os agentes que não são em si mesmas moralmente boas ou más, e sim neutras ou moralmente insignificante [Floridi 2013].

Ações moralmente distribuídas são resultante ou geradas a partir de interações entre os nós da rede com efeitos em um mais estados do sistema. Assim, a interpretação da rede possibilita o entendimento das ações moralmente distribuídas como o resultado de interações neutras entre os nós da rede (propagação para frente). Por outro lado, a propagação reversa (retropropagação), seria como a responsabilidade moral distribuída pode ser atribuída, tendo-se em vista a melhoria do estado do sistema afetado pelas ações moralmente distribuídas [Floridi 2016]. A granularidade da análise de Floridi, portanto, não está nas ações individuais, mas sim, nas ações geradas a partir das interações entre nós e seus efeitos ou consequências para um ou mais estados de um sistema.

Esta concepção de ação moral distribuída (AMD) responde a um novo regime de reconfiguração e avaliação moral dos contextos acionais contemporâneos fortemente baseados em interações envolvendo agentes humanos, artificiais e híbridos, e das ações

moralmente imputáveis [Gonzalez de Gomez 2020], na medida em que se privilegia na análise, as características ou estados do sistema que se quer perseguir ou evitar. Em muitos casos, é apenas agregando e fundindo os cursos de ação individuais que se faz a diferença moral [Floridi 2016].

Na abordagem de responsabilidade moral distribuída, o sentido de 'responsabilidade' é etiológico, ou seja, remete à fonte (causalmente responsável) de um estado do sistema. Os efeitos das decisões e as ações baseadas em IA são muitas das vezes resultado de incontáveis interações entre muitos atores. Com a agência distribuída vem a responsabilidade distribuída [Taddeo e Floridi 2018].

De particular importância para a abordar a questão da agência de entidades artificiais (como algoritmos), bem como de coletivos (organizações ou empresas) é a consideração de Floridi e Sanders [2004], para quem o agente possível de uma ação moral deve reunir três condições necessárias e suficientes para ser considerado como tal: autonomia, interatividade e adptabilidade. Dessa forma, embora os seres humanos, em grande medida, satisfazem essas condições, neste entendimento nem o livre arbítrio nem as intenções são considerados necessários para a agência.

A interatividade refere-se à mútua ação do agente e o meio ambiente; a autonomia na medida em que o agente tem certo controle sobre suas ações, implicando na sua capacidade de iniciar livremente uma ação e de mudar de estado sem que seja uma resposta a uma interação; e a adaptabilidade indicando a capacidade de aprendizagem, certo grau de autorreflexão, o que significa que o agente, numa interação, e de acordo a sua experiência, pode mudar as regras de transição conforme as quais muda de estado [Durante 2017; Floridi 2013].

A questão crucial ou o grande desafio que se coloca é entender como se pode alocar responsabilidade moral distribuída (RMD) para ações morais distribuídas que surjam de ações totalmente neutras, de modo que as ações corretas sejam facilitadas, promovidas, ampliadas e recompensadas, e as ações erradas impedidas, mitigadas ou punidas em reparação? [Floridi 2016].

O escopo de aplicabilidade da abordagem RMD são todos os agentes individuais envolvidos. É certo que alguns agentes podem compartilhar diferentes graus de responsabilidade, incluindo nenhum. Dessa forma, a princípio, este tipo de abordagem parece produzir situações de injustiças frente à atribuição de responsabilidade e responsabilização. No entanto, uma das vantagens do mecanismo de "responsabilidade por *default*", ou seja, da extensão da responsabilidade, em tese, para todos os nós/agentes relevantes na rede, é gerar um efeito em que um número maior de pessoas busquem ações moralmente boas, como também que para alguns, aumente a chance de se tornarem mais cautelosos [Floridi 2016].

> O que importa é a mudança no sistema causada pela AMD, seja boa ou má; e se for má, que se possa retificá-la ou reduzi-la, tratando a rede como responsável por ela e, assim, propagar de volta a responsabilidade para todos os seus nós/agentes para melhorar o resultado [Floridi 2016, p.7].

Na base de sustentação e plausibilidade da responsabilidade por *default* está o que Floridi denomina de infraética – um neologismo para expressar uma infraestrutura ética

que embora não seja moralmente boa ou má em si mesma, pode facilitar e promover decisões e ações moralmente boas [Floridi 2013]. Trata-se de uma estrutura de primeira ordem de expectativas, atitudes e práticas implícitas [Floridi 2013], que se assumidos valores moralmente bons, podem ter efeitos positivos para a ação moralmente distribuída.

Infraética pode ser entendida como um conjunto de capacitadores morais que devidamente planejados, coordenados, compreendidos e compartilhados podem atuar como promotores e facilatores do bem ou, no limite, podem neutralizar ou pelo menos limitar o mal, do ponto de vista moral [Floridi 2013]. Portanto, cabe perguntar e buscar identificar qual é a natureza e a lógica do tipo certo de infraética, suas interações e operações dentro de um sistema dinâmico e multiagente, como é o caso do desenvolvimento e uso de sistemas de IA.

Disponibilidade, acessibilidade e abertura de dados e informações, confiança, transparência, privacidade, segurança, dignidade, justiça, autonomia, liberdade de expressão, concorrência justa, atitudes na direção de promover iniciativas voltadas para o desenvolvimento de competências e qualificações para lidar com tecnologias de IA são alguns exemplos de capacitadores morais que podem apoiar o tipo certo de moralidade distribuida pautada no diálogo interdisciplinar e intercultural.

4. Considerações finais

A discussão sobre IA ética e responsável é sobretudo uma discussão sobre responsabilidade. A construção de sistemas de IA se dá em ambientes distribuídos e heterogêneos, envolvendo uma extensa rede de agentes humanos, artificiais e híbridos, de incontáveis interações e ações morais, amplificando assim a preocupação e a dificuldade de localização e atribuição de responsabilidade moral, que, em grande medida, emergem somente quando algo dá errado, não atende às expectativas ou resultados esperados ou quando estes são inesperados ou imprevistos.

A crítica de Floridi dirige-se justamente para a natureza antropocêntrica das grandes tradições éticas centradas no agente, em que o julgamento das ações humanas se dá pela avaliação da intenção do agente ou da consequência de sua ação. Tais premissas, para Floridi, fragilizam e restringem as investigações éticas em ambientes distribuídos, onde é cada vez mais comum que uma ação moralmente boa ou má resulte da ação moral de um miríade de agentes. Portanto, uma ética baseada na intenção não é a única disponível nem a mais adequada frente as novas configurações e complexidade das interações multiagentes contemporâneas [Floridi 2013; Floridi 2016].

Na concepção de Floridi, a mudança de perspectiva deve ser na direção de uma ética orientada para o agente que se preocupa com o desenvolvimento dos indivíduos, do bem-estar social, do planeta e dos direitos humanos fundamentais. O foco da ação moral, portanto, passa a ser o paciente da ação.

A ideia de infraética, ou seja, de uma infraestrutura ética pautada em valores, expectativas, atitudes e práticas que se bem desenhados e alinhados com um conjunto de capacitadores morais podem orientar, promover e incentivar o diálogo interdisciplinar e intercultural que deem conta de lidar com diferentes conhecimentos, bem como diferentes condições de possibilidades de conhecer, que busque formas aceitáveis de lidar com diferentes visões de mundo, ideologias e crenças e, sobretudo, possibilite modos e meios

de comunicação, de compartilhamento de informações, conhecimentos e práticas. A ideia de infraética remete à necessidade premente de se construir e promover ambientes que favoreçam as ações moralmente boas e restrinja ou neutralize o efeito das ações moralmente más. Remete à importância de se promover uma cultura de responsabilidade, de forma a contribuir para a garantia dos benefícios e mitigação dos riscos da IA.

As reflexões e contribuições conceituais de Luciano Floridi, apontam algumas direções e possibilidades interessantes para (re)pensarmos os nossos ambientes de construção de artefatos computacionais, em especial, artefatos inteligentes baseados em IA, cujos efeitos morais podem ter expressivos impactos na vida de pessoas, de grupos, bem como para o meio ambiente e para a sociedade como um todo.

As concepções apresentadas por Floridi, demandam investigações e ações em níveis macro, meso e micro dos contextos de concepção, produção e usos de artefatos baseados em inteligência artificial, abarcando, por exemplo, desenvolvimento de arcabouços normativos, regulatórios e avaliativos com aplicações global e local, políticas e ações educacionais que promovam o desenvolvimento de competências críticas e a formação de sujeitos com uma consciência crítica de forma a lidar e agir em um mundo cada vez mais mediado por artefatos tecnológicos baseados em inteligência artificial.

5. Referências

Angwin, J. et al. (2016) "Machine Bias", https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing,

Bielby, J (2014) "Information Ethics III: Concerning Intercultural Information Ethics".

Chaves, V. and Bernardo, C. (2020). História (29): p. 1-18, https://doi.org/10.1590/1980-4369e2020017

Coeckelbergh, M. (2019) "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability", Science and Engineering Ethics (26): p. 2051-2068.

Durante, M. (2017) "Ethics, Law and the Politics of Information: A Guide to the Philosophy of Luciano Floridi". Dortrecht, Springer.

Floridi, L. (2013) "Distributed Morality in an Information Society. Sci Eng Ethics (19): p.727-743.

Floridi, L. (2016) "Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions". Phil.Trans.R. Soc. A (374).

Floridi, L. and Sanders, J. W (2004) On the Morality of Artificial Agents. Minds and Machines 14: p. 349-379.

Gonzalez de Gomez, M.N. (2020) "A ética da informação de Luciano Floridi: Nas Trilhas da Filosofia [Preprint], http://eprints.rclis.org/42284/.

González de Gomez, M.N. (2022) Jogos Morais do Século XXI: Ética da Informação de Luciano Floridi. In Ciência da Informação: Sociedade, Crítica e Inovação, p. 323-348. Rio de Janeiro:IBICT.

Nunes, M.G.V and Soares, T. A. and Ferro, M. (2023) “Questões Éticas em IA e PLN”. In Caseli, H.M.; Nunes, M.G.V. (Ed.). Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português, BPL, p: 467-474.

Sánchez Vázquez, A. (2020) Ética. 39th edition. 304p.

Silva, T. and Mintz, A. (2020) “APIs de Visão Computacional: Investigando Mediações Algorítmicas a partir de Estudo de Bancos de Imagens”. Logos, 1(27): p.25-54.

Stahl, B. C. (2023) “Embedding Responsibility in Intelligent Systems: From AI Ethics to Responsible AI Ecosystems”. Sci Rep (13): p.7586.

Taddeo, M. and Floridi, L. (2018). How AI Can Be a Force for Good. Science, 361 (6404): p.751-752.

Zuboff, S. (2020) “A Era do Capitalismo de Vigilância: A Luta por um Futuro Humano na Nova Fronteira do Poder”, 1st ed., Rio de Janeiro, Intrínseca. 800p.