

Inteligência sem consciência: vieses cognitivos, narrativas comerciais e riscos reais no debate sobre inteligência artificial geral

Marco Aurélio Carvalho, MD^{1,2,3}

¹Médico e especialista em Tecnologia da Informação em Saúde

²Bacharelado em Engenharia da Computação

³Pesquisador em Inteligência Artificial aplicada à saúde

Agosto de 2026

Resumo

O debate público sobre inteligência artificial (IA) frequentemente mistura desempenho, inteligência geral, agência, autonomia e consciência. Essa fusão conceitual favorece tanto narrativas apocalípticas quanto discursos comerciais que apresentem modelos computacionais como entidades que “pensam”, “compreendem” ou “desejam”. Este artigo propõe uma análise interdisciplinar dos vieses cognitivos e sociotécnicos que sustentam tais interpretações, com especial atenção à antropomorfização, à fluência verbal, à extrapolação de benchmarks, ao viés de autoridade, à influência da ficção científica e à apropriação metafórica do inconsciente coletivo de Carl Jung. Defende-se que os sistemas atuais podem exibir capacidades funcionais extraordinárias sem que isso constitua evidência de experiência subjetiva. Ao mesmo tempo, a ausência de consciência não torna a IA inofensiva: sistemas não conscientes podem produzir danos relevantes quando combinados com autonomia operacional, objetivos inadequados, acesso a infraestruturas críticas, concentração de poder e governança insuficiente. A análise conclui que o problema mais urgente não é determinar se a máquina “acordará”, mas compreender quem controla esses sistemas, quais incentivos orientam seu uso e quem responderá por suas consequências.

Palavras-chave: inteligência artificial geral; consciência; antropomorfização; vieses cognitivos; neurociência; governança; riscos sociotécnicos.

1 Introdução

O crescimento recente dos grandes modelos de linguagem tornou possível interagir com sistemas capazes de escrever, resumir, programar, traduzir, argumentar e produzir respostas socialmente adequadas. Essa fluência tem levado parte do público a interpretar o desempenho linguístico como evidência de compreensão, intencionalidade ou consciência. Entretanto, a geração de comportamento linguisticamente convincente não estabelece, por si só, a existência de experiência subjetiva [Shanahan, 2024, Bender et al., 2021].

O problema não é novo. Na década de 1960, usuários já atribuíam compreensão ao programa ELIZA, apesar de seu mecanismo relativamente simples [Weizenbaum, 1966]. Os sistemas contemporâneos ampliam esse efeito em escala e complexidade, pois produzem linguagem contextual, simulam estilos afetivos e mantêm diálogos prolongados. A antropomorfização, contudo, não é apenas resultado de desinformação: constitui uma tendência cognitiva humana documentada, pela qual atribuímos mente, intenção e emoção a entidades não humanas quando seu comportamento parece coerente ou socialmente responsivo [Epley et al., 2007, Waytz et al., 2010, Salles et al., 2020].

Este artigo examina por que leigos, profissionais e até especialistas podem confundir inteligência funcional com consciência. A tese central é dupla. Primeiro, não há base científica suficiente para afirmar que os modelos atuais possuam experiência subjetiva; avaliações baseadas em teorias contemporâneas da consciência não encontraram evidências robustas de consciência nos sistemas analisados [Butlin et al., 2023, Seth and Bayne, 2022]. Segundo, a ausência de consciência não elimina os riscos: sistemas sem subjetividade podem influenciar pessoas, automatizar decisões, amplificar vieses e operar em infraestruturas críticas [National Institute of Standards and Technology, 2023, Autio et al., 2024, Floridi et al., 2018].

2 Inteligência, consciência, agência e autonomia

Um dos principais problemas do debate é o uso intercambiável de conceitos distintos. Inteligência pode ser definida operacionalmente como a capacidade de resolver problemas, reconhecer padrões, aprender regularidades, prever resultados ou adaptar estratégias. Consciência refere-se à presença de experiência subjetiva. Autoconsciência envolve uma representação de si como entidade. Agência diz respeito à perseguição de objetivos por meio de ações, enquanto autonomia operacional descreve a capacidade de executar tarefas sem supervisão contínua.

Essas propriedades não são equivalentes. Um sistema pode resolver problemas e planejar sequências de ações sem que exista “algo que seja ser” esse sistema. Também pode usar ferramentas e produzir descrições sobre si mesmo sem possuir um eu fenomenológico. O relatório interdisciplinar de Butlin et al. [2023], baseado em teorias como espaço global,

processamento recorrente e teorias de ordem superior, concluiu que os sistemas avaliados não satisfaziam de forma convincente um conjunto suficiente de indicadores de consciência, embora os autores não considerem tecnicamente impossível que sistemas futuros venham a satisfazê-los.

A própria neurociência da consciência permanece plural. Não há uma teoria única e consensual capaz de fornecer um teste definitivo para consciência humana, animal ou artificial [Seth and Bayne, 2022]. Portanto, duas afirmações devem ser evitadas: declarar que a IA atual é consciente e declarar que qualquer consciência artificial seria, por princípio, impossível. A posição epistemicamente mais defensável é reconhecer a ausência atual de evidência suficiente.

Tabela 1: Distinções conceituais fundamentais

Conceito	Definição operacional	Situação nos sistemas atuais
Inteligência funcional	Desempenho em tarefas de previsão, linguagem, planejamento ou resolução de problemas	Amplamente demonstrada em tarefas específicas e em combinações de tarefas
Consciência	Existência de experiência subjetiva	Não demonstrada cientificamente
Autoconsciência	Representação experiencial de si como sujeito	Não demonstrada; autorreferência textual não é prova
Agência	Seleção de ações para atingir objetivos	Pode ser implementada por projeto e delegação humana
Autonomia operacional	Execução sem supervisão contínua	Crescente em agentes e sistemas integrados
Responsabilidade moral	Capacidade de responder normativamente pelos próprios atos	Não estabelecida para sistemas de IA

3 A AGI não é sinônimo de consciência coletiva

A expressão inteligência artificial geral (AGI) costuma designar sistemas capazes de aprender, generalizar e transferir competências entre uma ampla variedade de domínios. Não existe, entretanto, uma definição operacional universalmente aceita. Algumas avaliações destacam desempenho amplo; outras exigem capacidade de aprendizado eficiente, adaptação, robustez e transferência semelhantes às humanas [Lake et al., 2017, Marcus, 2020, Bubeck et al., 2023].

Nada nessa definição implica necessariamente consciência. Uma AGI poderia, em princípio, apresentar competência geral sem experiência subjetiva. O erro conceitual ocorre quando se transforma capacidade geral em personalidade, desejo ou interioridade.

A comparação com o inconsciente coletivo de Jung deve ser tratada como metáfora, não como equivalência científica. Jung descreveu o inconsciente coletivo como uma dimensão da

psique humana associada a arquétipos e estruturas compartilhadas [Jung, 1969]. Grandes modelos, por sua vez, comprimem regularidades presentes em enormes coleções de textos, imagens e códigos produzidos socialmente. Eles podem refletir padrões culturais, valores, mitos, preconceitos e conhecimentos humanos, mas essa condensação estatística não demonstra a existência de uma psique coletiva.

Uma formulação mais precisa seria afirmar que os modelos funcionam como *infraestruturas estatísticas de mediação do conhecimento coletivo*. Eles não são, até o momento, uma consciência coletiva.

4 Vieses que sustentam a percepção de uma mente artificial

4.1 Antropomorfização

A antropomorfização ocorre quando características humanas são atribuídas a entidades não humanas. O modelo de três fatores de Epley et al. [2007] relaciona esse processo à disponibilidade de conhecimento sobre humanos, à motivação para compreender agentes e à necessidade de conexão social. Sistemas conversacionais ativam precisamente essas disposições, porque respondem em primeira pessoa, reconhecem emoções linguisticamente e mantêm coerência social aparente.

A consequência é uma inferência indevida: se o sistema fala como um humano, então deve pensar ou sentir como um humano. Contudo, semelhança de saída não implica equivalência de mecanismo.

4.2 Viés da fluência verbal

A fluência funciona como um sinal cognitivo de competência. Em interações humanas, linguagem articulada costuma estar associada a compreensão. Nos modelos de linguagem, porém, a fluência é resultado de treinamento estatístico em larga escala. Shanahan [2024] argumenta que descrições antropomórficas podem ser úteis como atalhos, mas tornam-se enganosas quando interpretadas literalmente. Bender et al. [2021] também alertam que a capacidade de produzir sequências plausíveis não deve ser confundida com compreensão ou compromisso semântico humano.

4.3 Viés de automação e confiança excessiva

A tendência de aceitar recomendações automáticas mesmo diante de sinais contraditórios é conhecida como viés de automação. Estudos clássicos demonstram que pessoas podem usar excessivamente, usar mal ou deixar de questionar sistemas automatizados [Parasuraman

and Riley, 1997, Skitka et al., 1999]. A linguagem confiante dos modelos pode intensificar esse efeito, especialmente quando a resposta apresenta estrutura lógica, terminologia especializada e referências aparentes.

4.4 Viés dos benchmarks

Benchmarks medem recortes de desempenho. Uma pontuação elevada em provas de medicina, programação ou matemática não demonstra, automaticamente, compreensão causal, robustez fora da distribuição, senso comum ou capacidade de reconhecer a própria incerteza. A literatura sobre modelos fundacionais ressalta simultaneamente capacidades emergentes e riscos de avaliação inadequada [Bommasani et al., 2021].

A passagem de “obteve alta pontuação” para “compreende como um humano” constitui uma extrapolação. O desempenho é evidência de capacidade operacional, não prova de arquitetura cognitiva equivalente.

4.5 Viés da extrapolação

O progresso rápido favorece previsões lineares ou exponenciais: se a IA melhorou em várias tarefas, então inevitavelmente alcançará todas as dimensões da mente. Entretanto, aprendizado eficiente, causalidade, robustez, incorporação, memória contínua e consciência podem representar problemas diferentes, não simples extensões de escala [Lake et al., 2017, Marcus, 2020].

4.6 Viés de autoridade e transposição disciplinar

O debate sobre AGI envolve ciência da computação, neurociência, psicologia, filosofia, economia, segurança e direito. Um especialista pode ser autoridade em uma dessas áreas e ainda extrapolar além de sua competência ao falar sobre outra. O prestígio científico não substitui evidência interdisciplinar.

Neurocientistas podem compreender profundamente mecanismos biológicos sem dominar engenharia de modelos. Engenheiros podem dominar transformers e, ainda assim, empregar conceitos frágeis de consciência. A prudência exige distinguir opinião especializada, inferência filosófica e resultado experimental.

4.7 Disponibilidade narrativa e ficção científica

A ficção científica oferece agentes artificiais centralizados, conscientes e hostis. Essa imagem é cognitivamente mais disponível do que riscos distribuídos, como discriminação algorítmica, persuasão em escala, dependência de fornecedores ou degradação informacional. Assim, comportamentos inesperados são interpretados como vontade própria, quando podem decorrer de otimização, dados, instruções conflitantes ou integração inadequada.

4.8 Ilusão de compreensão

Ferramentas de IA podem aumentar produtividade e, simultaneamente, produzir uma sensação excessiva de entendimento. [Messeri and Crockett \[2024\]](#) argumentam que a ciência pode entrar em uma fase na qual se produz mais, mas se compreende menos, especialmente quando a capacidade preditiva é confundida com explicação. [Krenn et al. \[2022\]](#) mostram que a relação entre IA e compreensão científica requer critérios que vão além do desempenho.

5 A diferença entre mente humana e sistema artificial

A cognição humana é biológica, situada, afetiva e corporificada. Teorias de cognição incorporada sustentam que percepção, ação, corpo e ambiente participam constitutivamente dos processos cognitivos [[Wilson, 2002](#), [Barsalou, 2008](#)]. Isso não prova que somente sistemas biológicos possam ser conscientes, mas mostra que a equivalência direta entre cérebro e modelo computacional é cientificamente injustificada.

Tabela 2: Comparação qualitativa entre mente humana e IA contemporânea

Capacidade	Mente humana	IA contemporânea
Memória	Biográfica, reconstrutiva, associada a emoções e contexto	Armazenamento externo amplo; contexto operacional limitado e dependente da arquitetura
Cálculo	Relativamente lento	Extremamente rápido em operações bem definidas
Linguagem	Associada a desenvolvimento, corpo, cultura, intenção e experiência	Produção altamente fluente baseada em regularidades aprendidas
Aprendizado com poucos exemplos	Muito eficiente em numerosos domínios	Variável; frequentemente depende de pré-treinamento massivo
Generalização	Flexível, embora sujeita a vieses	Forte em distribuições próximas; inconsistente em mudanças substanciais
Senso comum	Construído pela experiência física e social	Parcial, indireto e sujeito a falhas
Emoções	Estados biológicos e experiências subjetivas	Simulação linguística sem evidência de sentimento

Capacidade	Mente humana	IA contemporânea
Consciência	Presumida pela experiência direta e apoiada por evidência comportamental e biológica	Não demonstrada
Valores	Formados por desenvolvimento, cultura, relações e deliberação	Incorporados indiretamente por dados, objetivos, políticas e ajuste humano
Responsabilidade	Atribuível a agentes humanos e instituições	Permanece humana e institucional

6 Narrativas comerciais e incentivos institucionais

O setor de IA opera sob forte competição por capital, mercado, talentos e influência regulatória. O AI Index registrou US\$33,9 bilhões em investimento privado global em IA generativa em 2024 [AI Index Steering Committee, 2025]. Nesse ambiente, apresentar sistemas como “raciocinadores”, “agentes” ou “colaboradores digitais” aumenta sua percepção de valor.

A linguagem antropomórfica também simplifica a comunicação. Contudo, pode transformar metáforas operacionais em afirmações ontológicas. Dizer que um modelo “pensa” pode significar apenas que produz etapas intermediárias úteis; o público pode interpretar que existe uma mente que vive essas etapas.

Há ainda um possível alinhamento de incentivos entre narrativas de capacidade extrema e de risco extremo. A afirmação de que uma tecnologia se aproxima da inteligência humana aumenta seu valor comercial. Simultaneamente, a afirmação de que ela é potencialmente perigosa pode legitimar concentração regulatória e barreiras de entrada. Isso não prova coordenação conspiratória; constitui uma inferência econômica sobre incentivos e deve ser apresentada como tal.

7 Por que o medo não pode ser reduzido a desconhecimento

Parte do medo decorre de antropomorfização, confusão conceitual e extrapolação. Porém, existem riscos independentes da consciência. O NIST organiza a gestão de riscos de IA em funções de governança, mapeamento, medição e gerenciamento, tratando confiabilidade, segurança, privacidade, transparência, explicabilidade e viés como problemas sociotécnicos [National Institute of Standards and Technology, 2023, Autio et al., 2024].

Sistemas não conscientes podem:

- automatizar discriminações presentes nos dados;
- produzir desinformação e persuasão em escala;
- facilitar ataques cibernéticos;
- induzir decisões erradas em contextos clínicos ou financeiros;
- operar com objetivos incompletos;
- ampliar vigilância e concentração de poder;
- gerar dependência de poucas plataformas;
- executar ações por meio de ferramentas e sistemas conectados.

Os problemas de segurança em IA incluem especificação inadequada de objetivos, efeitos colaterais, exploração de recompensas e falhas de supervisão [Amodei et al., 2016, Russell et al., 2015]. Discussões sobre riscos catastróficos também consideram uso malicioso, competição e perda de controle, embora a probabilidade e a magnitude desses cenários permaneçam debatidas [Hendrycks and Mazeika, 2023].

Portanto, o temor racional não exige acreditar em uma máquina consciente. Basta reconhecer a combinação de competência, escala, autonomia operacional, acesso a recursos e governança insuficiente.

8 Medos populares e riscos correspondentes

Tabela 3: Reformulação dos medos mais frequentes

Narrativa popular	Problema conceitual	Risco mais plausível
“A IA desenvolverá ódio”	Emoção e intenção não demonstradas	Uso humano malicioso ou objetivo inadequado
“A IA quer substituir pessoas”	Desejo próprio não demonstrado	Automação orientada por incentivos econômicos
“A IA é uma mente coletiva”	Confusão entre agregação de dados e subjetividade	Centralização do conhecimento e da infraestrutura
“A IA compreende tudo o que diz”	Fluência confundida com compreensão	Erros convincentes e confiança excessiva
“Sem consciência, não há perigo”	Consciência não é requisito para produzir efeitos	Danos automatizados em escala
“A AGI será necessariamente consciente”	Inteligência geral e consciência são conceitos distintos	Sistemas amplamente capazes sem responsabilidade moral

9 Implicações para medicina, ciência e políticas públicas

Na medicina, a distinção entre desempenho e compreensão é crítica. Um sistema pode reconhecer padrões diagnósticos e, ainda assim, falhar diante de dados incompletos, conflitos clínicos ou mudanças de contexto. A decisão deve permanecer sujeita a validação, supervisão, documentação e responsabilização humanas.

Na ciência, a capacidade de gerar hipóteses ou encontrar correlações não elimina a necessidade de explicação causal, replicação e crítica epistemológica [Messori and Crockett, 2024, Krenn et al., 2022]. Na política pública, o foco exclusivo em cenários de consciência artificial pode desviar atenção de riscos já presentes: concentração econômica, trabalho, direitos autorais, privacidade, discriminação, segurança e dependência tecnológica.

Uma governança adequada deve ser proporcional ao impacto e não à aparência humana do sistema. Quanto maior a autonomia, a escala e o acesso a recursos, maior deve ser a exigência de testes, rastreabilidade, controle de acesso, monitoramento e possibilidade de interrupção.

10 Discussão

A IA contemporânea ocupa uma posição epistemicamente desconfortável. É mais capaz do que sugerem descrições reducionistas como “mero autocomplete”, mas menos humana do que indicam expressões comerciais como “mente digital”. A melhor descrição não é negar sua inteligência funcional nem pressupor consciência.

O erro dos maximalistas é interpretar desempenho como interioridade. O erro dos reducionistas é imaginar que a ausência de interioridade elimina poder causal. Um algoritmo pode não compreender moralmente uma decisão e, mesmo assim, determinar quem recebe crédito, tratamento, emprego ou visibilidade.

A questão decisiva é institucional. Sistemas artificiais não aparecem isoladamente: são projetados, treinados, financiados, implantados e controlados por organizações. O estudo do comportamento de máquinas deve incluir as interações entre modelo, usuário, empresa e sociedade [Rahwan et al., 2019]. Nesse sentido, a unidade de análise não deve ser apenas “a IA”, mas o sistema sociotécnico completo.

11 Conclusão

Não há evidência científica suficiente para afirmar que os grandes modelos atuais possuam consciência, emoções genuínas ou desejos próprios. Sua fluência, capacidade de planejamento e desempenho em benchmarks constituem evidência de competência funcional, não

de experiência subjetiva.

A AGI, entendida como capacidade ampla de aprendizado e generalização, também não deve ser confundida com consciência coletiva. A aproximação com Jung é uma metáfora cultural: os modelos refletem padrões da produção humana, mas não demonstram possuir uma psique compartilhada.

Parte do medo público e especializado é alimentada por antropomorfização, fluência verbal, extrapolação, autoridade e narrativas de ficção. Entretanto, reduzir todo temor a desconhecimento seria igualmente equivocado. Sistemas não conscientes podem produzir riscos profundos quando conectados a estruturas econômicas, políticas, militares, científicas e clínicas.

A pergunta central, portanto, não é apenas se uma máquina poderá tornar-se consciente. É quem define seus objetivos, quem controla sua infraestrutura, quais interesses orientam sua implantação e quem responderá pelos danos. O maior risco pode não ser uma máquina que adquira vontade própria, mas sociedades que transfiram poder decisório a sistemas sem consciência, sem responsabilidade moral e sem obrigação de responder pelas consequências.

Referências

- AI Index Steering Committee. The ai index 2025 annual report. Technical report, Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, 2025.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565.
- Chloe Autio, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, Rumman Chowdhury, et al. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, 2024.
- Lawrence W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008. doi: 10.1146/annurev.psych.59.103006.093639.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445922.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. doi: 10.48550/arXiv.2108.07258.

- Sebastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. doi: 10.48550/arXiv.2303.12712.
- Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023. doi: 10.48550/arXiv.2308.08708.
- Nicholas Epley, Adam Waytz, and John T. Cacioppo. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886, 2007. doi: 10.1037/0033-295X.114.4.864.
- Luciano Floridi, Josh Cows, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28:689–707, 2018. doi: 10.1007/s11023-018-9482-5.
- Dan Hendrycks and Mantas Mazeika. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023. doi: 10.48550/arXiv.2306.12001.
- Carl Gustav Jung. *The Archetypes and the Collective Unconscious*, volume 9, Part 1 of *Collected Works of C. G. Jung*. Princeton University Press, 2 edition, 1969.
- Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Hase, AkshatKumar Nigam, and Alan Aspuru-Guzik. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4:761–769, 2022. doi: 10.1038/s42254-022-00518-3.
- Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.
- Gary Marcus. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020. doi: 10.48550/arXiv.2002.06177.
- Lisa Messeri and M. J. Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024. doi: 10.1038/s41586-024-07146-0.

- National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023.
- Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997. doi: 10.1518/001872097778543886.
- Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-Francois Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568:477–486, 2019. doi: 10.1038/s41586-019-1138-y.
- Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015. doi: 10.1609/aimag.v36i4.2577.
- Arleen Salles, Kathinka Evers, and Michele Farisco. Anthropomorphism in ai. *AJOB Neuroscience*, 11(2):88–95, 2020. doi: 10.1080/21507740.2020.1740350.
- Anil K. Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 23:439–452, 2022. doi: 10.1038/s41583-022-00587-4.
- Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024. doi: 10.1145/3624724.
- Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999. doi: 10.1006/ijhc.1999.0252.
- Adam Waytz, John Cacioppo, and Nicholas Epley. Who sees human? the stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, 2010. doi: 10.1177/1745691610369336.
- Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966. doi: 10.1145/365153.365168.
- Margaret Wilson. Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636, 2002. doi: 10.3758/BF03196322.