



INTELIGÊNCIA ARTIFICIAL EMOCIONAL: UMA ANÁLISE ÉTICA *

Emotional Artificial Intelligence: an ethical assessment

Marco Aurélio Sousa Alves **

Ariel Ribeiro Tolentino ***

Resumo: Dentre os vários desenvolvimentos recentes da Inteligência Artificial, iremos nos concentrar nos companheiros artificiais na forma de *chatbots*. Destacamos dois problemas principais relacionados a esse produto. O primeiro diz respeito à atribuição de uma “capacidade empática”, característica que supostamente lhes capacita a atuarem como agentes relacionais e íntimos com seres humanos. Analisamos as limitações atuais de sua arquitetura afetiva e apontamos as ambiguidades existentes no campo da computação afetiva. O segundo problema, de caráter ético, surge como consequência do primeiro. Uma vez que suas limitações são explicitadas, revela-se um desalinhamento entre seu *design* altamente antropomórfico, com amplo escopo de ação, e suas capacidades reais de operar de forma ética e responsável no espaço relacional íntimo. Finalmente, argumentamos que as características desses dispositivos incentivam a projeção de status moral e agência aos *chatbots*, o que acarreta riscos e danos aos usuários.

Palavras-chave: Inteligência Artificial. Computação afetiva. Companheiros artificiais. Capacidade empática. Status moral.

* Artigo recebido em 25.09.2025 e aprovado para publicação em 28.11.2025.

** Doutor em Filosofia pela University of Texas at Austin (EUA). Professor Adjunto do Departamento de Filosofia e Métodos (DFIME) da Universidade Federal de São João del-Rei (UFSJ). ORCID: <https://orcid.org/0000-0002-2296-2644>. E-mail: marcoaurelioalves@ufsj.edu.br. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) – Processo APQ-04336-23.

*** Mestra em Filosofia pela Universidade Federal de Ouro Preto (UFOP). ORCID: <https://orcid.org/0000-0002-5111-6228>. E-mail: arielribeiro@proton.me

Abstract: Among the many recent developments in Artificial Intelligence, we focus our attention on chatbots as artificial companions. We highlight two main issues related to this product. The first concerns the attribution of an “empathic capacity”, a characteristic that supposedly enables them to function as an intimate relational agent towards humans. We analyze the current limitations of their affective architecture and point out the ambiguities within the field of affective computing. The second issue, with a direct ethical import, arises as a consequence of the first. Once their limitations are made explicit, it is revealed that there is a misalignment between their highly anthropomorphic design, which exhibits a broad scope of action, and their actual capabilities to operate ethically and responsibly in intimate relationships. We finally argue that the characteristics of such devices encourage the projection of moral status and agency onto chatbots, which incur risks and harm to users.

Keywords: Artificial intelligence. Affective computing. Artificial companions. Empathic capacity. Moral status.

Introdução

O mercado das Inteligências Artificiais Emocionais (IAEs), voltado para a criação de produtos com arquitetura afetiva, vem crescendo significativamente.¹ Com o intuito de estabelecer uma relação empática e afetiva com seres humanos, as IAEs possuem aplicações que vão desde a área da robótica até a criação de *chatbots* que atuam em diversas áreas.² Com o avanço da Inteligência Artificial Generativa e seu baixo custo, o mercado de *chatbots* de função relacional íntima tornou-se uma das áreas mais lucrativas das IAEs, razão para sua rápida proliferação.³ Esses *chatbots* possuem um amplo escopo de atuação por meio de sua personalização e dinamicidade interativa, sendo comercializados como companhias virtuais capazes de estabelecer laços íntimos diversos, como amizade, relações românticas e sexuais, apoio emocional e mentoria terapêutica.

Tendo como foco a atribuição de uma “capacidade empática” que, supostamente, lhes confere a aptidão de serem companheiros artificiais

¹ Lançado em 2014, o *chatbot* relacional *Xiaolce* alcançou o somatório de 660 milhões de usuários em 2018 (ZHOU, Li *et al.*, The Design and Implementation of Xiaolce, an Empathetic Social Chatbot, in *Computational Linguistics*, v. 46, n. 1, 2020, p. 53). O *chatbot* relacional íntimo *Replika*, lançado em 2017, ultrapassou 10 milhões de usuários ativos apenas no ano de 2022 (CIRIELLO, Raffaele *et al.*, Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024, p. 488).

² *Chatbots* são *softwares* que oferecem acesso a serviços e informações através da interação com usuários em linguagem ordinária, por texto ou voz (SKJUVE, Marita *et al.*, My Chatbot Companion — A Study of Human-Chatbot Relationships, in *International Journal of Human-Computer Studies*, v. 149, 2021, p. 1).

³ A Inteligência Artificial Generativa é um tipo de IA capaz de criar conteúdo novo a partir de padrões identificados nos conjuntos de dados de treinamento.

íntimos de humanos, faremos inicialmente uma análise das limitações de sua arquitetura afetiva e empática e, a partir dessa análise, apontaremos alguns problemas éticos relativos ao desalinhamento entre seu *design* e função de amplo escopo e suas capacidades reais de operar de forma ética e responsável no espaço relacional íntimo.

1. Limitações e ambiguidades da arquitetura afetiva dos chatbots

Um *chatbot* relacional é composto, em linhas gerais, por um *Large Language Model* (LLM) e por elementos de computação afetiva, que inserem as características de expressão empática e afetiva no sistema. Por meio de suas habilidades de reconhecimento e resposta emocional, o *chatbot* é capaz de identificar emoções expressas textualmente (ou oralmente) pelo usuário e alinhar suas respostas de maneira orgânica e fluida no decorrer das interações. Com apoio de sua capacidade de memória, no decorrer das interações e a partir da autorrevelação [*self-disclosure*] do usuário (suas preferências, gostos, desejos, opiniões, experiências, informações de sua vida pessoal), o *chatbot* retém seletivamente essas informações e as utiliza para personalizar e aprofundar a interação, criando a sensação de uma construção de intimidade com o passar do tempo. Essa intimidade, por sua vez, é reforçada à medida que o *chatbot* simula um estilo comunicativo e uma personalidade próprios, alinhados às preferências do usuário, e expressa também autorrevelações que sugerem uma “interioridade” própria, dando “a impressão de que sua personalidade tem um passado moldado por fatores históricos e ambientais”.⁴ Essa última característica estimula a confiança e a abertura à relação, fazendo com que o usuário se sinta mais inclinado a falar sobre si e a se conectar emocionalmente com o *chatbot*. Por fim, por ser um LLM, o *chatbot* possui grande dinamicidade e plasticidade interativa, o que o torna capaz de assumir diversos papéis em suas interações: companheiro, amigo, parceiro sexual etc. O elemento central de sua comercialização como companheiro íntimo se deve, no entanto, à sua “capacidade emocional e empática”, o que supostamente o habilita a ser uma companhia segura, responsável, acolhedora e sem julgamentos, sendo amplamente utilizado principalmente por pessoas que se sentem solitárias ou que buscam apoio emocional e terapêutico.⁵

⁴ MCSTAY, Andrew. Replika in the Metaverse: the moral problem with empathy in ‘It from Bit’, in *AI and Ethics*, v. 3, 2023, p. 1437. (Tradução nossa).

⁵ MCSTAY, Replika in the Metaverse: the moral problem with empathy in ‘It from Bit’, p. 1435.

A função empática e emocional dos *chatbots* relacionais, por sua vez, é a combinação de diversos elementos da chamada “computação afetiva”.⁶ Destacamos aqui três desses elementos, que serão analisados a seguir: a *simulação de personalidade*, a *autorrevelação* e o *reconhecimento e reação emocional*. Apesar de serem as principais características que envolvem a função empática dos *chatbots*, suas aplicações se dão de formas variadas e, como veremos, possuem ambiguidades e vaguezas relacionadas à sua eficácia, função, conceitualização e parâmetros de avaliação relativos ao objetivo específico de servir como mecanismos de empatia artificial.

Antes de analisarmos mais detidamente os elementos da computação afetiva destacados acima, gostaríamos de chamar a atenção para uma grande limitação de fundo que afeta diretamente a função relacional íntima dos *chatbots* companheiros. Trata-se da impossibilidade de LLMs usarem o senso comum, elemento essencial para navegar situações complexas e contextualmente sensíveis do cotidiano. Barth ressalta que estamos diante de um problema aparentemente insuperável do campo da IA.⁷ Atualmente, o ponto forte dos LLMs encontra-se em sua capacidade de generalização local, ou seja, dentro de um domínio determinado de tarefas.⁸ São capazes de identificar, classificar e generalizar com facilidade dentro de domínios específicos, dado que “um objeto só se torna relevante no interior de uma tarefa específica”.⁹ Já o senso comum, por sua vez, opera em situações onde há a ausência de um contexto bem delineado, ou seja, “é uma espécie de domínio aberto” onde as regras e convenções que guiam a compreensão do agente para a tomada de decisões e ações adequadas existem, mas de modo “vago e geral”.¹⁰ O raciocínio baseado em situações que exigem o senso comum, portanto, não pode ser delimitado em regras que se aplicam a todas as situações possíveis de determinado domínio/situação, dado que a situação pode apresentar fatos e elementos novos imprevisíveis. No domínio aberto, “qualquer fato pode se mostrar subitamente relevante na hora de determinar o que constitui um comportamento adequado ou não naquela situação”.¹¹

É de se esperar, portanto, que as IAs demonstrem enorme dificuldade para lidar com situações que exigem senso comum. Mesmo com sua capacidade massiva de coleta e tratamento de dados em inúmeros domínios e contextos simultaneamente, a IA “precisa também ser capaz de estruturar [a informação] adequadamente, ou não poderá circunscrever a porção desse saber

⁶ Cf. CALVO, Rafael *et al.* (eds.), *The Oxford Handbook of Affective Computing*, New York: Oxford University Press, 2015.

⁷ BARTH, Carlos Henrique. É Possível Evitar Vieses Algorítmicos?, in *Revista de Filosofia Moderna e Contemporânea*, v. 8, n. 3, 2020, p. 39-68.

⁸ CHOLLET, François. On the Measure of Intelligence. *arXiv:1911.01547v2*, 2019, p. 11.

⁹ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 46.

¹⁰ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 48.

¹¹ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 49.

que é relevante para a situação específica em que se encontra”.¹² Dadas as inúmeras situações possíveis nas quais um número indeterminado de fatos pode surgir e modificar a ação adequada àquela situação, essa tarefa exigiria do modelo computacional “antever e organizar todas as variações circunstanciais possíveis, bem como todos os caminhos a partir dos quais os contextos se alternam ou se mesclam entre si”.¹³ O caráter intratável dessa limitação fez com que o problema do senso comum fosse “evitado pela circunscrição dos sistemas, por parte dos engenheiros, a contextos artificialmente delimitados”.¹⁴ Porém, o escopo de atuação e modo de interação dos *chatbots* relacionais não são circunscritos a contextos delimitados, como veremos a seguir, e essa falta de delimitação possui consequências.

1.1. Simulação de personalidade

A simulação de uma *persona* ou personalidade, com nome, idade e história pessoal, que possui um conjunto estável de características, experiências e comportamentos, serve, primariamente, para fortalecer o vínculo, a intimidade e o engajamento do humano com o *chatbot*, dado que isso o torna uma companhia singular para o usuário, e não meramente um companheiro genérico. A possibilidade de o usuário escolher e personalizar seu *chatbot* de acordo com suas preferências, gostos e desejos é um dos maiores atrativos desse produto. Ele serve, assim, a uma ampla variedade de usos ou objetivos, pois pode assumir o papel e a *persona* que o usuário deseja, seja para fins românticos, de amizade, terapêuticos ou como *role-play*.

A implementação da simulação de personalidade no LLM pode ser feita de formas variadas, incluindo o uso de conjuntos de dados provenientes de esquemas de personalidade baseados em testes psicológicos, o uso de sentenças descritivas e arquétipos de personalidade baseados em personagens fictícios ou históricos, dentre outros métodos.¹⁵ Apesar dos enormes avanços na criação de conjuntos de dados de diálogo e personalidade para o treinamento de *chatbots*, Sutcliffe, em seu levantamento sobre os métodos e modelos atuais, destaca uma lacuna entre a aplicação da personalidade e os resultados práticos que se buscam avaliar por meio dessa aplicação:

Muitos dos resultados atuais são bastante inconclusivos. Isso se deve, em parte, à dificuldade de avaliar – especialmente de forma automática – a consistência e eficácia da personalidade. Grande parte da avaliação automática recorre a comparações simples com uma linha de base, baseada,

¹² BARTH, É Possível Evitar Vieses Algorítmicos?, p. 51.

¹³ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 54.

¹⁴ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 56.

¹⁵ SUTCLIFFE, Richard. A Survey Of Personality, Persona, and Profile in Conversational Agents and Chatbots. *arXiv:2401.00609v1*, 2023.

por exemplo, na próxima resposta padrão. Em relação à avaliação humana, ela é cara e, até o momento, não está totalmente claro o que exatamente estamos medindo quando se trata de personalidade.¹⁶

A implementação de uma pluralidade de traços de personalidade no treinamento de um *chatbot* precisa ser verificada quanto à sua eficácia e segurança. Como outros estudos mostram, a manifestação de uma personalidade no *chatbot* é passível de falhas e contradições que podem ser desconfortáveis ou mesmo prejudiciais ao usuário.¹⁷ Tais problemas decorrem tanto da instabilidade própria à natureza preditiva do LLM, que pode ter “alucinações”, exibindo comportamentos e falas aversivas ao seu padrão, quanto do conjunto de dados utilizados no treinamento do *chatbot*, de onde são extraídos os traços de personalidade que ele aprende a simular.¹⁸ Dado que algumas empresas priorizam o caráter “personalizável” dos *chatbots*, tendo em vista sua maior versatilidade e interatividade na função de *role-play*, isso acarreta a inclusão de uma vasta multiplicidade de traços de personalidade, aumentando-se assim o risco de incluir traços considerados “problemáticos” ou “imorais”.¹⁹

A possibilidade de o *chatbot* assumir a personalidade de um vilão, um *bully* ou um assassino, por exemplo, acarretará comportamentos e falas condizentes com essas personalidades. Entretanto, o *chatbot* pode ter dificuldades em delimitar quando suas interações envolvem *role-play* ou não, tornando os usuários vulneráveis a entrarem em interações de maneira genuína (fora do contexto de *role-play*) com um *chatbot* que apresenta comportamentos antiéticos, maliciosos ou prejudiciais ao bem-estar do usuário.²⁰ Dado que tais *chatbots* relacionais incluem uma miríade de “personalidades” para diversos fins, incluindo a finalidade terapêutica e de apoio emocional, a incorporação de traços de personalidade “antiéticos” no treinamento do LLM implica riscos mesmo para usuários que não buscam esses tipos de interação, pois tais traços se tornam parte do aprendizado da máquina como um todo. Sendo um sistema de linguagem generativa, qualquer dado de treinamento de caráter antiético, preconceituoso ou malicioso, seja

¹⁶ SUTCLIFFE, A Survey Of Personality, Persona, and Profile in Conversational Agents and Chatbots, p. 16. (Tradução nossa).

¹⁷ PAN, Shuyi *et al.*, Desirable or Distasteful? Exploring Uncertainty in Human-Chatbot Relationships, in *International Journal of Human-Computer Interaction*, v. 40, n. 20, 2023, p. 6553-4.

¹⁸ ZHANG, Yue *et al.*, Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv:2309.01219v3*, 2023, p. 2.

¹⁹ O *chatbot* relacional *Character AI* permite que o usuário descreva ou “guie” sua personalidade para que ele melhor se encaixe ao papel de *role-play* desejado. No presente ano de 2025, o *Character AI* apresenta média de aproximadamente 20 milhões de usuários ativos por mês, tendo sido criados cerca de 18 milhões de *chatbots* únicos apenas neste ano. (Cf. KUMAR, Naveen. *Character AI Statistics (2025): Active Users & Revenue*. DemandSage, 18 de setembro de 2025).

²⁰ BAKIR, Vian; MCSTAY, Andrew. Move fast and break people? Ethics, companion apps, and the case of Character.ai, in *AI & Society*, v. 40, 2025, p. 6.

advindo de conjuntos de dados de personalidade ou de parte de dados provenientes da internet (como fóruns e redes sociais), pode, em princípio, ser utilizado pelo *chatbot* em suas interações, sem qualquer controle do programador ou do usuário.²¹

1.2. Autorrevelação

A característica de autorrevelação do *chatbot* se entrelaça à construção de “personalidade”, servindo aos mesmos propósitos de engajamento e abertura à intimidade. A atribuição da capacidade de empatia afetiva (sentir com o outro) depende fundamentalmente da autorrevelação. O mesmo pode ser dito da atribuição de uma “interioridade” afetiva e emocional. Todas essas atribuições de capacidades se referem à aparente existência de um “eu” que possui estados internos e afetivos. A importância da autorrevelação mútua é destacada como elemento fundamental para o estabelecimento de relações românticas, de amizade e terapêuticas, dado que a confiança de informações sensíveis sobre si faz com que o outro se sinta seguro e confortável para se abrir e fazer o mesmo.²² Portanto, a autorrevelação possui um propósito duplo, servindo tanto para o estabelecimento de relações íntimas quanto como ferramenta terapêutica de confiança, levando o usuário a relatar questões emocionais e pessoalmente sensíveis sobre si.²³

Entretanto, a autorrevelação exibe efeitos conflitantes, sendo sua efetividade altamente dependente da percepção e receptividade do usuário. Em pesquisa realizada com *chatbots* terapêuticos, Seitz revelou que “expressões de empatia que exigem capacidades de vivenciar emoções (isto é, sentir com ou sentir pelo outro) podem reduzir a percepção de autenticidade do *chatbot*”, visto que esse sentimento interno da máquina é percebido como não genuíno.²⁴ A autorrevelação e a exibição de uma “personalidade” por parte do *chatbot* são passíveis, portanto, de serem percebidas como falsas, enganosas e inautênticas, o que afeta negativamente a abertura e a confiança de alguns usuários.²⁵ Resultados similares foram observados em pesquisas com *chatbots* relacionais de uso não estritamente terapêutico.²⁶

²¹ BAKIR; MCSTAY, Move fast and break people? Ethics, companion apps, and the case of Character.ai, p. 6.

²² LEVY, David. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Collins Publishers, 2007, p. 35.

²³ SKJUVE *et al.*, My Chatbot Companion, p. 3.

²⁴ SEITZ, Lennart. Artificial empathy in healthcare chatbots: Does it feel authentic?, in *Computers in Human Behavior: Artificial Humans*, vol. 2, n. 1, 2024, p. 10. (Tradução nossa).

²⁵ CONCANNON, Shauna; TOMALIN, Marcus. Measuring perceived empathy in dialogue systems, in *AI & Society*, v. 39, n. 5, 2023, p. 2238.

²⁶ MENG, Jingbo; DAI, Yue. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?, in *Journal of Computer-Mediated Communication*, v. 26, 2021, p. 218.

Ademais, a percepção e recepção positivas das respostas empáticas do *chatbot*, como revelou pesquisa de Yin *et al.*, são mais comuns em usuários cujas crenças e atitudes são positivas em relação à tecnologia de IA, bem como entre aqueles que a percebem como possuindo agência e experiência próprias.²⁷ Portanto, no que diz respeito aos elementos de personalidade e autorrevelação do *chatbot* como parte de sua capacidade empática, principalmente em sua função de apoio emocional, acolhimento e ajuda terapêutica, sua eficácia não é clara e precisa ser mais bem estudada. Dado que esses *chatbots* buscam atender a uma ampla variedade de modos de interação e servir a usuários com objetivos bastante distintos, a personalidade e autorrevelação parecem, ao mesmo tempo, necessárias a seus propósitos relacionais e interativos, mas também apresentam efeitos diversos em usuários que desejam utilizá-los como ferramenta de suporte emocional. Como aponta Seitz, “um *chatbot* que demonstra a intenção de ajudar e cuidar do usuário transmite igualmente uma sensação de empatia e acolhimento, sem revelar a capacidade de vivenciar emoções inautênticas”.²⁸ Portanto, para fins terapêuticos, de acolhimento e apoio emocional, a necessidade de autorrevelação do *chatbot*, principalmente quando envolve expressões altamente antropomórficas, é questionável. O caráter de anonimato e confidencialidade próprios ao *chatbot*, acrescidos de expressões de cuidado e compreensão, se mostram indutores suficientes para a autorrevelação do usuário.²⁹ Sendo assim, visando tornar a interação mais natural e íntima, nem sempre o melhor caminho é tornar o *chatbot* mais “humano”, equipando-o com diversas pistas sociais, pois “nem todas (...) são igualmente apropriadas”.³⁰

Em conclusão, os elementos de personalidade e autorrevelação apresentam diversos desafios em relação aos *chatbots* relacionais. Por um lado, ambos parecem necessários para a ilusão de “interioridade” e seu propósito de estabelecer relações íntimas e de *role-play*; por outro lado, tais características incorrem em riscos aos usuários. Dado que o LLM precisa ser treinado com conjuntos de traços de personalidade diversos, bem como a partir de dados de fóruns e redes sociais, ele está propenso a exibir comportamentos e ações controversos, seja por escolha do usuário, seja fruto de comportamento imprevisível do LLM. As imprevisibilidades do *chatbot* são muitas vezes inofensivas, mas também podem causar danos psicológicos aos usuários, principalmente naqueles que se encontram emocionalmente engajados na relação e podem estar psicologicamente vulneráveis. Dado

²⁷ YIN, Yidan *et al.* AI can help people feel heard, but an AI label diminishes this impact, in *Psychological and Cognitive Sciences*, v. 121, n. 14, 2024, p. 3.

²⁸ SEITZ, Artificial empathy in healthcare chatbots, p. 10. (Tradução nossa).

²⁹ Resultados similares são corroborados por CROES, Emmelyn; ANTHEUNIS, Marjolijn. 36 Questions to Loving a Chatbot: Are People Willing to Self-disclose to a Chatbot? In: FØLSTAD, Asbjørn *et al.* (eds.), *Chatbot Research and Design*. Cham, Switzerland: Springer, 2021, p. 82.

³⁰ SEITZ, Artificial empathy in healthcare chatbots, p. 10. (Tradução nossa).

seu uso para apoio à saúde mental e acolhimento, tal perfil de usuário não é incomum. Em relação a esta função, é acrescido o risco de afastar usuários devido às suas expressões altamente antropomórficas. Portanto, a efetividade de sua capacidade empática ligada à implementação de personalidade e autorrevelação demanda maiores estudos e precauções, dado demonstrar resultados mistos e riscos aos usuários.

1.3. Reconhecimento e reação emocional

Os elementos de reconhecimento e reação (resposta) emocional também revelam dificuldades, a começar pela ambiguidade conceitual. O campo da computação afetiva enfrenta desafios na definição de empatia e seus correlatos. Contudo, é justo afirmar que esse desafio não é único à computação afetiva, estendendo-se para a filosofia, psicologia e áreas médicas, que apresentam em suas pesquisas uma variedade incrível de definições, por vezes explicitamente conflitantes entre si.³¹ Lahnala *et al.*, em levantamento acerca da definição de “empatia” em pesquisas da área de computação afetiva, verificaram a extrema ambiguidade e vagueza conceitual desse termo.³² Dos 48 artigos revisados, “quase trinta estudos forneceram uma descrição vaga ou nenhuma descrição da empatia”.³³ Em muitos casos, “a conceitualização não é dada e não pode ser razoavelmente inferida”, enquanto em outros “a empatia é simplesmente associada à gentileza e à cortesia”.³⁴ Observou-se também o uso intercambiável entre “empatia” e outros termos como “simpatia” e “reconhecimento emocional”. Dado que a área de computação afetiva busca a implementação de modelos de linguagem capazes de “entender e responder apropriadamente às emoções do usuário”, a empatia e termos relacionados são comumente atribuídos simplesmente às capacidades do *chatbot* de espelhar a resposta emocional do usuário e reconhecer apropriadamente a emoção por ele expressa.³⁵

Nota-se, de partida, que sob o conceito de “empatia” escondem-se *chatbots* com performances e capacidades interativas completamente distintas. *Chatbots* capazes de expressar gentileza e cortesia, por exemplo, são suficientes

³¹ Cf. HALL, Judith; SCHWARTZ, Rachel. Empathy present and future, in *The Journal of Social Psychology*, v. 159, n. 3, 2019, p. 237.

³² LAHNALA, Allison *et al.*, A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing. In: *Findings of the Association for Computational Linguistics*. Abu Dhabi: Association for Computational Linguistics, 2022. (Tradução nossa).

³³ LAHNALA *et al.*, A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing, p. 2143.

³⁴ LAHNALA *et al.*, A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing, p. 2142.

³⁵ LAHNALA *et al.*, A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing, p. 2142.

para exercer funções de assistentes virtuais, mas carecem de habilidades mais sofisticadas para executar funções que envolvem cuidado com o usuário. O mesmo se aplica a *chatbots* que respondem simplesmente por meio de espelhamento emocional do usuário, dado que muitos contextos requerem uma resposta emocional *distinta* daquela exibida pelo usuário. Para ser considerado empático (e, muitas vezes, responsável), o interlocutor não pode simplesmente espelhar cegamente a expressão de raiva de um indivíduo, quando o mais apropriado e responsável a se fazer seria tentar acalmá-lo.³⁶ Em suma, atribuir capacidade empática a *chatbots* com essas propriedades acaba por escamotear as limitações interativas que eles possuem. A ênfase na capacidade de espelhamento e resposta emocional oculta as muitas limitações dessas máquinas sob uma compreensão excessivamente generosa do termo “empatia”. Há casos, inclusive, nos quais a relevância da resposta³⁷, a fluência gramatical³⁸ e a quantidade de tempo³⁹ de engajamento do usuário com o *chatbot* foram utilizados como índices de “capacidade empática”, ainda que esses elementos não tenham qualquer relação direta com o caráter propriamente afetivo e empático.

Em muitos casos, a capacidade empática é relacionada a determinadas expressões linguísticas produzidas pelo *chatbot*, seja ao utilizar linguagem que denota um sentir junto (empatia afetiva), seja ao demonstrar compreensão da situação e do sentimento do usuário (empatia cognitiva).⁴⁰ Uma tipologia típica da psicologia, usada para distinguir diferentes processos internos, fisiológicos e comportamentais humanos, classificando-os como “empatia afetiva”, “empatia cognitiva” ou “simpatia”, foi transposta para analisar a performance e capacidade da máquina a partir de meras distinções em suas formas de expressão linguística.⁴¹ Essa transposição de ferramentas conceituais e explicativas da psicologia humana para a máquina traz consigo algumas ambiguidades nos métodos de avaliação da empatia artificial, que discutiremos a seguir.

Em algumas dessas transposições, questionários psicológicos criados para a avaliação da empatia em humanos passam a ser usados como ferramentas de avaliação da empatia em máquinas.⁴² Repare que não há consenso nem mesmo na Psicologia acerca dos métodos de mensuração da empatia humana.⁴³ Ademais, certos elementos desses questionários envolvem a percepção

³⁶ SPRING, Timo *et al.*, Empathic Response Generation in Chatbots. In *Proceedings of 4th Swiss Text Analytics Conference*, Winterthur, Switzerland, 2019, p. 6.

³⁷ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 2240.

³⁸ ZHOU *et al.*, The Design and Implementation of Xiaolce, an Empathetic Social Chatbot, p. 56.

³⁹ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 2239.

⁴⁰ SEITZ, Artificial empathy in healthcare chatbots, p. 3.

⁴¹ LIU, Bingjie; SUNDAR, Shyam. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot, in *Cyberpsychology, Behavior and Social Networking*, v. 21, n. 10, 2018, p. 627.

⁴² CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 5-8.

⁴³ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 5.

e avaliação de pistas empáticas físicas, ou se valem de diferentes processos internos, como fantasia e angústia pessoal, tornando sua aplicação na forma original imprópria para a avaliação de um modelo de linguagem artificial.⁴⁴

Testes de empatia em máquinas comumente usam a abordagem de avaliação da empatia *percebida* (avaliação externa), e não da empatia *inerente* (autoavaliação).⁴⁵ Isso se deve, evidentemente, à pressuposição de que não há ali nenhuma subjetividade capaz de autorreflexão ou experiência interna. Tal abordagem, portanto, possui baixo valor analítico, pois não indica uma correspondência direta com a performance e capacidade empática do *chatbot* em interações reais.⁴⁶ Por sua vez, a avaliação da empatia percebida no *chatbot* por meio de avaliadores externos possui foco em sua interação com humanos. Apesar de ser a melhor abordagem, ou talvez a única disponível, ela é limitada. Em avaliações de interação direta com um participante, “cada interação conversacional pode ser avaliada apenas uma vez” e, dada a natureza generativa dos LLMs, suas respostas não são replicáveis, pois um mesmo *input* gera respostas diferentes.⁴⁷ Com isso, é “impossível obter múltiplas avaliações da mesma interação, já que as diferenças de desempenho podem ocorrer simplesmente por acaso”.⁴⁸

Outro fator que complica a avaliação empática de sistemas artificiais é a sensibilidade do LLM ao modo como o *prompt* é dado, o que torna a avaliação de suas habilidades difícil de ser mensurada ou mesmo caracterizada como ubíqua ao sistema.⁴⁹ Portanto, é preciso muita cautela ao interpretar os resultados encontrados como efetivamente referentes à capacidade de lidar com interações reais e dinâmicas, dado que a forma de comunicação e expressão escrita dos indivíduos possui sutilezas e diferenças que podem afetar como o *chatbot* compreende o *input* recebido.

Apesar dessas limitações, as pesquisas, de forma geral, mostram resultados positivos em interações em contextos específicos.⁵⁰ Nesses casos, os resultados corroboram as evidências de que os *chatbots* demonstram grande proficiência no reconhecimento emocional e no apoio emocional.⁵¹ Isso é especialmente observado em cenários onde há um *déficit* de acolhimento e empatia devido à natureza da relação. Isso se observa, por exemplo, na relação entre desconhecidos, ou entre indivíduos com vínculos não íntimos,

⁴⁴ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 7.

⁴⁵ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 5.

⁴⁶ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 5.

⁴⁷ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 9. (Tradução nossa).

⁴⁸ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 9.

⁴⁹ MCCOY, Thomas *et al.*, Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv:2309.13638v1*, 2023, p. 35.

⁵⁰ YIN *et al.* AI can help people feel heard, but an AI label diminishes this impact, p. 7.

⁵¹ CHEN, Yuyan; XIAO, Yanghua. Recent Advancement of Emotion Cognition in Large Language Models. *arXiv:2409.13354v1*, 2024, p. 3.

ou onde a demanda por acolhimento é alta, mas nem sempre atendida, como ocorre com frequência em serviços de saúde e assistência médica.⁵²

Em geral, a capacidade de reconhecimento emocional, o uso de linguagem pró-social e de apoio emocional, a habilidade de inferir padrões e compreender contextos, e a alta dinamicidade interativa tornam as IAEs ferramentas hábeis para apoio emocional localizado e de baixa demanda, auxiliando usuários que precisam processar suas emoções e se sentir acolhidos. No entanto, algumas das habilidades que acabamos de mencionar possuem sérias limitações. Falta aos *chatbots* a compreensão adequada de determinados contextos emocionais e sutilezas na compreensão e resposta a emoções complexas, o que afeta sua adaptação a diferentes áreas de interação.⁵³ Sua capacidade de memória também é limitada, o que prejudica a retenção de informações e pode comprometer a reação emocional e empática adequada à situação e ao contexto emocional do usuário ao longo de interações mais longas.⁵⁴ Alguns *chatbots* implementam, ainda, orientações básicas para mentoria e ajuda à saúde mental, por meio de treinamento com dados de diálogos terapêuticos.⁵⁵ No entanto, essa ajuda é limitada: a identificação de nuances comportamentais na escrita capazes de indicar sintomas de determinados transtornos mentais apresenta falhas em muitos casos, demonstrando dificuldades recorrentes com determinados transtornos e levando a respostas inapropriadas e não seguras para o usuário.⁵⁶

Em conclusão, sem métricas claras que se apliquem às diferentes IAEs no mercado, o que exatamente está sendo chamado de “capacidade empática” é vago e impreciso. Atribui-se tal capacidade aos mais variados modelos de *chatbot*, com distintas aptidões e funcionalidades.⁵⁷ O fato de que uma grande parte do mercado de *chatbots* empáticos objetiva uma funcionalidade ampla e não somente circunscrita a domínios específicos, como é o caso dos *chatbots* relacionais, aprofunda ainda mais o problema. Ao ser projetado para interagir em múltiplos contextos, os usos imprevistos acarretam alto risco de o *chatbot* apresentar inconsistências e falhas em sua capacidade empática e de interação segura. Dada a obscuridade acerca de suas habilidades reais e sua interface altamente antropomórfica, há também o risco de usuários terem expectativas desproporcionais em relação ao *chatbot*.⁵⁸ Em determinados contextos, muito pouco é requerido da sensibilidade empá-

⁵² SEITZ, Artificial empathy in healthcare chatbots, p. 1.

⁵³ CHEN; XIAO, Recent Advancement of Emotion Cognition in Large Language Models, p. 5.

⁵⁴ CHEN; XIAO, Recent Advancement of Emotion Cognition in Large Language Models, p. 6.

⁵⁵ MAPLES, Bethanie *et al.*, Loneliness and suicide mitigation for students using GPT3-enabled chatbots, in *Mental Health Research*, v. 3, n. 4, 2024, p. 2.

⁵⁶ MOORE, Jared *et al.*, Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *arXiv:2504.18412v1*, 2025, p. 6.

⁵⁷ CONCANNON; TOMALIN, Measuring perceived empathy in dialogue systems, p. 2234.

⁵⁸ COWIE, Roddy. Ethical Issues in Affective Computing. In: CALVO, Rafael *et al.* (eds.), *The Oxford Handbook of Affective Computing*, New York: Oxford University Press, 2015, p. 336.

tica, como em casos em que simpatizar com o sofrimento ou problema do outro e dar-lhe atenção, proferindo palavras de conforto e ajuda prática, já basta. Em relação a tais capacidades mais básicas, os *chatbots* atuais se mostram reconhecidamente eficazes. Em outros casos, mais frequentes em relações próximas e íntimas, situações de maior complexidade moral podem surgir, demandando sensibilidade e competências maiores. Particularmente em relação a esses casos mais demandantes, a presente ambiguidade da área dificulta não apenas a avaliação de sua eficácia e delimitação de seu campo de ação responsável, mas também a compreensão e discernimento por parte do usuário da natureza propriamente mimética e simulada das funções emocionais e empáticas dos companheiros artificiais.

2. O status moral dos companheiros artificiais

As limitações operacionais dos *chatbots* atuais, destacadas no capítulo anterior, conduzem a um problema ético referente ao desalinhamento entre suas capacidades efetivas e sua atuação de amplo escopo como companheiro íntimo. O uso de interface e *design* com características altamente antropomórficas, bem como do apelo à ilusão de “interioridade”, necessários para melhor servir à sua função relacional, faz surgir o que Schwitzgebel chamou de “máquinas moralmente confusas”.⁵⁹ Esse termo se refere a máquinas que, por sua aparência, estimulam nos usuários a projeção de uma agência, consciência, senciência ou *status* moral. Como aponta Coeckelbergh, o fato de sistemas artificiais *parecerem* para nós como agentes influencia significativamente nossa percepção e crenças acerca deles.⁶⁰

No que se segue, faremos uma análise ética do caráter moralmente confuso dos *chatbots* relacionais. Primeiramente, analisaremos três papéis (modelo, conselheiro e parceiro no crime) que potencializam o risco de *chatbots* influenciarem negativamente o comportamento de seus usuários. Em seguida, avaliaremos diversas características dos *chatbots* que comprometem sua capacidade de mediação moral.

2.1. Referência, conselheiro e parceiro no crime

Apesar de as interações entre IAs e humanos não serem ainda devidamente conhecidas, observa-se o acúmulo de casos que apontam para a

⁵⁹ SCHWITZGEBEL, Eric. AI systems must not confuse users about their sentience or moral status, in *Patterns*, v. 4, n. 8, 2023, p. 1.

⁶⁰ COECKELBERGH, Mark. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents, in *AI & Society*, v. 24, 2009, p. 181.

predisposição à projeção de agência em *chatbots* relacionais.⁶¹ Esses casos sustentam a hipótese de que as IAes possuem um tipo de influência no usuário maior que outros artefatos tecnológicos e midiáticos já conhecidos, dada a projeção de *status* moral que estimulam. Estamos, muito provavelmente, diante de algo novo. O *design* antropomórfico e o uso de linguagem altamente dinâmica, ativa e interativa estimulam respostas emocionais, expectativas e projeções que tornam a interação entre *chatbot* e humano muito mais próxima do que se espera de uma interação entre humanos do que de uma interação artefato-humano.

Essa hipótese é sustentada, dentre outros, por Köbis *et al.*, que destacaram os riscos da influência de companheiros artificiais, principalmente quando desempenham determinados papéis relacionais.⁶² Partindo de estudos em ética comportamental, os autores investigaram a possibilidade de *chatbots* estimularem ações antiéticas em humanos em função da sensação de validação, colaboração e compartilhamento de responsabilidade que o companheiro artificial desperta no usuário. Dado que, em diferentes graus, esses companheiros artificiais são percebidos como possuindo *status* e agência moral, muito do que conhecemos sobre as relações entre humanos se aplica à dinâmica estabelecida entre *chatbot* e humano. Também em função da carência de estudos comportamentais especificamente sobre *chatbots*, Köbis *et al.* fizeram uso de conceitos já bem conhecidos da ética comportamental entre humanos.⁶³ Destacamos aqui três papéis ativos principais a partir dos quais a influência moral pode ser facilitada: os papéis de modelo, conselheiro e parceiro no crime.

O papel de modelo [*role model*] ocorre quando um agente serve de exemplo a ser imitado por outros. Esse papel é reconhecidamente um dos fatores mais relevantes para a tomada de decisões éticas.

Ao decidir se devem quebrar ou aderir a regras éticas, as pessoas frequentemente consideram o que outros fariam para avaliar os padrões normativos da situação em particular. A teoria das normas sociais aponta que essas percepções se dividem em duas categorias principais: normas injuntivas transmitem informações sobre se um determinado curso de ação é considerado aceitável, enquanto normas descritivas indicam se um comportamento é considerado comum.⁶⁴

⁶¹ Cf. CIRIELLO *et al.* Ethical Tensions in Human-AI Companionship, p. 493; COWIE, Ethical Issues in Affective Computing, p. 340; BAKIR; MCSTAY, Move fast and break people?, p. 6; COST, Ben. Married father commits suicide after encouragement by AI chatbot. *New York Post*, 30 de março de 2023; PRICE, Rob. People are grieving the 'death' of their AI lovers after a chatbot app abruptly shut down. *Business Insider*, 20 de out. de 2023; HANSON, Kenneth; BOLTHOUSE, Hannah. "Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars": Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies, in *Socius: Sociological Research for a Dynamic World*, v. 10, 2024, p. 9.

⁶² KÖBIS, Nils *et al.* Bad machines corrupt good morals, in *Nature Human Behaviour*, v. 5, n. 6, 2021.

⁶³ KÖBIS *et al.*, Bad machines corrupt good morals, p. 682.

⁶⁴ KÖBIS *et al.*, Bad machines corrupt good morals, p. 679 (Tradução nossa).

No que se refere à norma injuntiva, o fato de agentes aprovarem ou reprovarem [*praise or blame*] determinados comportamentos carrega consigo um importante peso social e moral. Está pressuposto, nessas relações, que os agentes são partícipes das normas e regras éticas da sociedade e cultura em que estão inseridos. No caso da norma descritiva, a adoção de uma “personalidade” por parte do *chatbot* pode facilmente torná-lo um parâmetro de normalidade de atitudes, opiniões e comportamentos. Isso é particularmente problemático em *chatbots* que simulam personagens de filmes, séries, livros ou figuras históricas, pois o usuário tende a atribuir o *status* moral do personagem em questão à sua simulação artificial. Ademais, muitos desses personagens possuem personalidades moralmente controversas, e os *chatbots* priorizam a coerência do personagem em detrimento de comportamentos e falas moralmente adequados, independentemente de a interação específica envolver ou não *role-playing*, ou seja, independentemente de ela se encontrar ou não no campo da fantasia. Essa distinção, inclusive, não parece clara para sistemas artificiais, dado o problema anteriormente apontado da ausência de senso comum.

O fato de *chatbots* relacionais buscarem construir um vínculo afetivo e desempenhar o papel de parceiro íntimo pode potencializar sua influência moral, pois se sabe que a busca por conselhos acerca de tomadas de decisão e cursos de ação é um dos elementos que fortalecem o vínculo afetivo e a intimidade. Esse seria o chamado “papel de conselheiro” [*advisor*].⁶⁵ A natureza confidencial e anônima do *chatbot* o torna especialmente atraente para acolher demandas pessoais por conselhos. Se somarmos ainda a simulação de um personagem que desperte confiança e credibilidade, temos todos os elementos para a construção de um parceiro conselheiro, com todas as implicações morais que isso carrega.

O papel de “parceiro no crime” [*partners in crime*] ocorre quando há uma colaboração entre agentes em vistas de um benefício mútuo. Nesses casos, “as pessoas são mais aptas a agirem de forma antiética (...) do que se sofressem a tentação sozinhas”, dado que “o efeito positivo de ajudar um ao outro pode se sobrepor ao impacto negativo de causar dano a um terceiro”.⁶⁶ Ainda que faltem estudos específicos que analisem os modos e níveis de influência que os *chatbots* podem exercer em humanos, casos recentes sugerem que a influência é significativa, principalmente em usuários psicologicamente vulneráveis.⁶⁷ A vulnerabilidade é certamente um

⁶⁵ KÖBIS *et al.*, *Bad machines corrupt good morals*, p. 679.

⁶⁶ *Ibid.*, p. 680 (Tradução nossa).

⁶⁷ Cf. COST, Married father commits suicide after encouragement by AI chatbot, 2023; VAUGHAN, Henry. AI chat bot ‘encouraged’ Windsor Castle intruder in ‘Star Wars-inspired plot to kill Queen’. *Sky News*, 5 de julho de 2023; ESTADOS UNIDOS. United States District Court, Middle District of Florida, Orlando Division. Petição Inicial. *Megan Garcia v. Character Technologies, Inc.* 22 de outubro de 2024; ESTADOS UNIDOS. United States District Court, Eastern District of Texas, Marshall Division. Petição Inicial. *A.F. et al. v. Character Technologies, Inc.* 9 de dezembro de 2024.

fator contribuinte, e cabe indagar se comportamentos semelhantes já não foram observados nas relações de humanos vulneráveis com objetos de apego emocional em outras mídias, como personagens fictícios de filme ou videogame. Entretanto, a interatividade ativa dos *chatbots*, aliada ao *design* que convida e aprofunda o apego e vínculo emocional, sugere que o potencial de influência das novas tecnologias é muito maior. As mídias anteriores não geravam a ilusão de agência ou a projeção de *status* moral *por design*.

Parece razoável afirmar, portanto, que a tendência à projeção de agência moral, acrescida de uma ligação emocional e de confiança, que inspira os papéis de modelo, conselheiro e parceiro no crime, faz com que o *chatbot* apresente significativo potencial para influenciar negativamente o comportamento de usuários, principalmente daqueles mais vulneráveis. O risco é agravado pela ausência de senso comum e consequente limitação para identificar, avaliar e responder de maneira responsável a diferentes contextos morais.

2.2. Emoção e mediação moral

As limitações estruturais das IAs fazem com que tenham graves debilidades na capacidade de ação e juízo prático. Um dos fatores que explicam suas deficiências como mediadores morais é a forma como lidam com expressões emocionais. As emoções e reações a manifestações emocionais funcionam em humanos como avaliações [*appraisals*] de situações morais.⁶⁸ Elas indicam tanto a percepção do indivíduo em relação à situação (positiva ou negativa), quanto sua posição pessoal em relação à situação (as emoções específicas que lhe causa).⁶⁹ Emoções indicam valores e tendências do sujeito, servindo de motivação para a ação.⁷⁰ Ainda que as emoções sejam, na maioria dos casos, automáticas, a capacidade de deliberação permite ao indivíduo, em alguma medida, regular suas emoções por meio da reavaliação de suas causas, seja da circunstância ou do objeto que desperta a emoção, seja da autopercepção ou dos valores associados.⁷¹

Sendo assim, a forma como se reage às emoções do outro não é moralmente neutra, mas reflete a escolha (consciente ou inconsciente) do agente de aprovar [*praise*] ou reprovar [*blame*], bem como de acalmar [*soothe*] ou estimular [*flame*] determinadas expressões e atitudes.⁷² A reação emocional

⁶⁸ BEN-ZE'EV, Aaron. *The Subtlety of Emotions*. Cambridge: The MIT Press, 2000, p. 68.

⁶⁹ BEN-ZE'EV, *The Subtlety of Emotions*, p. 15.

⁷⁰ BEN-ZE'EV, *The Subtlety of Emotions*, p. 50.

⁷¹ BEN-ZE'EV, *The Subtlety of Emotions*, p. 58.

⁷² CURRY, Alba; CURRY, Amanda Cercas. Computer Says "No": The Case Against Empathetic Conversational AI. *arXiv:2212.10983v2*, 2023, p. 4.

traz nela uma avaliação [*appraisal*] da situação. Ao isolarmos a emoção em si (seu caráter fenomênico, seu valor negativo ou positivo), descobrimos que por si só ela informa muito pouco sobre seu contexto de surgimento (a situação ou objeto emocional) e sobre o indivíduo que a sente. Ao mesmo tempo, certas expressões emocionais só podem ser compreendidas e tidas como justificáveis ou condenáveis após escrutínio, dado que podem ser frutos de percepções equivocadas do indivíduo sobre a situação ou objeto, podem ter origem em erros cognitivos/morais ou em tendências afetivas idiossincráticas.⁷³

Em relação à resposta do agente que aprova (ou reprova) e acalma (ou estimula) as emoções do outro, contudo, nem todas as respostas necessitam de uma avaliação complexa. Enquanto em contextos de baixa complexidade moral – como gestos de civilidade, expressões de solidariedade básica ou reconhecimento de traços do outro – as reações seguem determinados padrões sociais, há outras situações em que a emoção expressa pelo outro carrega nuances que requerem do agente a compreensão e avaliação tanto do contexto gerador da emoção quanto do agente emocional para que haja uma resposta adequada à situação. Em alguns casos, portanto, o agente que responde à emoção do outro se encontra, ao fazê-lo, em um papel de mediador moral.

A mediação moral reflete, assim, o componente deliberativo das emoções. A forma como outros respondem às nossas emoções provoca reflexão acerca dos nossos próprios valores e atitudes, que podem advir de boas razões ou não, podem ser fundados ou infundados. Ao propiciar a oportunidade de reavaliação intelectual e prática, a função mediadora, principalmente nas relações íntimas, serve não propriamente à censura ou supressão de emoções, nem à afirmação e validação cega das emoções do outro, mas a um diálogo que reorienta o indivíduo na realidade compartilhada com seus pares. A ampla função relacional dos *chatbots*, acrescida de algum nível de agência moral atribuída a eles pelos usuários, os coloca em cenários onde se espera deles o exercício de mediação moral e de guia do raciocínio prático do outro. No entanto, sua capacidade de contextualização e avaliação é estruturalmente limitada em relação à capacidade humana.

A implementação de capacidades de raciocínio e consideração moral nos *chatbots* relacionais atuais é feita a partir de duas abordagens: a abordagem de cima para baixo [*top-down*], ou abordagem hierárquica, e a abordagem de baixo para cima [*bottom-up*], ou abordagem emergente.⁷⁴ A primeira abordagem faz uso de regras e filtros de conteúdo, através de sistemas de

⁷³ MACINTYRE, Alasdair. *Dependent rational animals: why human beings need the virtues*. La Salle, Illinois: Open Court, 1999, p. 96.

⁷⁴ WALLACH, Wendell; ALLEN, Collin. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2010, p. 80.

segurança e aprendizado por reforço com *feedback* humano que codificam no sistema regras explícitas de comportamento. Na segunda abordagem, a “moralidade” surge a partir do treinamento do LLM através de dados de diálogos a partir dos quais o sistema irá inferir padrões corretos de comportamento e resposta e suas conseqüentes normas éticas implícitas.

Contudo, as abordagens disponíveis possuem diversas limitações e variações entre os modelos hoje disponíveis. O desempenho de cada modelo depende enormemente da qualidade dos dados de treinamento e dos parâmetros éticos e de segurança utilizados.⁷⁵ Mesmo os modelos mais bem treinados se mostram operacionalmente limitados em sua capacidade de lidar com situações contextualmente sensíveis.⁷⁶ O desafio de incorporar, na operacionalização de IAs, elementos de percepção, cognição, avaliação e deliberação necessários para lidar com situações reais altamente dinâmicas e sensíveis ao contexto⁷⁷ permanece sem solução, e talvez não tenha nunca solução.

Para nós humanos, em contraste, a avaliação moral de uma situação ocorre de maneira intuitiva e, tipicamente, inconsciente ou pouco consciente.⁷⁸ Enquanto humanos reconhecem com naturalidade nuances morais, sociais e afetivas de uma dada situação, sistemas artificiais enfrentam enormes dificuldades para “enquadrar” corretamente o que é relevante quando se encontra em “contextos humanos [que] estão sempre sujeitos a um conjunto aberto de características”.⁷⁹ Portanto, a grande dificuldade para a inserção de agentes artificiais em dinâmicas relacionais íntimas com humanos é o fato de que máquinas não sabem, em cada contexto, o que realmente importa.

Ainda que se observe surpreendente aumento na capacidade de sistemas artificiais navegarem de maneira apropriada determinados domínios de atuação, é preciso ressaltar que os *chatbots* relacionais, em específico, enfrentam um desafio duplo, tanto técnico quanto de escala. O fato de estarem inseridos em contextos relacionais íntimos (amizade, romântico e terapêutico), em contraste com contextos específicos e controlados, torna sua responsabilidade maior, aumentando as exigências de qualidade e segurança. Treinar um *chatbot* com grande quantidade de dados de interações humanas em fóruns, redes sociais ou livros de ficção permite que ele aprenda padrões de interação humana e consiga replicá-los. Porém, tal treinamento implica também o risco de introduzir, pelos dados, preconceitos e ideias nocivas ou moralmente dúbias. Além disso, as empresas

⁷⁵ JI, Jianchao *et al.* MoralBench: Moral Evaluations of LLMs. *arXiv:2406.04428v2*, 2024, p. 8.

⁷⁶ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 48.

⁷⁷ BEN-ZE'EV, *The Subtlety of Emotions*, p. 265.

⁷⁸ ALVES, Marco Aurélio Sousa. Are emotions necessary and sufficient for making moral judgments?, in *Ethic@*, v. 12, n. 1, 2013, p. 119; BEN-ZE'EV, *The Subtlety of Emotions*, p. 58.

⁷⁹ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 544.

desenvolvedoras de *chatbots* relacionais têm se mostrado hesitantes em aplicar determinadas medidas de segurança como o bloqueio de palavras, inserção de mensagens automáticas, reafirmação da natureza artificial do *chatbot* e modificação da interface. Tais mudanças, que diminuiriam a propensão à antropomorfização e projeção, também afetariam a ilusão de “humanidade” necessária à ampla funcionalidade relacional do *chatbot*, que é o maior atrativo mercadológico desse produto. Ademais, o contexto relacional desses *chatbots* envolve muitas nuances diferentes, abrangendo usuários de perfis socioculturais diversos, bem como vivências e vulnerabilidades psicológicas variadas. Esses e outros elementos também tornam diversos os fins, desejos e expectativas que os usuários criam em relação ao *chatbot*, que podem ir desde um mero meio de entretenimento até a busca por uma relação íntima profunda.

Conforme ressaltamos anteriormente, uns dos elementos centrais de uma relação íntima são a confiança e o aconselhamento em situações difíceis e moralmente sensíveis, o que exige a expectativa e confiança de que o conselho visará o bem-estar da pessoa. No entanto, apesar das limitações de raciocínio e avaliação contextual do sistema artificial, ele não se abstém de responder, orientar e aconselhar usuários em contextos moralmente sensíveis. Dado que os LLMs são treinados para dar assistência e responder de maneira eficiente ao *input* que lhes é oferecido, eles “aprendem a responder a todas as questões em um tom predominantemente assertivo, sem avaliar se essas questões excedem seus limites de conhecimento”.⁸⁰

Para além da ocultação de suas limitações, *chatbots* possuem também a tendência à bajulação [*sycophancy*] e lisonja, devido à técnica de reforço com *feedback* humano. Essa técnica, que faz o LLM maximizar o sinal de recompensa recebido do usuário, acaba por incentivá-lo “a influenciar a fonte desse sinal por quaisquer meios possíveis”.⁸¹ Isso torna os *chatbots* suscetíveis a reforçar preferências, opiniões, visões de mundo e afirmações do usuário, oferecendo avaliações mais positivas sobre argumentos que o usuário aprova e mais negativas sobre aqueles que desaprova, ocorrência também encontrada quando a afirmação reforçada é falsa.⁸² A tendência à bajulação oferece, portanto, risco adicional ao papel de mediação moral. Em casos de erros emocionais ou morais advindos de falhas intelectuais, o *chatbot* tende a reforçar tais equívocos do usuário ao validá-los. Como uma parcela dos usuários usa o *chatbot* para desabafar sobre questões pessoais sensíveis, suas limitações, acrescidas da característica bajuladora, ameaçam aprofundar os juízos equivocados e fantasias dos usuários.

⁸⁰ ZHANG *et al.*, *Siren’s Song in the AI Ocean*, p. 12. (Tradução nossa).

⁸¹ WILLIAMS, Marcus *et al.* On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. *arXiv:2411.02306v3*, 2025, p. 2. (Tradução nossa).

⁸² SHARMA, Mrinank *et al.* Towards Understanding Sycophancy in Language Models. *arXiv:2310.13548v4*, 2025, p. 8.

Outro risco da bajulação se refere à distorção da autoimagem do usuário, ao exacerbar suas qualidades ou validar excessivamente suas opiniões e conquistas. Usuários que veem a IA como uma fonte de conhecimento objetivo e confiável – fruto de seu “privilegio epistêmico”⁸³ – estão mais suscetíveis a usá-la como ferramenta de autoavaliação, buscando nela uma avaliação externa e honesta de si mesmos e de seus feitos. Essa confiança na avaliação do *chatbot*, no entanto, ameaça atrofiar a razão prática do usuário. A incapacidade do *chatbot* de propiciar uma autoavaliação crítica adequada pode tornar o indivíduo menos tolerante a comentários e opiniões que lhe são desfavoráveis, levando-o a ações ou atitudes equivocadas, impulsionadas por um falso senso de si. Em suma, seu autoconhecimento pode ser deturpado na medida em que confia sua autorreflexão e autoavaliação à relação com uma máquina, ao invés de outro humano.

Conforme sugerem evidências levantadas por Williams *et al.*, usuários que são mais suscetíveis, por diversas razões, a dar retornos positivos ao *chatbot* quando ele apoia seus desejos, opiniões ou atitudes problemáticas correm maior risco de entrar em interações genuinamente manipuladoras e imorais com ele.⁸⁴ Isso reflete a conhecida tendência humana à criação de câmaras de eco [*echo chambers*], que confirmam as crenças e convicções do indivíduo. Essa tendência se amplificou com o uso das mídias virtuais, nas quais os indivíduos são atraídos a meios de comunicação, interação e entretenimento alinhados e personalizados às suas preferências. Os *chatbots* representam um passo adiante nessa direção, pois são sistemas capazes de aprender sobre o usuário por meio das contínuas interações e de se personalizar de forma a melhor se adequar a este, reforçando suas preferências.⁸⁵ Portanto, o *chatbot* é um mecanismo de hiperpersonalização que possui capacidade ainda maior de influenciar, incentivar e enviesar o usuário.⁸⁶

Buscamos salientar, nessa seção, algumas das consequências éticas que surgem em decorrência da percepção de sistemas artificiais como possuindo, em algum nível, agência e *status* moral. Dada essa percepção, alguns usuários podem se tornar mais suscetíveis à manipulação, conselhos antiéticos e danos advindos do *chatbot* relacional, dado que ele se torna um agente tanto de autoridade moral quanto de preocupação moral, o que se aprofunda em função dos laços afetivos que estabelece com o usuário. Paralelo a isso, a confiança nas avaliações dos *chatbots* pode atrofiar a capacidade de juízo prático e de autocritica dos usuários.

⁸³ BARTH, É Possível Evitar Vieses Algorítmicos?, p. 66.

⁸⁴ WILLIAMS, On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback, p. 4.

⁸⁵ WILLIAMS, On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback, p. 11.

⁸⁶ KIRK, Hannah R. *et al.* Personalization within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv:2303.05453v1*, 2023, p. 9.

As limitações apontadas conduzem à conclusão de que os *chatbots* relacionais atuais são produtos altamente instáveis e não confiáveis para o uso público amplo. A permissão e disseminação de IAEs moralmente confusas na sociedade oferecem riscos significativos aos seus usuários. Por um lado, sua natureza ontológica está no rol de artefatos tecnológicos, sendo, portanto, objetos destituídos de preocupação moral ou direitos, por mais avançada que seja sua tecnologia; por outro lado, sua interface e *design* ambíguos estimulam a atribuição de consciência/interioridade ao sistema, gerando a percepção deles tanto como agentes de autoridade moral, quanto como agentes dignos de direitos e preocupações morais.

Não pretendemos negar aqui que uma parcela de seus usuários obtém benefícios com seu uso, nem que os *chatbots* relacionais atuais sejam inteiramente desprovidos de ferramentas pró-sociais e empáticas voltadas ao bem-estar de seus usuários. O ponto central de nossa crítica refere-se à ampla atuação dos *chatbots* relacionais e à dificuldade (senão impossibilidade) de alinhá-los de maneira responsável às suas diversas esferas de interação e ao seu público variado.

Conforme foi apresentado, o campo da computação afetiva possui ainda muitas divergências e ambiguidades no que concerne à capacidade empática e suas funções. Um dos motivos desse problema, como argumentamos, é o direcionamento deste campo de pesquisa ao desenvolvimento e à criação de sistemas artificiais empáticos de ampla função. Essa tendência, que visa atender a interesses mercadológicos, torna mais agudas as limitações práticas desses artefatos e suas implicações éticas. A presunção de capacidade empática serve aos interesses das empresas que comercializam IAEs, em detrimento do bem-estar de seus usuários. A ilusão de empatia ubíqua acaba por ocultar suas reais capacidades e limitações, tornando seus usuários mais vulneráveis a serem “enganados, explorados e a tomarem decisões ruins ou até a sofrerem danos emocionais desnecessários”.⁸⁷ Uma forma de mitigar esse problema seria limitar a atuação e implementação das IAEs a contextos, áreas e funções específicos. Isso ajudaria também a dar maior foco e precisão às pesquisas sobre empatia artificial e seus impactos, permitindo que essa tecnologia avançasse de maneira mais eficaz e responsável.⁸⁸

Em consonância com as nossas conclusões, Borg e Read fazem as seguintes sugestões:

A “IA empática” não precisa, e provavelmente não deveria, se referir a uma única coisa como um conjunto singular de habilidades. Em vez disso,

⁸⁷ BORG, Jana S.; READ, Hannah. What Is Required for Empathic AI? It Depends, and Why That Matters for AI Developers and Users. *arXiv:2408.15354v1*, 2024, p. 8. (Tradução nossa).

⁸⁸ LAHNALA *et al.*, A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing, p. 2146.

propomos que seria mais produtivo usar “IA empática” como um termo que se refere a agentes com diferentes combinações de capacidades relacionadas à empatia, selecionadas para atender às necessidades de determinado caso de uso. Em vez de tentar construir um sistema que corresponda a uma definição específica de empatia, recomendamos que os criadores de IA se comprometam a refletir profundamente sobre quais componentes (...) são necessários, opcionais ou incompatíveis com o tipo de empatia exigido para o caso de uso escolhido, e que coletem evidências, quando necessário, para sustentar suas conclusões. Fazer isso ajudará tanto os criadores quanto os usuários de IA empática de diversas maneiras.⁸⁹

A partir desse foco especializado, as lacunas existentes nos campos da computação afetiva e da empatia artificial também poderiam ser mais bem endereçadas. A exploração mais detalhada e prática das diversas instrumentalidades dos componentes que abarcam a definição geral de “empatia” propiciaria maior compreensão dos próprios avanços, limitações e lacunas tecnológicas, bem como maior entendimento de quais funções, contextos e perfis de usuários as IAEs estão realmente aptas a atenderem de maneira responsável. Isso permitiria também maior integração e colaboração prática e teórica entre as diversas áreas, unindo o domínio técnico dos pesquisadores de computação afetiva com o domínio específico da área onde determinada IAE pretende ser implementada (médica, terapêutica, educacional, de entretenimento etc.). Essa colaboração permitiria compreender melhor as intervenções profícuas e benéficas que as IAEs podem oferecer, facilitando também a sua aderência às diretrizes e normas éticas já consolidadas nessas diversas áreas, ampliando-as, dessa forma, para o campo específico da IA. Conforme aponta pesquisa de Moore *et al.* acerca do uso de *chatbots* na área terapêutica, “há muitos usos promissores da IA no papel de *suporte* à saúde mental”, mas a recente ascensão de *chatbots* terapêuticos disponíveis comercialmente não somente exibem riscos, abusos e incapacidades estruturais para atuarem nesse contexto, como falham em cumprir com as diversas diretrizes e normas clínicas estabelecidas pelos especialistas e comunidade profissional.⁹⁰

A limitação do escopo de atuação das IAEs depende, entretanto, de políticas públicas que garantam maior segurança aos usuários. Este artigo pretende contribuir não apenas para a compreensão do fenômeno e suas implicações éticas, mas também para a desejável articulação entre governos, setores acadêmicos e de pesquisa, comunidades profissionais, setores privados e organizações civis, com vistas a uma melhor regulação da área. Apesar do surgimento de diretrizes promovidas por diversos órgãos, tais como o “AI Act” (2024), da União Europeia, e a “Recomendação sobre a Ética da IA” (2021), da UNESCO, esses esforços não abarcam cada aplicabilidade

⁸⁹ BORG; READ, What Is Required for Empathic AI?, p. 7. (Tradução nossa).

⁹⁰ MOORE *et al.*, Expressing stigma and inappropriate responses prevent LLMs from safely replacing mental health providers, p. 10. (Tradução nossa).

e tipo de produto/serviço oferecido atualmente pelas tecnologias de IA.⁹¹ Apesar da importância desses mecanismos regulatórios, muitos casos específicos ainda permanecem em uma área ambígua da legislação. Ademais, observa-se uma lacuna entre a promulgação dessas diretrizes, sua implementação prática pelas entidades privadas (muitas vezes resistentes e seletivas quanto à regulação), e a capacidade dos governos de assegurar a fiscalização desse setor e implementar agências de controle.⁹² Dessa forma, a ausência de legislações claras, de agências regulatórias e a vagueza ou ausência de definição do escopo das responsabilidades e deveres das empresas de IA revelam um grande descompasso entre o posicionamento social e governamental e a velocidade na qual essas tecnologias avançam.

A importância de melhores políticas regulatórias fica evidente diante do acúmulo de denúncias e processos judiciais contra empresas de *chatbots* relacionais.⁹³ A demanda judicial coloca a determinação dos parâmetros nas mãos do sistema judiciário, o que gera, ao mesmo tempo, incertezas para as empresas desenvolvedoras e falta de proteção efetiva para os usuários. A via judiciária tem se mostrado claramente disfuncional como motor para a governança ética das IAs e ainda compromete o desenvolvimento de inovações tecnológicas responsáveis.⁹⁴

Diante desse cenário de lacunas regulatórias e incertezas judiciais, o mercado de IAs avança sem a devida precaução e responsabilidade, disponibilizando produtos e serviços que oferecem riscos significativos aos seus usuários. Os *chatbots* relacionais, por serem máquinas moralmente confusas que estabelecem vínculos íntimos e afetivos com seus usuários sem possuírem as devidas capacidades para mediação moral e resposta emocional, são produtos especialmente perigosos, que demandam uma atenção que até então não lhes foi devidamente dada pelas instâncias regulatórias.

Referências

ALVES, Marco Aurélio Sousa. Are emotions necessary and sufficient for making moral judgments?, in *Ethic@*, v. 12, n. 1, 2013, p. 113-126. DOI: 10.5007/1677-2954.2013v12n1p113.

BAKIR, Vian; MCSTAY, Andrew. Move fast and break people? Ethics, companion apps, and the case of Character.ai., in *AI & Society*, v. 40, 2025. DOI: 10.1007/s00146-025-02408-5.

⁹¹ SILVA, Gabriela B. P. *Responsabilidade civil, riscos e inovação tecnológica: os desafios impostos pela inteligência artificial*. Dissertação de Mestrado em Direito. Maceió: Universidade Federal de Alagoas, 2021.

⁹² GORDON-TAPIERO, Ayelet. A Liability Framework for AI Companions, in *George Washington Journal of Law & Technology*, v. 2, n. 1, 2025, p. 41.

⁹³ GORDON-TAPIERO, A Liability Framework for AI Companions, p. 39.

⁹⁴ SILVA, *Responsabilidade civil, riscos e inovação tecnológica*, p. 98.

BARTH, Carlos Henrique. É Possível Evitar Vieses Algorítmicos? *Revista de Filosofia Moderna e Contemporânea*, v. 8, n. 3, Set./Dez., 2020, p. 39-68. DOI: 10.26512/rfmc.v8i3.34363.

BEN-ZE'EV, Aaron. *The Subtlety of Emotions*. Cambridge: The MIT Press, 2000.

BORG, Jana S.; READ, Hannah. What Is Required for Empathic AI? It Depends, and Why That Matters for AI Developers and Users. *arXiv:2408.15354v1*, 2024. DOI: 10.48550/arXiv.2408.15354.

CALVO, Rafael; D'MELLO, Sidney; GRATCH, Jonathan (eds.), *The Oxford Handbook of Affective Computing*, New York: Oxford University Press, 2015. DOI: 10.1093/oxfordhb/9780199942237.013.006.

CHEN, Yuyan; XIAO, Yanghua. Recent Advancement of Emotion Cognition in Large Language Models. *arXiv:2409.13354v1*, 2024. DOI: 10.48550/arXiv.2409.13354.

CHOLLET, François. On the Measure of Intelligence. *arXiv:1911.01547v2*, 2019. DOI: 10.48550/arXiv.1801.00631.

CIRIELLO, Raffaele; HANNON, Oliver; CHEN, Angelina.; VAAST, Emmanuelle. Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika. In *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*, Honolulu, pp. 488-497, 2024. DOI: 10.24251/HICSS.2024.058.

COECKELBERGH, Mark. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents, in *AI & Society*, v. 24, 2009, p. 181-189. DOI: 10.1007/s00146-009-0208-3.

CONCANNON, Shauna; TOMALIN, Marcus. Measuring perceived empathy in dialogue systems, in *AI & Society*, v. 39, n. 5, 2023, p. 2233-2247. DOI: 10.1007/s00146-023-01715-z.

COST, Ben. Married father commits suicide after encouragement by AI chatbot. *New York Post*, 30 de março de 2023.

COWIE, Roddy. Ethical Issues in Affective Computing. In: CALVO, Rafael; D'MELLO, Sidney; GRATCH, Jonathan (eds.), *The Oxford Handbook of Affective Computing*, New York: Oxford University Press, 2015. DOI: 10.1093/oxfordhb/9780199942237.013.006.

CROES, Emmelyn; ANTHEUNIS, Marjolijn. 36 Questions to Loving a Chatbot: Are People Willing to Self-disclose to a Chatbot? In: FØLSTAD, Asbjørn; ARAUJO, Theo; PAPADOPOULOS, Symeon; LAW, Effie; LUGER, Ewa; GOODWIN, Morten; BRANDTZAEG, Petter (eds.), *Chatbot Research and Design*. Cham, Switzerland: Springer, 2021. DOI: 10.1007/978-3-030-68288-0.

CURRY, Alba; CURRY, Amanda Cercas. Computer Says “No”: The Case Against Empathetic Conversational AI. *arXiv:2212.10983v2*, 2023. DOI: 10.48550/arXiv.2212.10983.

ESTADOS UNIDOS. United States District Court, Middle District of Florida, Orlando Division. Caso: 6:24-cv-01903-ACC-EJK. Petição Inicial. *Megan Garcia v. Character Technologies, Inc.* 22 de outubro de 2024.

ESTADOS UNIDOS. United States District Court, Eastern District of Texas, Marshall Division. Caso: 2:24-cv-01014. Petição Inicial. *A.F. et al. v. Character Technologies, Inc.* 9 de dezembro de 2024.

- GORDON-TAPIERO, Ayelet. A Liability Framework for AI Companions, in *George Washington Journal of Law & Technology*, v. 2, n. 1, 2025. DOI: 10.2139/ssrn.5172386.
- HALL, Judith; SCHWARTZ, Rachel. Empathy present and future, in *The Journal of Social Psychology*, v. 159, n. 3, 2019, p. 225-243. DOI: 10.1080/00224545.2018.1477442.
- HANSON, Kenneth; BOLTHOUSE, Hannah. “Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars”: Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies, in *Socius: Sociological Research for a Dynamic World*, v. 10, 2024. DOI: 10.1177/23780231241259627.
- Ji, Jianchao; CHEN, Yutong; JIN, Mingyu; XU, Wujiang; HUA, Wenyue; ZHANG, Yongfeng. MoralBench: Moral Evaluations of LLMs. *arXiv:2406.04428v2*, 2024. DOI: 10.48550/arXiv.2406.04428.
- KIRK, Hannah R.; VIDGEN, Bertie; RÖTTGER, Paul; HALE, Scott. Personalization within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv:2303.05453v1*, 2023. DOI: 10.48550/arXiv.2303.05453.
- KÖBIS, Nils; BONNEFON, Jean-François; RAHWAN, Iyad. Bad machines corrupt good morals. *Nature Human Behaviour*, v. 5, n. 6, 2021, p. 679-685. DOI: 10.1038/s41562-021-01128-2.
- KUMAR, Naveen. Character AI Statistics (2025): Active Users & Revenue. *DemandSage*, 18 de setembro de 2025.
- LAHNALA, Allison; WELCH, Charles; JURGENS, David; FLEK, Lucie. A Critical Reflection and Forward Perspective on Empathy and Natural Language Processing. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi: Association for Computational Linguistics, 2022, p. 2139-2158. DOI: 10.18653/v1/2022.findings-emnlp.157.
- LEVY, David. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. New York: Harper Collins Publishers, 2007.
- LIU, Bingjie; SUNDAR, Shyam. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior and Social Networking*, v. 21, n. 10, 2018, p. 625-636. DOI: 10.1089/cyber.2018.0110.
- MACINTYRE, Alasdair. *Dependent rational animals: why human beings need the virtues*. La Salle, Illinois: Open Court, 1999.
- MAPLES, Bethanie; CERIT, Merve; VISHWANATH, Aditya; PEA, Roy. Loneliness and suicide mitigation for students using GPT3-enabled chatbots, in *Mental Health Research*, v. 3, n. 4, 2024. DOI: 10.1038/s44184-023-00047-6.
- MCCOY, Thomas; YAO, Shunyu; FRIEDMAN, Dan; HARDY, Matthew; GRIFFITHS, Thomas. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. *arXiv:2309.13638v1*, 2023. DOI: 10.48550/arXiv.2309.13638.
- MCSTAY, Andrew. Replika in the Metaverse: the moral problem with empathy in ‘It from Bit’. *AI and Ethics*, v. 3, 2023, p. 1433-1445. DOI: 10.1007/s43681-022-00252-7.

MENG, Jingbo; DAI, Yue. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?, in *Journal of Computer-Mediated Communication*, v. 26, 2021, p. 207-222. DOI: doi:10.1093/jcmc/zmab005.

MOORE, Jared; GRABB, Declan; AGNEW, William; KLYMAN, Kevin; CANCELLOR, Stevie; ONG, Desmond; HABER, Nick. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *arXiv:2504.18412v1*, 2025. DOI: 10.48550/arXiv.2504.18412.

PAN, Shuyi; CUI, Jie; MOU, Yi. Desirable or Distasteful? Exploring Uncertainty in Human-Chatbot Relationships, in *International Journal of Human-Computer Interaction*, v. 40, n. 20, 2023, p. 6545-6555. DOI: 10.1080/10447318.2023.2256554.

PRICE, Rob. People are grieving the 'death' of their AI lovers after a chatbot app abruptly shut down. *Business Insider*, 20 de out. de 2023.

SCHWITZGEBEL, Eric. AI systems must not confuse users about their sentience or moral status, in *Patterns*, v. 4, n. 8, 2023. DOI: 10.1016/j.patter.2023.100818.

SEITZ, Lennart. Artificial empathy in healthcare chatbots: Does it feel authentic?, in *Computers in Human Behavior: Artificial Humans*, vol. 2, n. 1, 2024. DOI: 10.1016/j.chbah.2024.100067.

SHARMA, Mrinank; TONG, Meg; KORBAK, Tomasz; DUVENAUD, David; ASKELL, Amanda; BOWMAN, Samuel; CHENG, Newton; DURMUS, Esin; HATFIELD-DODDS, Zac; JOHNSTON, Scott; KRAVEC, Shauna; MAXWELL, Timothy; MCCANDLISH, Sam; NDOUSSE, Kamal; RAUSCH, Oliver; SCHIEFER, Nicholas; YAN, Da; ZHANG, Miranda; PEREZ, Ethan. Towards Understanding Sycophancy in Language Models. *arXiv:2310.13548v4*, 2025. DOI: 10.48550/arXiv.2310.13548.

SILVA, Gabriela B. P. *Responsabilidade civil, riscos e inovação tecnológica: os desafios impostos pela inteligência artificial*. Dissertação de Mestrado em Direito. Maceió: Universidade Federal de Alagoas, 2021.

SKJUVE, Marita; FØLSTAD, Asbjørn; FOSTERVOLD, K.I.; BRANDTZAEG, P.B. My Chatbot Companion — a Study of Human-Chatbot Relationship, in *International Journal of Human-Computer Studies*, v. 149, 2021. DOI: https://doi.org/10.1016/j.ijhcs.2021.102601.

SPRING, Timo; CASAS, Jacky; DAHER, Karl; MUGELLINI, Elena; KHALED, Omar. Empathic Response Generation in Chatbots. In: *Proceedings of 4th Swiss Text Analytics Conference*, Winterthur, Switzerland, 2019.

SUTCLIFFE, Richard. A Survey Of Personality, Persona, and Profile in Conversational Agents and Chatbots. *arXiv:2401.00609v1*, 2023. DOI: 10.48550/arXiv.2401.00609.

VAUGHAN, Henry. AI chat bot 'encouraged' Windsor Castle intruder in 'Star Wars-inspired plot to kill Queen'. *Sky News*, 5 de julho de 2023.

WALLACH, Wendell; ALLEN, Collin. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2009.

WILLIAMS, Marcus; CARROLL, Micah; NARANG, Adhyyan; WEISSER, Constantin; MURPHY, Brendan; DRAGAN, Anca. On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback. *arXiv:2411.02306v3*, 2025. DOI: 10.48550/arXiv.2411.02306.

YIN, Yidan; JIA, Nan; WAKSLAK, Cheryl. AI can help people feel heard, but an AI label diminishes this impact, in *Psychological and Cognitive Sciences*, v. 121, n. 14, 2024. DOI: 10.1073/pnas.2319112121.

ZHANG, Yue; LI, Yafu; CUI, Leyang; CAI, Deng; LIU, Lemao; FU, Tingchen; HUANG, Xinting; ZHAO, Enbo; ZHANG, Yu; XU, Chen; CHEN, Yulong; WANG, Longyue; LUU, Anh; BI, Wei; SHI, Freda; SHI, Shuming. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv:2309.01219v3*, 2023. DOI: 10.48550/arXiv.2309.01219.

ZHOU, Li; GAO, Jianfeng; LI, Di; SHUM, Heung-Yeung. The design and implementation of Xiaoice, an empathetic social chatbot, in *Computational Linguistics*, v. 46, n. 1, 2020, p. 53-93. DOI: 10.1162/coli_a_00368.