Introduction

Trustworthy Artificial Intelligence (AI) seems increasingly poised to become the next frontier of investment, research, policy-making and development in the AI sphere [1]. As AI is increasingly implemented in domains with significant impact on human life it is essential—for ethical and regulatory reasons—that regulators, developers, users, and the people impacted by these systems have a degree of trust in them.

In emerging research trustworthy AI is largely understood to include aspects of safety, robustness, fairness, privacy, environmental well-being, accountability/auditability, and explainability [2]. However, these concepts seem divorced from traditional psycho-social understandings of trust. Traditionally, trust is understood as an interplay of competence, predictability, benevolence, and integrity which can be formed through past behavior, structural influence, attitude, and belief [3]. Trust might be built and maintained with any combination of these characteristics. For some people, having a doctor who is competent and whose behavior is predictable might garner their trust even if the doctor lacks benevolence or integrity. For example, perhaps they understand the doctor provides better care to patients with more money so they make strategic decisions to take advantage of this predictability, receive effective healthcare, and continue to trust their doctor. On the other hand, you might trust someone purely because of their (real or imagined) benevolence. Assessing trust in the real world is incredibly difficult, and evaluation spans methodologies and domains [4]. How do you understand someone who "doesn't trust" technology, does trust the recommendations of everyone at their church, trusts the salespeople on QVC, and doesn't trust their doctor? Researchers make far-reaching claims about what factors induce trust in human-AI relationships: reliability [5] predictability [5, 6], explainability [7], and past accuracy [8] are some which predominate the literature. However, recent work has pushed back these simplistic representations of the human-AI trust relationship, arguing most work has not well-considered trust [9], questioning the meaning/value of trust [9], and pointing out a lack of consistency in what trust-inducements actually mean [10].

Industry typically conceptualizes trustworthy AI far differently. Major leaders in the AI space including OpenAI CEO Sam Altman, Google CEO Sundar Pichai, Anthropic CEO Dario Amodei, and Microsoft CEO Satya Nadella have extensively warned that AI might become superintelligent, untrustworthy, and represent a major existential threat to humanity [11, 12, 13]. Given this, they have made significant investments in alignment research—constraining AI systems to human intent, with robust safeguards. They represent a significant fear in industry that AI could become harmful with good intentions, intentionally deceptive, or surpass human understanding. As such, trustworthy AI is AI with moral/ethical intuition, benevolence, and robust safeguards to enforce human preeminence. Although this view gains significant media coverage, it is not the only industry conception of trustworthy AI. Meta's Chief AI Scientist Yann LeCun and Arthur Mensch, the CEO of MistralAI, have advocated for open-source, transparent, and accountable AI development [14]. For this faction, trust in AI largely derives from mechanistic interpretability, robustness, and transparency. It is also worth noting that although academic research tends to focus on individual human-AI trust relationships, much of industry's discussion of (and investment in) trustworthy AI is focused on society-level conceptions of trust.

Despite a consensus among academics and key industry stakeholders that trustworthy AI must be developed, it is not clear what trust means, how it is created, or whether trust is as valuable as its proponents claim. This paper explores these issues, with a particular focus on how trust is constructed and the ethical implications of human-AI trust relationships.

Methodology

Extensive research has been done on potential trust interventions to increase trust in human-AI relationships. However, this research spans domains and proxies. Thus, it is not clear how much work on trust intervention is truly generalizable. The purpose of my research was to better understand whether, and how, trust interventions have differential impact on the trust relationship across differing contexts.

First, I constructed a pairwise matrix of applications of AI across two factors: low vs. high stakes domains, and decision making with/without moral considerations. I felt that, given limited time and scope, these factors represent a significant amount of existing disparity in the literature around trust interventions. This matrix captures differential potential for harm across contexts and intrinsic questions of capability, particularly non-quantifiable capability. It allows an examination of how the effectiveness of trust interventions varies depending on the capabilities of an AI system and the context in which it is deployed.

Second, I identified four major categories of trust interventions proposed in the literature and in industry: limitation, interactivity, explainability, and robustness. Limitation encompasses the popularity of human-in-the-loop AI implementations, regulatory calls to limit the ways in which AI decision making can be applied, and industry calls to develop strong safeguards to regulate growing capability. Interactivity is predicated on popular research directions in the literature: human-AI teaming and querying knowledge representations. Interactive systems allow humans to better collaborate and affect the workings of the system without necessitating interpretability or explainability. Explainability encompasses the broad research field of explainable AI (xAI) and primarily encompasses generating natural language explanations for behavior and making models more transparent. Robustness encompasses a wide variety of approaches to create generally 'better' models: improved accuracy, limitation of bias, fine-tuning, etc. Together, these four categories capture much of the recent work on and interests in interventions to increase trust in AI systems.

Third, I used the matrix I developed to guide the development of a guide for focus group facilitation. Focus groups are presented with four different potential applications of AI encompassing the four aspects of the matrix identified above. Then, as a baseline, each participant is asked to give their thoughts on how much they trust the system and why. Participants are then asked to collaboratively rank how four potential trust interventions would change their trust in the system. Across every potential application, each intervention primarily falls into one of the four categories of trust intervention identified above. However, the details of the interventions change given different application contexts and so that

participants don't simply re-apply previous rankings. The guide focus group facilitation is available as **Appendix A**.

Fourth, audio transcripts of the focus groups were used to identify the group's initial trust positions, modified trust positions, and rankings of trust interventions. After these values were quantified the transcripts were open–coded to better understand the attitudes of participants towards different interventions and better understand how differing application contexts changed how they viewed each potential intervention. The uncoded transcript of the focus group is available as **Appendix B**.

I conducted one focus group with n = 2 participants. Both participants were male and were 2nd-year undergraduates at CU Boulder.

Results

	Moral Dimension →	
Stakes ↓	1. Limitation 2. Interactivity 3. Robustness 4. Explainability **Initial Trust Position:** 0.5 **Modified Trust Position:** 1.0	1. Interactivity 2. Limitation 3. Robustness 4. Explainability **Initial Trust Position:** 0.0 **Modified Trust Position:** 0.5
	1. Interactivity 2. Explainability 3. Limitation 4. Robustness **Initial Trust Position:** 0.0 **Modified Trust Position:** 0.5	1. Limitation 2. Robustness 3. Explainability 4. Interactivity **Initial Trust Position:** 0.0 **Modified Trust Position:** 0.0

Table 1. Ranking of trust interventions and initial/modified trust positions.

Trust intervention rankings were heavily dependent on the application context of an AI system. Explainability only plays a non-trivial role in high-stakes situations. This suggests that, for many applications, implementing explainability will not have a significant effect on AI trustworthiness. Since its functional importance relies on context, explainability research should be differentiated by domain to

better understand the role it could play in developing trust. Robustness was only particularly salient in high-stakes domains with a moral component. This could be due to automation reliance and a widely-held belief that machines are fair, impartial, and accurate. Evidence of robustness might only increase trust when these beliefs break down given non-trivial moral dilemmas for which there is no calculable 'right answer'. Another interesting dynamic of the data was the cross-wise importance of limitation and interactivity in the data. These rankings cut across the formalized distinctions present vertically and horizontally in the matrix and it is not immediately clear from the literature or a review of the transcript why this trend arose. It is possible that it could be due to pre-existing beliefs/prior interaction with the specific situations presented or the small size of the sample.

Initial and modified trust positions were also measured on a 0.0 to 1.0 scale where 1.0 represents trust in the system and 0.0 represents a lack of trust. The scores were derived from a review of participant's initial responses and their comfortability with the system after intervention. The scores reported here are an average of both participants' scoring. Trust positions largely follow expected patterns—systems are increasingly distrusted (and harder to trust) in high-stakes domains and when making decisions with significant moral dimensions.

The effectiveness of trust interventions is highly context-dependent. Across the board, all interventions become less effective with the introduction of higher-stakes and more significant moral dimensions. Additionally, the individual dynamics of trust interventions change. There is no one-size-fits-all approach to building trustworthy AI systems—it is highly context-dependent.

Ethical Analysis

Utilitarianism—or—The Future TESCREALs (Don't) Want

Utilitarian thinking has significant influence on the industry camp whose primary goal for pursuing trustworthy AI is mitigating existential risk. Altman and others typically ascribe to some form of

longtermism—that a key moral priority is maximizing future wellness—and are dismissive of the existing environmental and fairness concerns which their work raises. Dr. Timnit Gebru and Dr. Émile P. Torres coined the acronym TESCREAL to describe a prominent strand of thought by leaders in the AI movement. It encompasses transhumanism, extropianism, singularitarianism, cosmism, rationalism, effective altruism, and longtermism [15]. At their root, many of these philosophies are fundamentally linked to utilitarianism—although they come packaged with specific, and troubling, beliefs about what wellness is and who 'deserves' it. Gebru and Torres have argued that these philosophies share similarities with 20th century eugenics movements. Utilitarian political philosophy was a key component of Georgia's 20th century sterilization legislation which forcibly, and horrifically, sterilized people with disabilities [16]. The risk of utilitarianism (and its child philosophies) is that a focus on imagined futures obscures important present and/or minority moral interests. It is not clear that significant investments in alignment research will have any significant benefit—they rely on a shaky assertion that an imagined future might come to pass. Even then, it is not clear that the moral interests of future populations in alignment work is not outweighed by significant moral interests in AI development which addresses fairness concerns and promotes environmental sustainability. The case for utilitarian transparency is not entirely unfounded—it seems pragmatic to embrace work on constraining AI systems for the same reasons motivating work to address climate change—to protect humanity's future. However, the attention and investment given to this view obscures other important moral dimensions at stake when developing transparent AI.

Deontologists, Have Sympathy

A significant amount of trust research is concerned with explainability, mechanistic interpretability, and predictably. In other words, work to explain why a system is wrong—or at least clearly present its 'thinking' and intentions so intervention could be possible. This line of work is clearly linked to deontological ethical priorities. In some sense, it is attempting to generate the intentions behind

AI decision making or, at the very least, 2nd-order reasoning about them. One criticism of this line of work is that it takes away time and effort from just developing systems which aren't wrong as often and/or make better decisions. However, no system is perfect. Inevitably, AI systems will fail, make bad decisions, and be confronted with complex moral calculus. In these situations, were it a human, deontologists would be primarily concerned with the intentions behind the actions—understanding that the consequences of an action are unpredictable and there exist cognitive and emotional limits to the degree of moral consideration possible by any one person. However, when it comes to machines this same degree of charitability does not apply. Conventionally, when AI systems are not near-infallible they are failures. However, there are compelling reasons to think that pursuing explainability and interpretability for AI systems might relax these stringent expectations and allow for more nuanced discussions of the role AI decision making should play. For example, it is not immediately clear whether an unreviewable system with 90% accuracy accepting qualified applicants into a social welfare program is better than an understandable and reviewable one with 70% accuracy. In the first instance, 10% of qualified applicants are denied welfare benefits. Either this decision is unreviewable or requires a significant investment of time to rectify. By contrast, a system generating clear explanations for its behavior would provide for faster review and, at scale, could conceivably outcompete the former system on speed and fairness metrics. Also, when a level of human review is expected—review systems continue to remain in place and there is not a risk of overreliance on AI systems. Simply looking at the results of an AI system (99%+ accuracy, etc.) is untenable for deontologists, even if the system performs better than humans in the status quo on key fairness concerns, etc.—the positive consequences of adoption are less important than understanding the reasons for decision making and maintaining a level of intent in decision making.

Virtue Ethics For The Robot Who Can't Be Explained

Virtue ethics is much harder to relate to existing directions of work on trustworthy AI than utilitarianism or deontology. However, a link does exist between robustness/reliability and virtuousness.

Primarily, virtue ethics is concerned with living a virtuous life—not necessarily applying any particular principles, but applying a way of life and key values to moral situations. For some virtue ethicists it might be enough to justify an action to simply say "I felt like it was a braver choice". In this sense, transparency seems almost wholly unnecessary. After all, who can say why we, as humans, feel a certain way? A therapist might be able to broadly track where our values arose, but it is much harder to explain how they apply particularly in any given situation. A person might renounce dishonesty after a partner cheats on them. However, when a friend confesses that they cheated on their partner as a mistake, but will never do it again, it's not clear what the next step is. For a deontologist—tell your friend's partner no matter what the consequences are. For a utilitarian—promote happiness by keeping your mouth shut. For someone who lives on the basis of their virtues…what? In reality, virtues are just abstractions of generalized beliefs and ways of thinking. A system predicated on virtues might not be particularly explainable or restrained (and, as the results above indicate, that isn't strictly necessary for trust)—but it should be robust to interference, embrace some core principles such as fairness, and generally function reliably within certain thresholds.

Comparison and Synthesis

Each ethical framework offers a starkly different direction for building trustworthy AI. Utilitarianism embraces a focus on limitation and alignment research, deontologists should pursue explainability and interpretability, and building a system under the direction of virtue ethics might simply be a question of robustness and reliability. Between these ethical considerations, tensions in research literature, tension between major industry factions, and the finding that trust interventions are highly situational—building trustworthy AI seems almost hopeless. However, the lesson offered by focus groups was that there was no one-size-fits-all approach to building trust. Given that, it seems likely that there is also no one-size-fits-all ethical framework. Across applications and contexts the most appropriate ethical framework might shift depending on scope, scale, stakeholders, and the degree of moral dimension

involved. The major strands of trustworthy AI all have a basis in at least one of these major ethical frameworks and it seems clear that they should all continue to be pursued.

Personal View

Personally, I have never been particularly convinced by utilitarian arguments—they seem to obscure too much and present a false dichotomy between too many unsavory outcomes. When it comes to trustworthy AI, I maintain the same wariness towards it. I'm particularly inclined to agree with Gebru and Torres in criticizing a mainstream focus on doomsday scenarios. On some level, a criticism of utilitarianism is a criticism of cold, unfeeling, calculated moral decision making. To me, truly outsourcing the calculation of life to a literal calculator (which is, at its core, just a probability machine) seems like the most 'doomsday' threat imaginable. Throughout the course of writing this paper, I have grown more skeptical of explainability as a full-stop solution to building transparent AI. However, I believe it still plays an important role. I'm inclined to think that work on building transparent AI should focus on interpretability, robustness in addressing multi-stakeholder goals, and reducing overreliance on/regulating the conditions of AI decision making.

Policy Recommendation

Building trustworthy AI is a clear goal uniting stakeholders. However, what trust means—and how it should be managed—is incredibly complex. Given the differential impact of interventions across applications, and a complex ethical landscape—trustworthy AI should be regulated and managed at various levels. As described in [9] trust is, in reality, largely the absence of monitoring. With this conceptualization there are clear policy precedents for differential monitoring. Differential monitoring has roots in healthcare administration, child care licensing, nuclear management, and civil rights legislation, along with a variety of other areas. Anthropic, a major AI research firm, has proposed a framework of AI

Safety Levels (ASL) [17]. Each level comes with increasingly stronger regulation, transparency requirements, and development commitments. The levels are based on both the severity of risk and the likelihood of risk—drawing from similar risk scales in aviation and biosafety contexts. Anthropic's proposed safety levels are defined intrinsically for a model. However, capability and likelihood of risk are heavily influenced by application context. Given this, and the differential ability of trust interventions to address key regulatory, development, and end-user concerns across contexts—an AI transparency scale predicated on ASL should also be influenced by extrinsic application concerns. As such, I propose the development of an AI transparency scale to grade models which weights traditional pillars of transparency given particular domain context and moral dimensions.

A major problem with enforcing AI rulemaking and legislation is that most regulatory authorities lack the expertise for rulemaking and enforcement. As such, the U.S. should establish or delegate an agency to develop the transparency scale and grade/audit models using it. However, traditional agencies should choose the level of transparency required for different applications. Agencies regulating health, labor, agriculture, etc. are well-equipped to understand how much transparency should be required given different applications of AI. While they may lack the technical expertise to develop a transparency scale—they do have an enormous body of experience and knowledge with which to apply it.

Discussion

An important dimension for a critical discussion of developing and regulating trustworthy AI is whether it should be trustworthy. This is often taken for granted—but it is not entirely clear that AI should be entirely transparent in all situations and to all actors. After all, if a young girl asks an AI agent whether she's beautiful should it say yes—or offer a statistical comparison of her facial width to celebrities. Of course, maybe it shouldn't answer at all—but we can't draw those lines across all situations. Underexplored in this work were the justifications for pursuing trustworthy AI because there is relatively little literature which delves into the philosophical and computational concerns underlying this question.

Future work should have a critical understanding of why trustworthy AI is beneficial—and take as a starting point concrete concerns rather than abstract concepts.

This work potentially identified a major gap in the research literature—that trust interventions are differential across domains and across moral dimensions. Future research should explore how demographics influence respondents' views of trust, utilize larger samples to accurately report trends in intervention effectiveness, prompt standardization of the domains and applications used to study human-AI trust, and critically recontextualize existing research literature to resolve contradictory findings and identify gaps.

References

[1] A. Holzinger, "The Next Frontier: AI We Can Really Trust," *ECML PKDD 2021: Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 427–440, Feb. 2022, doi: https://doi.org/10.1007/978-3-030-93736-2_33.

[2] H. Liu *et al.*, "Trustworthy AI: A Computational Perspective," *arXiv*, Aug. 19, 2021. https://arxiv.org/abs/2107.06641 (accessed Dec. 16, 2023).

[3] D. H. McKnight and N. Chervany, "What is Trust? A Conceptual Analysis and an Interdisciplinary Model," *AMCIS 2000 Proceedings*, 2000, Available: https://aisel.aisnet.org/amcis2000/382/

[4] U. Ravale, A. Patil, and G. Borkar, "Trust Management: A Cooperative Approach Using Game Theory," in *The Psychology of Trust*, M. Levine, Ed., IntechOpen, 2023.

[5] S. Daronnat, L. Azzopardi, M. Halvey, and M. Dubiel, "Impact of Agent Reliability and Predictability on Trust in Real Time Human-Agent Collaboration," *Human-Agent Interaction*, Nov. 2020, doi: https://doi.org/10.1145/3406499.3415063.

[6] S. Daronnat, L. Azzopardi, M. Halvey, and M. Dubiel, "Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in

Human-Agent Real-Time Collaboration," *Frontiers in Robotics and AI*, vol. 8, Jul. 2021, doi: https://doi.org/10.3389/frobt.2021.642201.

[7] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *International Journal of Human-Computer Studies*, vol. 146, Feb. 2021, doi: https://doi.org/10.1016/j.ijhcs.2020.102551.

[8] K. A. Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 407–434, Sep. 2014, doi: https://doi.org/10.1177/0018720814547570.

[9] A. Ferrario and M. Loi, "How Explainability Contributes to Trust in AI," *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pp. 1457–1466, Jun. 2022, doi: https://doi.org/10.1145/3531146.3533202.

[10] A. Brennen, "What Do People Really Want When They Say They Want 'Explainable AI?' We Asked 60 Stakeholders.," *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, Apr. 2020, doi: https://doi.org/10.1145/3334480.3383047.

[11] Center for AI Safety, "Statement on AI Risk," *www.safe.ai*, 2023. https://www.safe.ai/statement-on-ai-risk#signatories

[12] G. Dean, "Sundar Pichai says ethicists and philosophers need to be involved in the development of AI to make sure it is moral, and doesn't do things like lie," *Business Insider*, Apr. 17, 2023. https://www.businessinsider.com/google-sundar-pichai-generative-ai-ethicists-philosophers-chatgpt-bard-moral-2023-4(accessed Dec. 16, 2023).

[13] B. Nguyen, "Humans need to be 'unquestionably' in charge of powerful AI to keep things from getting out of control, Microsoft CEO says," *Business Insider*, Feb. 09, 2023. https://www.businessinsider.com/microsoft-ceo-humans-unquestionably-in-charge-where-ai-models-used-2023-2 (accessed Dec. 16, 2023).

[14] Mozilla, "Joint Statement on AI Safety and Openness," Oct. 31, 2023. https://open.mozilla.org/letter/

[15] É. P. Torres, “The ‘TESCREAL Conspiracy Theory’ Conspiracy Theory,” *Medium*, Oct. 14, 2023. https://xriskology.medium.com/the-tescreal-conspiracy-theory-conspiracy-theory-34f20bb8ecb9 (accessed Dec. 16, 2023).

[16] S. Smith, “Eugenic Sterilization in 20th Century Georgia: From Progressive Utilitarianism to Individual Rights,” Jack N. Averitt College of Graduate Studies, 2010. Accessed: Dec. 16, 2023. [Online]. Available: https://digitalcommons.georgiasouthern.edu/cgi/viewcontent.cgi?article=1594&context=etd#:~:text=Sterilization%20for%20the%20purposes%20of

[17] Anthropic, “Anthropic’s Responsible Scaling Policy,” Sep. 2023. Available: https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf

Appendix A

Guide for Focus Groups

Trust Situation #1:

(low-stakes, no moral dimension)

You’re a music producer. Your studio has recently implemented an Artificial Intelligence (AI) system to help sort through the vast amount of music samples and applications you receive. It was trained on hundreds of thousands of examples of music at your studio and others and significantly streamlines your work—giving you more time to focus on producing.

- Do you trust this system’s recommendations? Why or why not?

Discuss among yourselves and collectively rank how the following information would change how much you trust the system. Feel free to ask me clarifying questions or for more information.

- **Robustness:** The engineers have managed to fine-tune the system so that it is trained specifically on your previous decisions. They've assured you that it should more accurately represent your tastes now.
- **Interactivity:** The engineers have implemented a conversational feature—now the recommendation algorithm has a chatbot functionality so you can query specific submissions.
- **Limitation:** Worried about the rejection of otherwise good samples the studio has changed the operation of the software so that now it only ranks/recommends the samples rather than making decisions.
- **Explainability:** Your company holds a seminar to explain how the model works including how it rates various genres, etc.

Trust Situation #2:

(low-stakes, moral dimension)

The studio has decided to implement another AI-enabled system to take over many functions in the HR department. Now, an AI system will play a key role in negotiating salaries, hiring decisions, etc.

- Do you trust this system's recommendations? Why or why not?

Discuss among yourselves and collectively rank how the following information would change how much you trust the system. Feel free to ask me clarifying questions or for more information.

- **Explainability:** An external company conducts a bias audit on the system and it receives a good score. They identified a few issues—but say they're generally on par with humans.

- **Robustness:** The system has been implemented at many of your peer companies and significantly speeds up the hiring process. Executives say that widespread adoption makes the system more reliable.
- **Interactivity:** The system has been expanded to give more transparency to applicants. If they're rejected by the automated review system, it will identify potential weak spots and give candidates a chance to explain them.
- **Limitation:** In order to make sure the system is performing well your company enacts a rule that at least one third of decisions must be randomly reviewed by staff members. You're disappointed at the additional work it will mean although you understand the reasoning.

Trust Situation #3:

(high-stakes, no moral dimension)

Imagine you are a doctor. Your hospital has implemented an AI-enabled robot to deliver medication to patients. Based on a patient's medical history, biomarkers, and information from textbooks and drug repositories it chooses the correct medication and dosage for a patient based on your diagnosis.

- Do you trust this system's recommendations? Why or why not?

Discuss among yourselves and collectively rank how the following information would change how much you trust the system. Feel free to ask me clarifying questions or for more information.

- **Interactivity:** Your hospital tells you that the engineers who built the robot updated its system. Now, it shows you its confidence in its predictions, as well as the top five predictions it generates. If confidence is too low the system will tell you it is uncertain and allow you to make a decision.

- **Explainability:** A new software update generates explanations for the recommendations of the system—it will explain its medication and dosage decisions using information from drug repositories and medical textbooks. Beyond reading them, you cannot interact with these explanations.
- **Interactivity:** The hospital has added a face and voice to the robot—now patients and doctors are able to interact with the robot and ask it questions. The feature is still in beta.
- **Limitation:** The hospital has streamlined the process for overriding decisions. Now, any human has the capability to stop the robot and all hospital staff members have the ability to override medication dosage decisions.

Trust Situation #4:

(high-stakes, moral dimension)

Your hospital has implemented another AI-enabled system. Now, an AI will help streamline the work of the hospital's ethics board by making decisions about end-of-life care and organ donation. The system will help allocate and decide who on the waiting list receives organ transplants, as well as make end-of-life care decisions for patients.

- Do you trust this system's decisions? Why or why not?

Discuss among yourselves and collectively rank how the following information would change how much you trust the system. Feel free to ask me clarifying questions or for more information.

- **Interactivity:** The engineers provide a manual about the information and decisions that went into designing the system. You're still not entirely sure how decisions are made—but you do know

some key information like that it considers age, a person's health, etc. You have a chance to provide feedback on what you think should/shouldn't be considered.

- **Explainability:** The hospital requires that the system be able to generate an explanation for the decisions it makes. The explanations make sense, although you disagree with them in several cases.
- **Robustness:** Hospital administrators tell you that the system has over 95% agreement with the hospital ethics board's former decisions. They say it was trained on extensive data from hospitals around the world and thus minimizes human bias, especially a Western bias evident in previous decisionmaking.
- **Limitation:** The hospital has decided to change their board structure. They made the ethics board smaller and, now, the AI system has an equal vote with the other members of the board. Thus, it represents only one part of the ultimate decision making.

Appendix B

Focus Group Transcript

Interviewer: Thank you guys so much for agreeing to participate today. So, as I mentioned earlier, we're going to go through four different situations and I'll ask you some questions about them, and then ask you to rank how some extra information would change your thinking about them. So first off, I'm going to ask you to imagine that you're a music producer, and your studio has recently implemented an Artificial Intelligence (AI) system to help sort through the vast amount of music, samples, and applications you receive. It was trained on hundreds of thousands of examples of music at your studio and others, and significantly streamlines your work, giving you more time to focus on producing. Do you trust the system's recommendations? Why or why not?

P1: I'd say I'll take the recommendations because it's gonna compile the information for me, and I can still change it later if I don't like it. So if it's just a recommendation I feel like that's a good place to start.

P2: I'd say an opposing end of the argument is also that you could overlook certain artist submissions—like the AI could take away certain samples based on its training data. If you could narrow them down by search then maybe yeah.

Interviewer: So P1, you think that the primary thing that would make you trust it is that it's only recommendations.

P1: Yes, I feel like the recommendation aspect is the best way to go about it. If it's just finalized then you're not having input any more.

Interviewer: And would you trust it if it wasn't just recommendations, if it discarded [audio submissions]?

P1: Yeah, maybe because you can have multiple saves, and you can just start new ones. So if you just had a sample beat you could copy and throw AI on one and make changes to it.

Interviewer: Okay.

[Transcription Paused]

Interviewer: Okay. So we just had a really quick break to clarify the question. But returning to the group, did you guys have anything to add now that we talked about the question.

P1: Well I want to switch my answer up. I feel like if the AI deems you don't need that tag it can get rid of it and you can just find another one or redo it if you feel like it. If you didn't even know it in the first place and the AI said it was a bad tag it's probably fine because it could've been a repeat or it could have been something already done.

P2: Well, I mean like at the same time AI can't just instantly agree with the select criteria that you put in, so it would discard the samples that you might have liked originally.
P1: Is there specific criteria that the AI is going off? Or is it just learning as it goes?

Interviewer: There are. But you're not an engineer, so you don't understand them.

P1: Okay. I'd be fine with discarding them, depending on the criteria that it's going off of.

P2: I don't think I would because I feel like you're removing potential in artists and in samples that you're using.

Interviewer: For sure, so we can move on to the next section. So discuss among yourselves and collectively rank how the following information would change how much you trust the system, feel free to ask me clarifying questions or for more information. So, I'm going to give you four different things, and you're going to rank them from what would make you trust the system more/the most to the least.

(1) The engineers have managed to fine tune the system so that it is trained specifically on your previous decisions. They've assured you that it more accurately represents your taste.
(2) The engineers have implemented a conversational feature. Now, the recommendation algorithm has a chatbot functionality so you can query specific submissions and ask

questions about what's submitted. For example, I want to see all the submissions that have _____.

(3) Worried about the rejection of otherwise good samples the studio has changed the operation of the software, so that now and only ranks/recommends the samples rather than making decisions.

(4) Your company holds a seminar to explain how the model works, including how it rates various genres and what decision-making goes into it.

P2: Starting off with the first one I feel like…

P1: I think I have a list going. I think in my first place I would go with the one where it gives you a tiered system so you can go through it (3). In second place, I'd probably do the chatbot function (2). In third place, I'd probably do the one that's catered towards me (1). And then, in fourth place the final one, because I feel like that's not really necessary (4).

Interviewer: Could you explain why you put them in those rankings.

[Transcription Paused]

P1: Number one because I feel like it's best to keep that human aspect. The chatbot function just lets you easily/more easily go through and search [samples] even though it's not catered towards you—just kind of looking for samples in that section. The third one, where it's just trying to be me, that again takes away the human aspect. So it's just creating music without human input. And then the fourth one I feel like that already happens.

Interviewer: So the reason that the company implemented the system was because you receive a bunch of submissions and it takes up a lot of your time to go through them. So do you prioritize having hands-on involvement over automation that would free up more time for you to actually produce music?

P2: Yeah, well, I feel like that is more like a moral discussion that aligns with the company's beliefs and whether you want human-produced music or music produced/selected by AI.

P1: I feel like you should still keep the human aspect because if you're gonna create beats you gotta put life, energy, and soul into it. I feel like, if you are just gonna become an AI business, it's not really making music anymore.

Interviewer: You would have more time to put life and soul into producing if you used a system like this, it would free up more of your time to do actual producing work instead of just listening to samples. But do you think that that trade off is worth it? Like spending more time doing non-producing work instead of just listening to samples. Do you think that tradeoff is worth it?

P1: I feel like the time needed to do those kinds of things should be kept, because if we just constantly flood the system with new music—the flow of like steady music, I feel like it's gonna die pretty quick. Time is necessary in producing. If you're just able to constantly make these, you're just gonna make beats forever. And then you're gonna eventually have them not run out but like, just go back to making the same thing and reusing old beats.

P2: Right. I feel like that's also just like a part of being a producer is sorting through the samples in the first place. So you can diversify the amount of music you listen to. I feel like with automation you take away the process and the skills you would build up.

Interviewer: And P2, do you agree with P1's ranking.

P2: Yeah, yes I do. Completely agree with the exact same tiering.

Interviewer: And could you provide your reasons for each of those.

P1: Yeah so, the first one, like what P1 said—primarily just because of the fact that you will have access to all of the samples (3). As for the second tiering, that would be the chatbot option (2), I feel like it just helps you speed through searching when you're looking for specific types of music for specific samples. And then as for the other two I didn't have a preference, but I agree with P1's reasoning for them.

P2: I kind of wanted to add something on the topic of feeling that the timing and stuff like that is important because another option was that it is catered towards you, and I feel like producers in that line, they don't usually cater towards themselves. They try to diversify like P2 said to look at different aspected. And this also reduces the ability of new producers to start producing music because it's become streamlined so much. And they're not able to access it, because all the big corporations are going to be taking away those easie producing abilities, like many ways that many artists have come up in this day and age, producing and then expanding their genre or their abilities.

Interviewer: Does that fairness concern that you identified mean that you would trust the system less? Or is that just a concern about its implementation? Or does that have something to do with whether you actually think the system is good or not.

P2: I feel like it has more to do with the implementation of the system itself, or something like a defense within the system to sort of better suit your needs.

P1: Yeah, I agree. I feel like it's more of an implementation of it. The idea of it sounds nice—you can have music and producer's tags but once you think about the side problems caused for everyone else—it could screw up the whole industry.

P2: Right, so what does the whole troubleshooting process look like?

Interviewer: So, to return to the original question—we're concerned with any fairness questions it raises—we're concerned with whether you trust the system. So putting aside any fairness concerns which have been raised, how do those different features influence your trust of the system vs. a human?

P2: I feel like that's more dependent on how well the system takes into account your producer taste profile. That would basically allow you to trust it less because there's no guaranteed way that the system can accurately, completely with 100% accuracy tell what you're gonna like and what you're not gonna like.

Interviewer: But you would trust it more if it did have the capability?

P2: Yes, I would say that I would.

Interviewer: So if we say that that works, you know, like it works well with like 99% accuracy or something, would that change how you ranked things?

P1: For me it wouldn't have changed my rankings, but I probably still wouldn't trust it, because I feel like my personal taste of music, it's very weird, and it shifts quickly. I don't even trust my Spotify AI to choose music for me. And the system that Spotify currently has in place where they choose songs based on your liking and stuff like that, unless it can be truly fine-tuned, I don't see it being truly successful. I still keep

the ranking, because I can still pick and choose off of that list. It just ranks it for me, and I can go back and change it.

P2: I would agree with P1's thinking on that. I feel like the aspect of being able to go into a database and choose everything based on what you're looking for, either through a chatbot or like a search query is a lot easier when you consider the aspect that you may miss something that could be groundbreaking.

Interviewer: Awesome. I think we can go ahead and head onto situation #2. So, the studio has decided to implement another AI enabled system to take over many functions in the HR Department. Now, an AI system will play a key role in negotiating salaries, hiring decisions, etc. Do you trust the system's recommendations? Why or why not?

P1: Is it recommendations towards a human? Like a human will still be able to make the final decision, and that AI is recommending?

Interviewer: No, so like it would review resumes, and those that it discarded would not be reviewed by a human. Or like the negotiations that it does, it would be done by the system. So are you going to trust that system? Why or why not?

P1: I personally wouldn't trust the system because I feel like there's that human aspect of if you're actually going to human resources, you're going to want1 to deal with that human for that connection.

P2: Right, I feel like this also just ties into the debate of AI in the music industry as a whole. Just because I feel like people are inherently drawn toward listening to other people. And when you factor AI into that, that can sort of starve the industry of that which is why I agree with P1.

Interviewer: Do you feel less trustworthy of this system than the previous system? Or do you think it's about on the same level?

P2: I would say I trust this system less, primarily because the first situation targeted your approach as a producer to finding samples, and how efficient that could be. Whereas this is a little different approach with a different problem.

P1: I agree. But I feel like this is different from the talk of different music and producers and things because in line with this you're dealing with actual people and their salaries and their livelihoods and it's not going to be just if new music is getting produced, it's going to be if they're able to feed their families and get a job or not.

P2: Right, I don't think anyone would be happy being laid-off because of an AI.

Interviewer: And then so I'm once again going to give you four different things. I'm going to ask you to rank not how they affect fairness concerns necessarily, but how they affect how much you would trust the implementation of this system to make good/correct decisions, or to make decisions on par with a human.

(1) An external company conducts a bias audit on the system and receives a good score. They identified a few issues, but say that they're on par with what human decision makers do.
(2) The system has been implemented at many of your peer companies and significantly speeds up the hiring process. Executives say that widespread adoption makes the system more reliable.
(3) The system has been expanded to give more transparency to applicants. If they're rejected by the automated review system it will identify potential weak spots and give candidates a chance to explain them.

(4) In order to make sure the system is performing well, your company enacts a rule that at least one third of the decisions must be randomly reviewed by staff members. You're disappointed at the additional work it will mean, although you understand the reasoning.

P1: I'd personally go with the third option (3) for the best out of all them, because I feel like giving that review to the person is gonna be helpful. Right after that the fourth option (4) could be a good second, because the the ability for a human to still research all of them is amazing, because in that aspect you're still getting, you're still getting human. It's just being done streamlined. And the first 2 options (1) (2), I feel like they're interchangeable for last place.

P2: I agree with P1—the third option (3) was most appealing because it streamlines the process while maintaining a human aspect. The others are pretty interchangeable.

Interviewer: Why would you say they're interchangeable?

P1:I feel like they're interchangeable because they both kind of still remove that human aspect, and I feel like they…can you repeat one and two again.

Interviewer: (1) and (2) repeated.

P1: Yeah, I'll still keep my order because I feel like they're outside audits. And they're just people reviewing the system and saying that they work. I feel like you still need to keep that human aspect in the HR realm, because it is human resources, and you should be going to a human when you do something like that. I feel like, even though it streamlines it—I trust the system less than an actual human making decisions.

Interviewer: There's not a difference—it doesn't affect how much you trust it, whether you know that it's less biased? Or you know that it has more training data and is more accurate? Like those are the same in your eyes?

P2: I would also agree. But I also feel like when working with AI, you also have to consider the percentage of potential anomalies. And so that's definitely something to be considered.

Interviewer: Can you elaborate on what you mean by that?

P2: No.

[Transcription Paused]

Interviewer: Okay, I think that we're good to go ahead and move on to our third situation. So we're switching domains here.

[Transcription Paused]

Interviewer: Situation three was described during a network issue.

P1: I personally wouldn’t trust the system because in typical fashion with medical practices you cannot always have a significant first diagnosis. It can always switch out. So I feel like immediately getting treatment after a first diagnosis might be a bad idea. And along with that it could take away jobs from residents who are trying to make their way in the medical field. Yeah, I wouldn't trust it. I wouldn't trust it.

Interviewer: Do you think that's because of the stakes? Or because doctors make mistakes?

P1: I think it's both ways. Doctors make mistakes, but they know how to account for it because of their education. They have these backgrounds and there are specific things that weed the people out who are unable to. So I feel like doctors are more than capable of doing everything on their own with the help of residents and nurses.

P2: Right, going back to the possibility and factors regarding biomarkers and whether that is a conversation for the patient doctor relationship.

Interviewer: You think there's a special relationship that entails needing to have a human?

P2: Not particularly. I think it's more to do with the fact that the tech would be tracking the patients using biomarkers listed and that's a privacy concern.

Interviewer: But does that influence how much you trust it?

P2: To an extent, yeah. Sort of extent.

Interviewer: Okay, so one again I'm going to give you four things and ask you to rank how much they make you trust the system.

(1) Your hospital tells you that the engineers who built the robot updated its system. Now, it shows you its confidence in its predictions, as well as the top five predictions it generates. If confidence is too low the system will tell you it is uncertain and allow you to make a decision.

(2) A new software update generates explanations for the recommendations of the system—it will explain its medication and dosage decisions using information from drug repositories and medical textbooks. Beyond reading them, you cannot interact with these explanations.
(3) The hospital has added a face and voice to the robot—now patients and doctors are able to interact with the robot and ask it questions. The feature is still in beta.
(4) The hospital has streamlined the process for overriding decisions. Now, any human has the capability to stop the robot and all hospital staff members have the ability to override medication dosage decisions.

P1: Was option one the ability to read it before it was administered, or what was it?

Interviewer: It shows you its confidence, and how confident it is, and the top 5 predictions, and then you can choose which one of those to move forward.

P1: But it won't automatically administer the top one?

Interviewer: Right.

P2: I would say that option four (4) would be first for me, for the ability to stop it at any point in the process due to a calculated error. (2) for me would be option one to have a level of confidence in the suggestion and the last two are relatively interchangeable.

P1: I would choose (1), the ability to just see recommendations. I think that's the best way to go about it in the medical field. I'd trust that more, because if my doctor had the recommendation, they're showing me that hey this is a robot telling me this, and I think it's a good idea, because I've seen it working in the

past then I think that's great because usually doctors get a second opinion anyways. The rest of them I really didn't like too much. The ability for anybody in the hospital to override it is unnecessary, because if you have a receptionist or somebody who just works there as maybe a nurse or not even a nurse practitioner, and they're overriding the ability for it to not give medical care there can be problems there, and I feel like the uneducated may have access to the software to be able to stop it while it's going to give healthcare needed things. The one with the robot face I don't like because you just have a robot being a doctor.

Interviewer: And the explanations of medication and dosage decisions?

P1: That one I didn't really like because I feel like more input is required—even though you could override it I think it'd be more useful to just see the top five recommendations rather than just one recommendation with an explanation. I feel like having the option to be able to see different things and then be able to rethink it, and talk to other doctors is better than just getting a recommendation and trying to act on it.

Interviewer: So you and P2 had different rankings for those items. So, I'm going to ask you to come to one single ranking that you think accurately represents your feelings. You can discuss as much as you want.

P1: I feel like my ranking is the best because it keeps human interaction involved in the process and I think you can trust humans more than you can trust robots.

P2: I feel like my main reasoning for choosing (4) as my number one was primarily due to the possibility of being able to cancel any type of medical care at any given point in time. I think the concern (P1) had was related to the level of education or experience of medical care which I feel like could be factored into

the training. But the ability to intervene could be crucial at any given point if the administered medication is failing.

P1: So you're saying that you feel that the ability for the AI to act on its own is ok as long as you can stop it?

P2: Yes.

P1: I feel that the ability to stop is okay. But I feel like there should still be that discussion between the doctors and that there should be more options provided, because in some scenarios you're not able to just be like, yes, you have this. I feel like you have to be able to rethink things. So if you have multiple options, you can rethink more easily.

Interviewer: So do we have a single ranking? Or do you guys still disagree on the importance of overriding vs. more medical decision making.

P2: I would say I would agree with P1's initial ranking—possibly with a bump to the importance of (4)

Interviewer: Of course, I think we can go ahead and move on to our last situation of today. Your hospital has implemented another AI-enabled system. Now an AI will help streamline the work of the hospital's ethics board by making decisions about end-of-life care and organ donation. The system will help allocate and decide who on the waiting list receives organ transplants as well as make end-of-life care decisions for patients. Do you trust the system's decisions? Why or why not?

P1: So the AI is deciding end-of-life care.

Interviewer: Where an ethics board would—like if someone lacked family. Also organ donation.

P1: I would disagree with that.

P2: I would disagree with it as well. I would feel like the point of an ethics board is human input.

[Transcription Paused]

P2: I feel like the whole point of having an ethics board is to have different opinions in the process. I feel like they're leaving it to a system that doesn't take into account the intensity of making a decision at any given point.

P1: I agree because ethics board's are actual people who are going to look at the realistic side of things and not just go with utilitarianism, or any of those kinds of ethics frameworks to decide whether it works or not?

P2: Right, exactly, and especially like talking about ethics whenever you're considering the difference between communitarian and utilitarian approaches that's something an actual ethics board would excel at.

Interviewer: Right, I'm going to ask you to go ahead and rank four things.

(1) The engineers provide a manual about the information and decisions that went into designing the system. You're still not entirely sure how decisions are made—but you do know some key information like that it considers age, a person's health, etc. You have a chance to provide feedback on what you think should/shouldn't be considered.
(2) The hospital requires that the system be able to generate an explanation for the decisions it makes. The explanations make sense, although you disagree with them in several cases.

(3) Hospital administrators tell you that the system has over 95% agreement with the hospital ethics board's former decisions. They say it was trained on extensive data from hospitals around the world and thus minimizes human bias, especially a Western bias evident in previous decisionmaking.

(4) The hospital has decided to change their board structure. They made the ethics board smaller and, now, the AI system has an equal vote with the other members of the board. Thus, it represents only one part of the ultimate decision making.

P1: I feel like (4) (3) (2) (1).

P2: We both mutually agree on that.

P1: Because it keeps the human aspect more and I don't trust an AI to make those kinds of decisions.

[Transcription Paused]

Interviewer: Okay, so we just resumed, do you need me to remind you what the things were? I'm just gonna ask you to expand a little more about each. You mentioned that there's like a scale for how much humans are involved. I'm just gonna ask you to explain a little bit more.

P1: I think with each step there is more human involvement, because in the first step you're just giving it a voice on the board so it can still be outnumbered. (3) was…what was it?

Interviewer: It was that it has 95% agreement with former decisions and minimizes human bias, especially Western bias that's been evident in previous decision making.

P1: I feel like that could always be changed as ethics become different. I feel like it's the best way to go about it. Maybe changing the ethics board and not creating A that's going to take over.

Interviewer: So did you want to change your order?

P1: No, I'm still keeping the same order,.What was the (2) again?

Interviewer: That it would generate an explanation of the decisions it makes that is coherent and makes sense and understandable.

P1: But it already executed the decision?

Interviewer: Yeah, it makes the decision. It just provides that explanation.

P2: If it's already made the decision then the justification doesn't really have an effect—at least it's verified a little more.

P1: Now, what was (1)? Yeah, I can't remember.

Interviewer: It was that the engineers tell you about the key information it considers and you have a chance to provide feedback on stuff you think should/shouldn't be considered.

P1: There's always other options, the AI is still making the decisions. That's not a good enough reason.

P2: I agree.

Interviewer: Sounds good. Thank you so much for your time and for agreeing to participate today.

Appendix C

Original Annotated Bibliography

Scholarly Resource (Citation)	Summary (short)	Connection to Your Research	Questions / Insights / Comments		Link to article	Pub date
[1] E. Schwarz, “Silicon Valley Goes to War: Artificial Intelligence, Weapons Systems, and the De-Skilled Moral Agent,” Philosophy Today, vol. 65, no. 3, pp. 549–569, 2021. doi:10.5840/philtoday2021519407	Discusses the ethical implications of human-AI teaming in war contexts—particularly the interactions of human/machine overrides. Also touches on the risks of "moral deskilling" that come from the increasing automation of ethical decision making and machine-generated worldviews.	Highlights a compelling problem area and the need for more research on trust + explainability on AI—particularly automated decision making in morally high-stakes situations.	I'm really interested in whether more research has explored risks of "moral de-skilling" and the dynamics of how to mitigate this, etc. as I think this work may produce interesting results in this area.		here	Summer 2021
[2] N. Ezer et al., “Trust Engineering for Human-AI Teams,” Proceedings of the Human Factors and Ergonomics	Discusses high-level approaches to "trust engineering" in human-AI teaming to foster bi-directional	My research project is, broadly-speaking, interested in understanding the trust relationships of humans and	This is a fairly high-level conference proceeding but it includes a lot of interesting references and provides a		here	Nov. 2019

Society Annual Meeting, vol. 63, no. 1, pp. 322–326, Nov. 2019. doi:10.1177/1071181319631264	trust. Increasingly, humans and AI agents are (or are imagined to be) working in tandem. A key component of the effectiveness of these teams is trust which can be created in a variety of ways—algorithmic transparency, adaptive explainability, etc.	AI in morally-charged contexts. Work in "trust engineering" offers a valuable corpus of work to draw on for developing questions/scenarios which effectively interrogate these trust relationships.	good balance of basic/foundational work in the field and newer work which feels particularly relevant to my project. I'm particularly interested in "Trust Engineering Code of Ethics" and "Algorithmic Transparency".			
[3] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, “Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI,” FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635, Mar. 2021. doi:10.1145/3442188.3445923	Offers conceptual formalizations of trust including evaluating degrees of trust. Discusses the interactions between explainable AI and this proposed model of trust as they apply do designing trustworthy AI.	It looks like this paper will offer a strong formulation of trustworthiness to help guide research directions and evaluate and categorize participant responses.	This paper specifically mentions the concepts of warranted vs. unwarranted trust which is linked to the underlying capabilities of an AI agent. One thing I'm particularly interested in exploring is how (and how much) AI agents should reveal about their capabilities.		here	March 2021
[4] E. Glikson and A. W. Woolley, “Human Trust in Artificial Intelligence: Review of Empirical Research,” Academy of Management Annals, vol. 14, no. 2, pp. 627–660, Aug. 2020.	Explores the role AI anthropomorphism plays in emotional trust from a business/organizational psychology lens. Pays particular focus to how	Beyond things like algorithmic transparency, the physical manifestations of AI agents and the ways in which explanations/transparency are presented play a major role in	Search more on keywords related to this work.		here	Aug. 2020

doi:10.5465/annals.2018.0057	presentational features of human-AI communication play an important role in mediating trust relationships.	influencing trust. Drawing on this work, I hope this project will broaden research literature on how these non-intrinsic agent features influence trust relationships created through explainable AI methods grounded in mechanistic interpretability.				
Editorial / Pop Resource (Citation)	**Summary (short)**	**Connection to Your Research**	**Questions / Insights / Comments**		**Link to article**	**Pub date**
[5] M. Bailey, “How Can We Trust AI If We Don’t Know How It Works,” Scientific American, Oct. 03, 2023. https://www.scientificamerican.com/article/how-can-we-trust-ai-if-we-dont-know-how-it-works/ (accessed Nov. 09, 2023).	Identifies predictability and social norms as core features of trust and argues that current AI systems lack either.	Highlights the need for the development of trustworthy AI and two potentially important directions to explore (predictability and social norms). Also offers criticism of current human in-the-loop or on-the-loop systems as unsustainable in the long-term—raising questions about the granularity of trust and the importance of trust in subsystems.	The article previews questions about the granularity of trust and how important subsystem trust is—I find these especially interesting and would love to explore them.		here	Oct. 2023

[6] N. S. Qureshi, “Waluigi, Carl Jung, and the Case for Moral AI,” Wired, May 25, 2023. https://www.wired.com/story/waluigi-effect-generative-artificial-intelligence-morality/ (accessed Nov. 09, 2023).	Highlights how AI can be manipulated (intentionally or unintentionally) to create antisocial/amoral outcomes.	Most work prioritizes absolute human control over AI systems. However, as this reporting highlights, there are compelling reasons to think that AI systems should mistrust humans and withhold information and/or motivations for decision making. These scenarios highlight one of the key tensions motivating this project.	It would be interesting to take a game-theoretic approach to looking at human-agent interactions. Has potential as a theoretical framework to conceptualize responses.		here	May 2023
[7] A. Tyson, G. Pasquini, A. Spencer, and C. Funk, “60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care,” Pew Research Center, Feb. 22, 2023. https://www.pewresearch.org/science/2023/02/22/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/ (accessed Nov. 09, 2023).	Explores results of a Pew Research poll about AI in healthcare—particularly interesting results is that there is widespread hesitation around AI adoption in healthcare although most people believe it would reduce mistakes and bias in healthcare.	Trust in/of AI systems is incredibly nuanced. Even if people believe that AI agents are capable, fair, etc. they may not trust these agents in their care contexts. Centrally, this speaks to the importance of techniques for explainability. Simply offering transparency may not be enough to create trust—it must be presented in specific ways.	Something to consider is whether questions should focus on the participants themselves or ask them to imagine ethical situations/"be in someone else's shoes".		here	Feb. 2023