Understanding how we over-trust AI

sheds light on the human conditions

Manh-Tung Ho [1,2*] Hong-Kong T. Nguyen[2]

1. Institute of Philosophy, Vietnam Academy of Social Sciences, Hanoi, 100000, Vietnam
2. Centre for Interdisciplinary Social Research, Phenikaa University, Hanoi, 100803, Vietnam

<<Draft Manuscript No. VNAIethics-20250312#1>>

Figure 1: A kingfisher navigating through a sea of information and advanced tools. Imaged created by DALLE via ChatGPT.

Soon afterward, humans install new bodyguards that are great products of the digital technology era: robots. These humanoid figures, made of endurable materials, are very sophisticated and stand in the fields in all their haughty superiority. More amazing is the fact that these robots are armed with various tools to combat incoming birds…Experienced and quite updated on the new technologies, Kingfisher says: – You guys have met rivals of the 4.0 age. You would for sure fail if you didn't study them carefully. You must do thorough research on theirbehaviors to find a solution.

- In Boogeyman, *Wild Wise Weird: The Kingfisher Story Collection*, Vuong (2025)-

Abstract

In this essay, we argue understanding how we over-trust AI sheds light on what it means to be human. The troubling fact is that we seem to knowingly accept the use of AI products with questionable accuracy and privacy safeguards even in the most high-stake or most intimate situations such as AI uses in war zones or as virtual companionship. We offer five potential explanations for this puzzling fact based on emerging literature on human-AI interactions and evolutionary theory centered around our relationship with information and tools.

Key words: human-AI interaction; evolution; information-foraging behaviors

People knowingly accept flaws and downsides of AI systems

Emerging empirical research converges around an almost paradoxical fact, AI, despite its many flaws and uncertain implications, is over-trusted (Krügel et al., 2022). While text-based generative AI systems such as GPT-4, Gemini, and Copilot suffer from a multitude of serious problems such as hallucinations, jailbreaking, and implicit bias against the disadvantaged, a study by McKinsey in early 2024 shows 65% of respondents report their organizations are regularly using generative AI, which nearly doubles from the previous survey just ten months ago. Such flaws are not limited to generative AI but are also present in virtually all other AI systems. Specifically, all AI programs are subject to algorithmic bias, "data shift"—a mismatch between the training data and the data in the real world, and "underspecification"—wherein machine learning models might have passed benchmarking tests but their performance in practice still has huge variability.

Although the public are aware of these problems, AI systems are being deployed in not only high-stake situations but also the most intimate (e.g., the rise of AI girlfriend/boyfriend apps). Those include not only the installment of face-scanning algorithms in homes, workplaces, and cars with questionable accuracy and privacy safeguards, but also the deployment of AI-powered weapons in the Russia-Ukraine war (e.g., Ukraine's use of autonomous drone Vyriy in combat to hit Russian targets), or the Hamas's use of Chinese-made Da Jiang Innovation drones as suicide drones versus Israel's lethal autonomous and semi-autonomous weapon systems and biometric surveillance system in Gaza.

Five potential explanations

Understanding the psychology and philosophy behind the seemingly paradoxical willingly use and accept flawed AI systems *will illustrate what it means to be human.* This essay discusses five reasons that could account for our acceptance of AI at both the individual and collective levels.

Lack of alternatives

First, on the face of it, today's widespread adoption of AI regardless of its defects is inevitable given the limited options available to users, a problem rooted in the monopolistic dominance of a few tech powers in the global market. Industry giants, e.g., Google, Microsoft,

Meta, and Apple, are leading the global market with their efficiency in raising capitals and acquiring startups or smaller competitors, thereby, consolidating intellectual property and creating barriers to entry for smaller innovators. For example, in January 2014, Google bought London-based artificial intelligence company DeepMind, which gave the company a significant edge in AI development, particularly in machine learning and reinforcement learning. Similarly, in 2019, Meta acquired New York-based CTRL Labs, a startup that develops a wristband designed to transmit electrical signals from the brain into computer inputs. Other cases include Amazon's acquisition of Kiva Systems in 2012 to gain a competitive edge in robotics and warehouse automation, Apple's acquisitions of machine learning company Turi in 2016 and edge computing startup Xnor.ai in 2020, etc. Such moves grant big tech companies exclusive control over cutting-edge innovations and access to top-tier talent, enabling them to construct tightly integrated ecosystems. This consolidation of technological power marginalizes smaller innovators, who often lack the necessary resources and data scale to compete on an equal footing. Worse still, this cycle of dominance and dependency in the digital economy is further reinforced lately with the growing collaboration among themselves. For instance, Meta has been using Microsoft's Azure cloud supercomputing platform since 2021, and in 2023, Microsoft has agreed to integrate Meta's AI model, LLama 2, into Azure. Meanwhile, since 2019, Microsoft has invested more than USD13 billion in OpenAI—which insists that it remains an "independent company governed by the OpenAI Nonprofit." Then, just in June 2024, OpenAI and Apple announced a partnership that would integrate ChatGPT into Apple experiences. Thus, in the face of the tech giants' expansive presence and integrated AI systems, we either have to opt out entirely or go with the flow.

At the same time, our willingness to forgo and even accept the limitations of the tech companies and the AI products themselves reflects a much deeper psychology and philosophy relating to *how we process information* and *cocreate* our perceptions with tools.

Homo Faber

Second, the human species are unique in their innate desire and ability to generate explanatory knowledge and solve problems in the world, thus AI systems, flawed as they are, represent an opportunity and a continuation of what our species we always do: augmenting our cognitive ability with tools (Ho & Vuong, 2024). The *Homo Faber*—i.e., 'man the tool maker'—view posits that humans are not merely the products of nature and culture; rather, what makes us human is the fact that we make things, and in turn, things make us. This view strongly emphasizes the fact that technologies, from the most primitive forms, play defining roles during human evolution. For instance, projectile technology (such as stones, spears, and bows and arrows) allows multiple weaker individuals to defeat a stronger one from a distance with acceptable risk. This has exerted immense evolutionary pressure on humans to improve their communication abilities, form alliances, and resolve conflicts according to complexity and social evolution theorist Peter Turchin. Similarly, AI tools, which can perform many tasks independently and are often equipped with the

ability to gather and analyze our data, tap into our innate desire to augment our problem-solving abilities.

Information foraging and hoarding

Third, once we see how human beings relate to the infosphere through the lens of foraging behaviors, a fundamental behavior for animals, the seductiveness of informational technologies like AI becomes clear: it presents an opportunity to solve the tensions in the exploration versus exploitation tradeoff, i.e., whether to continue foraging for more information about unfamiliar options/strategies or to settle on a strategy with known reward. AI technologies help us deal with an expensive task: information foraging and hoarding, a strategy that makes perfect sense since our brain, for most of its evolutionary history, was in an information and energy scarce environment (Floridi, 2015). Here, even when an AI system fails to perform, it is always rewarding, at least in the short term, similar to the way ultra-processed food feels good to our calories-craving brain.

Perceived self-efficacy

Fourth, another psychological feature underlying the embrace of half-baked AI products is the fact that many people tend to have overstated self-efficacy beliefs in their ability tell the good, the bad, and the ugly of the technologies apart, so to speak. The self-efficacy belief has been shown to explain the privacy-personalization paradox, i.e., people's online behaviors belie their stated strong preferences for privacy. People are very willing to trade their personal data for the *vague* benefits of perceived personalization given that they have high perceived self-control and self-rated knowledge about the technologies (Mantello et al., 2023). The literature on human-computer interactions increasingly reports that users feel they have *tamed or successfully trained* their AI assistants or the recommender algorithms to provide them what they want in reliable manners. Here, in line with the previous point on our desire for augmentation, a rudimentary form of *human-AI symbiosis* emerges: AI algorithms become the users' extended mind, participating into generating their epistemology, self-perception, and perceptions of the world. Effectively, AI participates in *co-creating our* world views and social selves. Some thinkers even go so far as AI systems, albeit nonconscious, have rendered users' preferences predictable. However, this is really two sides of the same coin since humans also strive to make AI systems' behaviors predictable.

Fear of missing out

Fifth, acceptance of flawed AI is born out of the fear of missing out, a psychological trait rooted in our existence as networked beings under great uncertainty of what the networks will do, both at the group and individual levels. From the game theory perspective, which the physicist John von Neumann thought of as a mathematics of incomplete information, humans are always making a decision whether to cooperate with or defect against what they perceive as the social reality, i.e., whether other people are doing the same things and reaping some reward. Thus, we constantly feel the pressure to stay up to date, to avoid being perceived as uninformed or illiterate,

to compete, to save face, and embracing the latest AI products might be a strategy to hedge against these perceived failures in social spaces. We accept this bet even if it comes at the cost of overloading ourselves with a mixed quality of AI-generated and AI-recommended contents, and with losing the ability to make sustained deliberate mental efforts.

Conclusion

In conclusion, acceptance of flawed AI relaxes the worry voiced by many technologists that people will reject new technologies because their failures would produce visceral and salient public reactions. Such an observation of human psychology is of great relevance for policymaking since AI systems increasingly reach performance not of 100% success or accuracy rate but nonetheless better than their human counterparts. Policy-wise, if combined with a participatory approach to decision-making, this willingness to accept the initial potential misfunctions is promising for developing a healthy approach to adoption of AI applications that is not saddled by *over-expectation or over-pessimism*. Nevertheless, the uncritical attitude is worrying given the accelerating AI arm races among world powers and giant tech-companies, and various socio-political problems already take places the muddle the water including the health of democracies, climate change, population decline, and the lack of disadvantaged groups in decision-making bodies around the world.

Data availability

There is no data associated with this study.

Conflicts of interest

The authors declare no conflicts of interest.

References

Floridi, L. (2015). *The onlife manifesto: Being human in a hyperconnected era*. Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-04093-6

Ho, M.-T., & Vuong, Q.-H. (2024). Five premises to understand human–computer interactions as AI is changing the world. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01913-3

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. *Philosophy & Technology*, *35*(1), 17. https://doi.org/10.1007/s13347-022-00511-9

Mantello, P., Ho, M.-T., Nguyen, M.-H., & Vuong, Q.-H. (2023). Machines that feel: behavioral determinants of attitude towards affect recognition technology—upgrading technology acceptance theory with the mindsponge model. *Humanities and Social Sciences Communications*, *10*(1), 430. https://doi.org/10.1057/s41599-023-01837-1

Vuong, Q. H. (2025). *Wild Wise Weird. https://www.amazon.com/dp/B0BG2NNHY6* (4th ed.).