I’m a real boy! AI: the Pinocchio of technology?

Researcher perceptions on AI interaction and design through anthropomorphism, privacy, and purpose.

Sarah Vahlkamp

Dr. St. Jean

University of Maryland

INST 888 Spring 2021

AI is currently the field-a-la-mode. Researchers, conferences, think-tanks, and groups dedicated to its study are materializing regularly. While the field is shadowed by sci-fi myth and quixotic framing, more nuanced and considered research is undertaken by academics and industry professionals in the discipline. This review focuses on small aspects of interaction and design. It will delve into the discussions of intention, determinism, privacy, anthropomorphization, and teammate vs. tool with respect to AI. I will discuss areas in which researchers agree, and areas that are still being explored, including consciousness, embodiment, and surveillance. The review ends with an examination of the directions identified by researchers for further study, and a case for other directions AI research might take.

Based on the arguments put forward in these articles and books, I see a still relatively nascent field trying to find its footing on some central and fundamental issues. Currently, the most predominant theme is privacy. Privacy is looked at in a general and personal way, but there is room to take it into a more specialized realm, looking at issues of surveillance in the workplace in the face of ubiquitous AI. AI, by definition, requires and gets more accurate with more information. The trade-off is loss of privacy. People cede some of their privacy when signing terms of service. This is one thing when done for their own purposes, and when they make their own determination of a line (which is arguable in its own right, but beyond the scope of this paper.) When these systems are introduced in the work arena, though, some sovereignty is taken from the individual, and people are compelled to agree to gratuitous scrutiny or lose their livelihood.

In light of the pandemic, there is discourse about whether surveillance is inherently good or bad. Generally, surveillance is seen as bad, but public health workers might argue differently in the name of a thriving society. This merits ancillary research into the dichotomy. While I'm not advocating for a general pivot toward embracing surveillance, I think it's appropriate to

investigate if and when there are times that surveillance is actually beneficial, rather than a detriment to society.

Further, ethicists have focused questions of bias largely on algorithms, but there is still a fairly divisive debate around who or what holds responsibility for the creation and prevention of biased output. Is it primarily algorithms that introduce bias, or does it come down to data? Garbage in, garbage out is a relatively common maxim in machine learning and AI, and while true, does this miss a larger discussion on the way algorithms are designed and implemented? Ruha Benjamin looks at wider aspects of the tech industry, such as employee composition, shared history, and blind optimism that contribute to disproportionate adverse actions for some groups. (Benjamin, 2019) She describes what she terms “The New Jim Code”, and posits that while the effects may not be intentional, they are still perpetuating a system of inequality, and that the branding on new technologies as more objective plays fully into this system.

The debate over responsibility touches on the idea of determinism, a view aligned with Silicon Valley and akin to philosophical use of determinism. Related, but not to be confused with technological determinism, this use of determinism lands more on the side of technological evangelism, in that it adheres to the idealization of technology and nomological determinism. I will discuss how this is framed in the literature, with one side suggesting that nomological determinism is foundational to technology, and another suggesting that rather, this is a way to absolve oneself from having to put energy into understanding the implications of innovation. More discussion about this subject can be found by looking into arguments that surround the singularity.

This idea of determinism is linked to an area I see as relevant for more research, a look at the role of intention vs. outcomes in technology development. Discussed overtly by both Benjamin and Russell, themes of intention and outcomes pervade AI research, despite these themes not always being the explicit focus. Benjamin relates this directly to discourse in race and critical theory, suggesting that the issue was settled long ago in that field. (Benjamin, 2019) She

focuses on marginalized groups, seeing technology's place in concert with years of oppressive social policies and practices.

In the first book about AI safety written by a leading AI researcher, Russell addresses his concerns about the existential threat of AI to humanity, while keeping an open mind to the desire for, and in fact, benefits of continued exploration into artificial general intelligence (AGI). He discusses the AI model currently most closely associated with research, using it as a springboard to presenting his thoughts on changing the discourse from an intention-based to an outcome-based paradigm. He sets his argument as a defense against common critiques he has heard from leading AI researchers (leaving their names unpublished for saving face.) Despite framing this particular viewpoint as an outside opinion, Russell's arguments clearly come from a place of understanding, given his status within the field. Russell looks at the discussion from a more utilitarian perspective, arguing that developers need to think of third- and fourth-order consequences when programming AI, and understand that their intention won't necessarily be reflected in the outcome. Interestingly, he takes a view (famously associated with Nick Bostrom's paperclip example) that AI will so strictly adhere to the intent of the programmer that it will warp the outcome. (Russell, 2019) His focus on outcomes, rather than intentions, has been mirrored in other areas of the AI research community, seen in the section on safety and ethics now found on the deep mind site.

Russell's view maintains an explanatory nature of the debate, but really there needs to be further development on the issue of which matters more, intent or outcome, as in the technology realm, it isn't as settled as Benjamin suggests it is in other fields. This point will become central to other existential debates showing up around AI, particularly pertaining to legal matters.

Zuboff also addresses this matter, albeit framing her views in a discussion on economics. She argues that AI outcomes are the result of a class issue, going so far as to define a new stage of capitalism, one she terms Surveillance Capitalism. While I don't necessarily agree that she makes a strong enough case to defend this as a new form of capitalism, she does

unintentionally bring up the relationship between intent and outcomes by attributing both to a small cadre of powerful technologist leaders who at best hold malevolent intentions.

This literature looks at several facets of interaction. The first, related to anthropomorphism, is determining which form AI will take. Will it be embodied, or something less corporeal? In a small study looking at differences in human interaction with AI teammates, Yale researchers Bainbridge, Hart, Kim, and Scassellati examine human interaction with video-displayed AI vs. present, embodied AI. (Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. 2010) They identify similarities and differences in human judgment of AI between these forms. Although they saw cooperation with both forms of AI, they noted greater willingness to undertake what they term 'unusual requests' and more spatial concessions when interacting with the physically present, corporeal AI. While the study focused mostly on observation, it followed the interaction with a questionnaire, which indicated that subjects' attitudes reflected their behavior. The study makes the case that AI developers should consider physical presence in design decisions, and suggests that this impacts interaction enough to make it a central element in AI design.

Researchers Culley and Madhaven urge designers to consider issues of universalization, trust, and affect display when anthropomorphising AI. In their short paper, they identify a concern around attributing human intentions, feelings, motivations, and values. They convey discomfort with their assertion that these attributions are related to users assigning unwarranted trust to the AI. The area of trust in AI is discussed in much more detail in other research, but the takeaway from this article is that while anthropomorphism has its place, designers should be cautious about impairing users' decision-making abilities and misaligning trust calibration with the AI. (Culley, K. E., & Madhavan, P., 2013.)

This line of research implies a need for deeper looks into the way humans internalize their interactions with AI, and how to mitigate excessive unearned confidence. Culley and Madhavan also touch on the idea of universalizability and anthropomorphism. They point out subtle differences in natural language, body language, and cultural etiquette that belie the ability to

create universal embodied AI. Related to the heavily-researched concept of uncanny valley, these concerns imply a need to think about social psychology not just from the human to human interaction perspective, but also across human-AI interactions. The direction of this research also touches on a potential gap in literature around media and cultural influences, discussed later in the paper.

Can AI be an actual teammate, or is it just a tool? Dialogue surrounding this question is alive and well, even within my department. Shneiderman steps out against the possibility of AI teammates. In his “Three Fresh Ideas”, Shneiderman puts forth the ideas of Nass’s Fallacy, Shneiderman’s Conjecture, Teammate Fallacy, and Computers-in-the-loop Reality to summarize his beliefs. He cites cognitive psychologist Margaret Boden’s assertion that “Robots are simply not people” in defense of his stance. Others emphasize human-agent teaming, or human teaming with agents who are not anthropomorphized (Parasuraman, R., Sheridan, T. B., & Wickens, C. D. 2000 & Sukthankar, G., Shumaker, R., & Lewis, M. 2012.) Still others are looking at teammate structured interactions between humans and social robots. (Bartneck, C., & Forlizzi, J.2004 & Breazeal, C. 2003.) This discussion closely aligns with discussions about augmentation and replacement, although there seems to be more push toward the former. It’s not clear that there will ever be agreement on this subject, but I suspect that organizations are still interested in the idea, and if they are interested in it, it should be researched in order to avoid common and unforeseen pitfalls. Social and cognitive psychologists (like my pseudo-advisor, Dr. Susannah Paletz and team, paper still in development) are moving forward with discussions on human-AI teaming, and conducting studies that apply lessons learned from their fields to human interactions with AI, whether embodied, social, or agent-based.

At the core of all this discussion, there is a question of consciousness, which brings up questions on legal authorities and liabilities, moral and ethical issues, and relatedly, even AI rights. There is significant disagreement as to the ability to create consciousness, the ethics and morality behind it, and even the definition of consciousness. This topic engages the curiosity of

researchers in fields as diverse as computer science, psychology, philosophy, sociology, and theological studies. It has entire books dedicated to its study, so to boil it down, what I'll say here is that often, but not always, where a researcher lands on this subject will influence his or her views on the other subjects discussed in this review.

This review, and the literature it's based on, covers a lot of ground. Given the expanding interest in the field, and the number of calls for papers, conference invitations, and research invitations sent out related to AI, it may seem like all areas are being covered, but we are just starting to see good numbers of usable results. From the papers reviewed here, alone, I see opportunity to look deeper into the importance of intent vs. outcome, put stronger emphasis on privacy and surveillance in different contexts, and ideas of culpability with respect to determinism and existential threat.

The focus on whether AI can be called a teammate or not is myopic, and we need to address this from the underlying assumption that human-agent interaction is going to become important to all our lives. In some ways, it is already happening, but in others, there is still time to positively affect its emergence. I suggest there is room to move forward as if it is happening anyway and concentrate on elevating research of applied social science lessons to AI.

The final area for research I'll identify in this review is to look at cultural and media influences on perception of AI. Understanding that sci-fi literature, movies, and tv has obviously influenced how people welcome or push away AI, I think that can be distilled down in support of disagreements over anthropomorphization. For example, I started to look at cultural composition of various time period as it related to popular movies. This could be an interesting area to delve further into, and perhaps identify why and how humans create their AI constructs. This also addresses the cultural differences seen in areas like trust and anthropomorphization.

No matter the direction, AI is at the forefront of current technological research. It is looked at through the lens of machine learning, social robotics, and more and more, through social science constructs. The growing attention AI ethics has been getting recently, and particularly in

the past year, bodes well for more nuanced looks at human-AI interaction. Shifting tones in the computer science world, as shown through Stuart Russell's work, indicate an opening for more entrenched interdisciplinary research. Thought experiments and hypotheticals are becoming more real all the time, and real research on these subjects is being taken more seriously than ever before.

Citations

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2010). The Benefits of Interactions With Physically Present Robots over Video-Displayed Agents. *International Journal of Social Robotics*, *3*(1), 41–52. https://doi.org/10.1007/s12369-010-0082-7

Bartneck, C., & Forlizzi, J. (2004, September). A design-centred framework for social human robot interaction. In *RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No. 04TH8759)* (pp. 591-594). IEEE.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, *42*(3-4), 167-175.

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior*, *29*(3), 577–579. https://doi.org/10.1016/j.chb.2012.11.023

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Shneiderman, B. (2020a). Human-centered artificial intelligence: Reliable, safe and trustworthy. *International Journal of Human-Computer Interaction, 36(6),* 495-504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B. (2020b). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction, 12*, 1-16.

http://DOI:10.17705/1thci.00101

Sukthankar, G., Shumaker, R., & Lewis, M. (2012). Intelligent agents as teammates. *Theories of Team Cognition: Cross-Disciplinary Perspectives*, 313-343.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power:* New York: Public Affairs, 2019.

Yudkowsky, E. (2002). The AI-box experiment. *Singularity Institute*.