When Trust is Zero Sum: Automation's Threat to Epistemic Agency

Emmie Malone[1], Saleh Afroogh*[2], Jason D'Cruz[3], Kush R. Varshney[4]

1. Lone Star College, emmie.Siobhan.Malone@lonestar.edu
2. The University of Texas at Austin, saleh.afroogh@utexas.edu
3. The State University of New York at Albany, jdcruz@albany.edu
4. IBM Research, Thomas J. Watson Research Center, krvarshn@us.ibm.com

* Corresponding author: saleh.afroogh@utexas.edu

Abstract

AI researchers and ethicists have long worried about the threat that automation poses to human dignity, autonomy, and to the sense of personal value that is tied to work. Typically, proposed solutions to this problem focus on ways in which we can reduce the number of job losses which result from automation, ways to retrain those that lose their jobs, or ways to mitigate the social consequences of those job losses. However, even in cases where workers keep their jobs, their agency within those roles might be severely downgraded. For instance, human employees might work alongside artificial intelligence (AI) but not be allowed to make decisions or not be allowed to make decisions without consulting with or coming to agreement with the AI. Here, we argue that this is a kind of epistemic harm (which could be an injustice if it is distributed on the basis of identity prejudice). It diminishes human agency (in constraining people's ability to act independently), and it fails to recognize the workers' epistemic agency as qualified experts. Workers, in this case, aren't given the trust they are entitled to. This means that issues of human dignity remain even in cases where everyone keeps their job. Further, job retention focused solutions, such as designing an algorithm to work alongside the human employee, may only enable these harms. In response to this, we propose an alternative design solution, adversarial collaboration, which addresses the traditional retention problem of automation, but also addresses the larger underlying problem of epistemic harms and the distribution of trust between AI and humans in the workplace.

1. Introduction

Ethical discussions surrounding the introduction of artificial intelligence (AI) into the workplace are often dominated by concerns about the various costs and benefits of automation. For instance, there has already been considerable attention paid to the potential harms to human dignity that go along with humans competing with, and ultimately being replaced by, AI systems (Stefano, 2019). Humans derive a sense of meaning from their work, and from doing it well. Likewise, they feel invested in social cohesion when they are both allowed to contribute to social progress and those contributions are recognized as valuable by that society. If AI systems enter the workplace and compete for jobs with human workers, then those workers risk losing this sense of meaning and social buy-in. The possibilities for these technologies, like large-language models, to replace individuals' jobs through the automation of tasks which formerly required a person's labor are enormous. In turn, the potential for this general loss of meaningful employment to lead to a corresponding loss of workers' sense of identity and meaning is also correspondingly high. In addition, as AI replaces human roles, this disruption may lead to a loss of social cohesion and progress.

In addition, employment brings with it material consequences for individuals and communities that lose out on income as a result of automation. The displaced worker, their family and dependents, and their community suffer from a loss of income, potentially a loss of health insurance coverage, and downstream disinvestment in community resources (when automation affects a particular geographic region). These material consequences have the potential to result in serious secondary social harms as well. For instance, lack of access to economic opportunities may contribute to distrust of social institutions, resentment, and general social unrest (Acemoglu, 2021). These problems can be compounded when patterns of automation disproportionately affect already marginalized communities (Petersen et al., 2022). If the jobs that are automated most (or most readily) tend to be those primarily occupied by women, people of color, or those of a lower economic status, then the harms of automation are distributed in inequitable ways. In this way, automation through AI poses potential social and economic problems, but those problems can also be distributed in ways that raise concerns about social justice.

Concerns over equity and automation are not new, and ethicists and AI researchers have devoted significant attention to understanding and addressing these problems. However, ethical issues surrounding AI in the workplace are not exhausted by concerns about employment. That is, even if workers manage to keep their jobs, certain applications of AI alongside those workers still risk causing them significant harm and still risk perpetuating identity-prejudicial injustices. By thinking of these problems solely in terms of automation and employment, researchers risk ignoring the problems which remain even when continued employment for human workers is guaranteed, and risk endorsing design approaches which enable continued harms.

We maintain that concerns about employment do not exhaust the ethical concerns raised by AI in the workplace, because there are additional potential 'epistemic' harms which workers can and do face as a result of the adoption of AI. These epistemic harms concern workers' sense of dignity and meaning even when they are able to retain their jobs. A focus on job retention suggests a high-level design approach focused on human-AI collaboration. However, straightforward collaboration between humans and AI still runs the risk of perpetuating these epistemic harms. Instead, we advocate that we ought to also think about ethical issues surrounding AI in the workplace in terms of the distribution of trust (between humans and AI). Here, the distributive schema of the workplace trust economy can be more or less equitable in the same way that employment can.[1] We propose *human-AI adversarial collaboration* as a design approach to mitigate the harms of automation (even in, but not limited to, cases where workers retain their jobs) and to recognize the epistemic agency of workers. Adversarial collaboration aims to address these concerns by avoiding the circumstances in which human-AI conflict can arise within the workplace at the design-level instead of or in addition to at the organizational level.

In section one, we describe the problem of automation in terms of a distribution of trust and expertise within the workplace. In section two, we argue that epistemic harms can persist for workers even when they retain their employment and describe various scenarios in which biases in the distribution of trust within the workplace can lead to unjust and inequitable distributions of epistemic harms. Importantly, we argue that straight-forward design approaches centered on human-AI collaboration will fail to address (and could enable) those epistemic harms. Finally, in section three, we advocate for an adversarial collaboration design approach, which better recognizes the epistemic agency of workers. We describe a few ways in which this design approach could be realized and explain how they avoid the harms already discussed.

2. Automation & Trust

Historically, technological innovations have played an instrumental role in human development and were considered tools in assisting humans, as agents and decision-makers, in achieving their goals (Afroogh et al., 2021). However, AI, with its current rapid growth, has the potential to invert the historical tool-agent relation between technological production and human beings. AI systems may, in some contexts, use humans as tools to assist them in attaining their goals. For example, trainees in laparoscopic surgery have, in some cases, had their roles limited to simply 'docking' the surgical AI-powered robot and observing as it performs the entire procedure, a marked change from how trainees once contributed more robustly to these procedures (Beane, 2022). This effect might even occur in cases where a field faces automation generally or merely might, and a particular worker's position is not yet directly affected. For instance, human beings

[1]. The concept of an economy of trust refers to the distribution of trust between people and AI systems at work, describing how they share in trusting each other.

might think that they do not need to perform or further develop their capacity to perform tasks which will ultimately be better performed and developed by more advanced AI algorithms. Widespread public fatalism about the inevitability of machines to outperform humans at most tasks risks dealing a long-term blow to human creativity and ingenuity. In this way, advances in technology have and continue to pose, at a minimum, the public perception of a threat to the ultimate futility of our own human self-cultivation. Importantly, these harms will occur whether or not these worries are justified as long as the belief in AI's epistemic superiority is widespread. While these issues, when it comes to full automation a given job, have drawn considerable attention among the public and researchers, it is important to note the ways in which issues around automation are tied with issues of expertise, agency, and trust. By doing this, we can begin to identify ethical issues and questions of justice which are overlooked by focusing on complete automation of a professional role.

The central problem raised by automation is that employment in a particular job is often a rivalrous good.[2] At the smallest scale, you have a human worker and an AI competing for the same job. Within this competition, if an AI is capable of outcompeting the human worker on the basis of either performance or cost-savings, and since the contest for employment in this particular role is zero-sum, the worker risks losing out on all of the direct and indirect benefits that come with employment. On the basis of this description, we might adopt a few strategies to ameliorate the effects of automation. One solution might be to mitigate consequences for the worker by expanding the social safety net or by offering training in some other economic sector. An alternative, but not mutually exclusive, strategy would be to design the AI with the intent that the worker will be employed alongside the AI and collaborate with it, and thereby avoid the problem. However, and importantly, this solution will not fully address the harms associated with automation. To make clear why this is the case, it might be helpful to think about the problem of automation in broader terms.

Trust, as a disposition towards others which attributes competence and truthfulness, and which promotes reliance, is a rivalrous good. In addition to being an element of individual capital, trust is also an element of social capital (Claridge, 2020). Considered as individual capital, being trusted (especially on matters concerning one's expertise) is an important good. A systemic failure to be trusted when one is trustworthy is damaging to one's sense of dignity and achieving the trust of others allows for other goods.[3] For instance, trust is a fundamental aspect of all communication, whether it be human-to-human or AI-to-human. There is a remarkable value and function to trust in human communities, organizations, and society. Distinguished as a type of social capital, trust can also assist local leaders, administrators, and governments in reaching their goals by enhancing the efficiency of local communities, organizations, and societies (Afroogh, 2022). However,

[2] Of course, there are also often situations in which employment is non-rivalrous between humans and AI systems. That is just to say that all instances of automation do not raise 'the problem of automation'. Here, we limit ourselves to problematic cases, as we aim to offer a solution to the problem where it arises or could arise.

[3] For further reading on the personal and social costs of a failure to recognize one's expertise, see (Fricker 2007), which is discussed more fully later in the paper.

beyond the value of appropriately placed trust for the person being trusted, as a kind of social capital, proper trust is valuable for the person or institution who trusts. When relevant expertise is ignored, those in a position to have trusted the experts suffer a significant epistemic loss; they are deprived of valuable knowledge they would otherwise have had.[4]

Yet, trust can also be a scarce good and cannot be assigned to or recognized equally for all individuals or agents in a society. In these cases, there is a direct correlation between the scarcity of trust and its value. Individuals are in competition to be trusted. In some contexts, even if there are several trustworthy agents, not all will be credited with being the trustee. As an example, imagine a case in which two experts are consulted to offer solutions to the same problem. Assuming they are both paid a fee for their consultation, employment and its material consequences are non-rivalrous between them. However, if they offer two competing solutions, only one will ultimately be relied on. If the same two experts are pitted against one another repeatedly in various cases, with the same expert failing to be relied on every time, they might reasonably conclude that their expertise is not trusted.

Ultimately, in cases like this, the consulting party has to adopt one recommendation or the other where those recommendations are mutually exclusive. Insofar as experts are interested in having their recommendations adopted by the user and adoptions are rivalrous, various schemes of adoption open themselves up to being more or less just. Situations such as these are what philosophers like John Rawls call 'the circumstances of justice' (Rawls 1971). That is, these are the circumstances in which issues of distributive justice arise, and we are, accordingly, compelled to ask how a scare resource is to be distributed (in this case, trust). Granted, neither trust nor reliance are *always* rivalrous. There are certainly cases where one trusts several parties with respect to the same decision. Furthermore, there are cases in which one can take into account conflicting advice by synthesizing it into complex plan of action. (For example, you may decide to take risk while still hedging your bets). Nonetheless, there will be contexts whereby opting to follow the recommendation of one party, you thereby reject the recommendation of the other. In many such cases, the party whose advice is jettisoned may suffer a kind of harm which comes with the sense that one's expertise is not adequately recognized or trusted.

Importantly, this same situation may arise between a human expert and AI system. Suppose that a medical doctor prescribes an approach that is contrary to and incompatible with an AI system's recommendations. If this were to play out repeatedly, as in the case above, the human doctor might reasonably conclude that their expertise is not trusted. In these cases, relying on one party means not relying on the other, and reliance is a significant way to express and reestablish trust. As such, workers placed into reliance competitions of this kind against AI systems face the same kinds of harms as above, where their expertise might fail to be recognized, thus damaging

[4] The authors thank an anonymous reviewer for raising this point.

their sense of self, but also the overall epistemic structure of the healthcare systems they are involved in.

On this description, the problem raised automation is also one of how trust is distributed. Since adopting a particular recommendation in cases like this involves trusting one party over another, the question of whether a worker will retain their influence with the end user, and thus their employment, comes down to whether their expertise is recognized by that user. However, notice that even when the human worker retains their employment, they still may end up with a smaller share of trust than they started with as a result of the introduction of AI into the workplace. For instance, if employers keep their human workforce but supplement it with AI, then those AI-systems may take on a decision-making or decision-verifying role. In this case, a worker may continue to do their job, but the role that that job picks out may become increasingly circumscribed. If AI performs at a high level, then employers may demand that workers consult with, defer to, or come to agreement with the AI before acting. We argue that subordination to this circumscribed role still constitutes a harm to the worker, it is just that that harm is *epistemic*, rather than financial, in nature.

3. Epistemic Harm & Trust Equity

As we have said, we argue that the situation described above is an instance of what Miranda Fricker calls an 'epistemic harm' (Fricker, 2007). In this case, where workers are pushed into a subordinated and circumscribed role, involves an employer failing to recognize the epistemic agency of the worker. That is, as a relevant expert who is sufficiently trustworthy in and of themselves. Accordingly, an epistemic harm occurs when one does not recognize or respect the full extent of a person's epistemic agency. This model highlights the way in which the meaning and dignity that humans derive from their work isn't reducible to being able to say that they have a job or that they do work, nor to obtaining the material benefits that come with gainful employment. Rather, much of the meaning and dignity that workers feel is the result of the exercise of their agency. If workers are reduced to serving at the discretion of the algorithm, then they are liable to experience alienation even if they avoid the economic consequences of a loss of income. Recognizing that worker's epistemic agency means trusting them as a relevant and trustworthy expert to make decisions and be held accountable for those choices. When outcomes issue from the exercise of this agency, that worker can reasonably be accountable for their success and the benefits that result from it. This gives workers a sense of ownership and self-respect with regards to their work. This is what enables a sense of meaning and dignity, values that are not necessarily realized by guaranteeing continued employment.

In addition to the epistemic harms that may face individual workers, the distribution of these epistemic harms may result in more or less equitable patterns of trust. When one fails to recognize the complete epistemic agency of another on the basis of some identity prejudice, we

can think of this as an 'epistemic injustice' (Fricker, 2007). For instance, if we assign less credence to the testimonies of women than men on the basis of an (explicit or implicit) gender prejudice, then an epistemic injustice has occurred. So-called 'testimonial injustices' of this kind can be thought of as unjust distributions of trust in cases where those testimonies are mutually exclusive because we distribute that rivalrous good according to what we, upon reflection, deem to be an unjust bias.

This kind of unjust distribution can happen in terms of employment. For instance, consider a case in which two patients decide whether to consult with one of 1) a male human doctor, 2) a female human doctor, or 3) an AI system. In a situation where all three options are deemed sufficiently trustworthy, the decision of who each patient will trust might be rationally underdetermined all things considered. This leaves room for bias to be the difference maker in which of the three doctors is left with no patient. It has been shown that patients rate female doctors less favorably than male doctors (Wallace & Paul, 2016; Kauff et al., 2021). This means that an AI system could be trusted more than female doctors but less than male doctors. If this is reflected in patients' adoption rates, then that AI system will automate the roles of female doctors but not male doctors. While this situation already raises concerns when the competition is only between male doctors and female doctors facing an identity bias, the proliferation of AI systems which can compete in this marketplace of expertise only exacerbate the problem, as it introduces a number of additional competitors (AI systems) which are not subject to the same identity bias as the female doctors. Our point here is that we ought to be sensitive to this amplification of inequity and recognize that these epistemic harms are not necessarily distributed evenly or fairly across a population of experts. In cases where all three are trustworthy and patients trust male human doctors and AI systems to have the relevant medical knowledge (enough to endorse their recommendations), but not female human doctors, they fail to recognize and respect the female doctors' epistemic agency for reasons of identity prejudice.

Beyond potential inequalities within a particular professional role, inequalities of epistemic harm may also arise across varying professional roles. In these cases, the identity prejudice at play may conceal itself as a downstream consequence of the automation of certain roles and not others. For instance, if automation disproportionately affects industries dominated by marginalized groups, then inequitable distributions will arise within a society even when AI systems aren't outcompeting individuals at a biased rate within that domain. In this case, we might see differential rates of automation between male-dominated roles like medical doctors and female-dominated roles like nurses. The same patterns may result on the basis of class or race as well. Just as the male-dominated role of doctor might be automated at a different rate than the female-dominated role of nurse, individual discretion within these roles might be given over to AI at different rates. If doctors are allowed to use their own independent discretion when recommending diagnoses and treatment where nurses must defer to a consensus between themselves and an AI system, then their epistemic agency is diminished in ways that we might think represent an unequitable distribution of epistemic agency. This is an important additional harm that can occur even in cases where

workers retain their employment and income, and it represents an equity issue which concerns the dignity of workers as epistemic agents. Again, these same biases, but concerning race or class, might apply across or within industries. Importantly, the differential automation of roles across gender lines is not merely a hypothetical concern, while women make up only 47% of the workforce, some analysis suggests that they constitute 58% of the workers at highest risk of having their roles fully automated (IWPR 2019). Likewise, these differences are, at least in part, a result of gendered differences in occupation, as women outnumber men 10 to 7 in those jobs deemed 'high risk' for automation (IWPR 2019).

Importantly, the situations we are describing occur in cases in which humans retain their jobs by collaborating with AI systems. Yet, these partnerships could easily become hierarchical (with the human on the bottom) if the AI outperforms the human or if it is merely perceived that it does. As such, any solution to the problems of automation which focuses on designing the AI system to perform the same task as the human worker, even if this is initially intended to be alongside the human so that they can retain their employment, runs the risk of falling into a hierarchical arrangement in which the human's role is highly circumscribed, reliance is distributed to the AI, and the human worker's agency and expertise fail to be recognized. This seems to raise the question of what an alternative approach to AI design might look like, which might avoid the possibility of this problem.

4. Adversarial Collaboration

While there may be several strategies for addressing the problem described above, we argue that an optimal solution would be one in which all parties avoid falling into the circumstances of justice altogether. Our contention is that one way of doing this is by designing algorithms to engage in what is called 'adversarial collaboration' with fellow trustworthy recommenders (Kahneman 2003). A version of this approach has been developed and employed to coordinate otherwise competing teams of human researchers in a scientific context, and a similar model may work equally well between human experts and AI systems. On a standard scientific model, scientists often adopt an approach where one team of researchers designs an experiment in order to disprove another team's explanation of the data and the second team responds by designing and running a responding experiment designed to disprove the first team's explanation of the data. By contrast, adversarial collaboration in the human context involves the two teams working together to design an experiment which both parties see as moving the debate forward. However, this is not mere collaboration, in which colleagues with a common point of view work together to achieve their shared ends. Instead, this approach views the adversarial stance between the two parties as essential in generating progress. We propose a similar, but distinct, model to avoid the circumstances of justice in AI-human collaboration.

As discussed above, one model for designing AI with human collaboration in mind might be to aim for the algorithm to perform the same task as the human agent. For instance, an algorithm may look at X-ray scans and suggest the presence of a tumor or not on the basis. This suggestion could be offered to the human doctor who makes their own recommendation, or the human doctor might merely review the recommendations made by the AI for error. This model constitutes designing for ordinary collaboration. However, since the algorithm and human perform the same task, this opens up the possibility of eventual complete automation of the human doctor's role, or subordination of the human doctor in line with the cases of epistemic harm described above. While the workplace structure may place the human doctor in the position of epistemic authority (by giving them the final say in making a diagnosis), the design of the algorithm itself does not. Insofar as these algorithms are designed to perform the same task as the human, the actual agency the human doctor has in exercising their power to make the final decision can be curtailed or limited on the basis that the algorithm could have outperformed them at the same task. At this point, their final say is merely perfunctory.

Our proposal is that the algorithm should in some cases, not be designed to perform the same task as the doctor. Instead of establishing a straightforward partnership between the AI and the human, we could design the algorithm to occupy an adversarial relationship within their collaboration. The task of the AI, on this model, is not to offer an alternative recommendation, but to scrutinize the basis for the human's decision. This might be to identify counterevidence, to offer alternative explanations of evidence in favor of the human's decision, or to mount the best defense of the alternative position. All of these approaches would constitute models of adversarial human-AI collaboration.[5] The human agent could then engage in an iterative process with the AI for a number of cycles before arriving at a final decision.[6]

Our proposal resembles Kahneman's model in some ways and departs in others. As the name suggests, they share in common that the parties are aligned in their goal (arriving at an optimal decision) but are oriented toward critically scrutinizing each other's views. However, Kahneman models human-human collaboration in science and our proposal is designed to model human-AI collaboration. This introduces an important distinction. Human-human collaboration involves multiple agents, each with their own epistemic agency deserving of recognition. This puts both parties in an equal position of epistemic authority. Our model involves human-AI collaboration and is designed specifically to protect the epistemic agency of the human agent. This problem is solved by creating an asymmetry of epistemic authority which places the human in the

[5] We recognize that not all situations are amenable to this kind of an iterative process. For instance, emergency room procedures might require humans to act quickly for the benefit of a patient. Accordingly, human-AI adversarial collaboration of kind we are describing might be best suited for contexts in which the refinement of expert opinion is both needed and possible. This point is discussed further in the penultimate section of the paper. We thank an anonymous reviewer for their suggestion of this example.

[6]. An alternative approach to adversarially collaborative design might have human agents play the antagonistic role with AI systems are the lead recommender. This model has already been proposed for reasons unrelated to trust equity (Attenberg et al., 2015).

primary position, and in a way that recognizes their expertise and places trust in them accordingly by resisting the kinds of direct competition which engender their position becoming circumscribed. The human, in this model, places the central role in generating a recommendation, is the one ultimately making the decision, and the algorithm isn't designed to compete in this role, but to interrogate and, thus, sharpen the recommendations of the human agent.

In a model of adversarial collaboration between human and AI recommenders, both parties work together to achieve the user's adoption of their joint recommendation. By contrast, less adversarial forms of collaboration between humans and AI (in which, for example, an AI might offer a doctor a recommendation that that doctor could take or leave), we run the risk of making the doctor redundant should the AI outpace the doctor's performance. That is, there is no safeguard against collaborative AI and their human users entering into the circumstances of justice. Meanwhile, the adversarially collaborative AI's contribution to the recommendation is necessarily subordinate to the human who makes both the initial and final recommendation to the end user. Because this model avoids the situation where AI systems outcompete some or all workers for jobs as expert recommenders, it avoids the potential for total automation and the potential harms that go with it (in terms of human dignity and/or income). Likewise, because the human agent still plays an essential (and primary) role in providing end users with recommendations, adversarially collaborative design works to recognize the epistemic agency of expert workers and the epistemic injustices associated with a failure to do so. That is, if these epistemic harms cannot arise, then epistemic injustices relating to how these harms are distributed cannot either.

Additionally, recognizing workers' epistemic agency does not need to involve blind trust. Real expertise does not entail, in and of itself, that we always defer. Instead, it merely entails that we recognize that expertise and take it seriously. Indeed, the algorithm taking the human's reasons and recommendations seriously, and providing some scrutiny, is, in a way, a recognition of their epistemic agency, not just an instance where a failure to recognize that expertise is avoided. In this way, AI can contribute to human problem-solving. The moral demands of epistemic agency merely require that human be challenged by alternative evidence and explanations (which can be provided by AI), to incorporate that feedback, and take responsibility for their own resulting conclusion.

Importantly, such a model (but involving only human collaborators) has already been shown to be a productive method for addressing persistent weaknesses in social scientific norms, and similar but distinct strategies for AI-human collaboration have shown some success in medical imagining (Clark, 2023; Leibig et al., 2022). In this case, the AI system developed according to an adversarially collaborative design approach plays the role of a 'devil's advocate'. Beyond avoiding potentially harms to experts, to the extent that this model is productive, this increase in performance will generate better outcomes and these better outcomes will have knock-on effects in terms of social acceptance of AI systems. This is to say that the adversarial collaboration model removes expert workers from a zero-sum game of competition with an AI system and places them into a possible Pareto efficiency, raising the trustworthiness of both the AI and human recommender, while better serving the end user. The enhanced trustworthiness that comes with

better risk management strategies could lead to wider user adoption and enhanced public opinion of AI solutions, a problem which already attracts significant attention (Akkara & Kuriakose, 2020; Chen & Wen, 2021; Jackson & Panteli, 2021; Spiegelhalter, 2020; Tschopp, 2019). This is to say that adversarial collaboration, as a high-level design approach may address concerns about automation and epistemic harms to workers (and thus avoid potentially unequitable distributions of those harms), but also work to counter the problems of adoption that go with diminished trust in AI. Likewise, to the extent that the model we present here is human-centered and productive, end users will be able to point to a specific human agent who is empowered to take accountability, and to reliable results, as a basis for placing trust in the recommendations of a competent and trustworthy AI. This could lead to greater public adoption and trust in AI systems.

Finally, while epistemic harms to workers can be serious and weighty, they are not the only value that should be considered in most cases. While adversarial collaboration addresses the issue of epistemic harms, it may not be appropriate as a general design strategy in all circumstances, especially when epistemic harms might be outweighed by considerations of expediency. For instance, in high stakes decisions about health in emergency situations, what matters most is that the right answer is arrived at quickly. These circumstances might suggest a more straight-forward collaboration model because adversarial collaboration requires a timeline amenable to an iterative process. However, where decisions are relatively less time sensitive, epistemic harm may become a sufficiently salient moral feature when designing an algorithm for human-AI collaboration. Take, for example, a radio DJ who must determine which set list is most likely to get listeners dancing. A well-trained AI could reasonably outperform the human at this task in all cases, but the costs associated with a slightly suboptimal dance music playlist do not warrant the associated harms to worker agency. Likewise, adversarial human-AI collaboration may be appropriate in situations where human agents have special access to sources of information that an AI will not. For example, a human doctor or nurse will know things about a patient's values and preferences as well as their ability to follow through on treatment plans. While the AI is not making an 'error' in calculation in these cases, it comes to a worse judgment about how to proceed. Deference to the algorithm (as a result of automation or subordination of the human agent) in such a case is an insult to the expertise and agency of the healthcare worker and an injury to the patient as an end-user. This is just to say that both straight-forward and adversarial models may be optimific under different circumstances, and that a one-size-fits-all approach to collaborative design may not be appropriate.

In the same way, there are contexts in which epistemic harms are more salient than others. This is especially true with respect to workers who are of lower status, whether that is internally within the structure of a company, economically, or in terms of social esteem. The subordination of a nurse's expertise and agency in the workplace is not counteracted by significant social status as an expert the way that a surgeon's might be. In the same way, epistemic harms will matter more in cases where decisions have more bearing on workers' own work arrangements than on end users. For example, consider decisions about workplace shift scheduling or protocols that are determined by algorithm to maximize efficiency. The benefits of this gain in efficiency are often

secondary and minimal to the end user. However, the agency and autonomy that worker's exercise in organizing their own workplace and workflow are important in recognizing their epistemic authority. These factors should be considered in determining whether adversarial collaboration is the optimal design paradigm for a given situation and task.

5. Conclusion and Future Directions

We have identified a limitation of thinking about the problems of automation solely in the context of employment. Namely, that doing so will fail to capture the ways in which workers who retain their employment may still be harmed as epistemic agents when their discretion is diminished as they are required to defer decision-making authority to AI systems. Like the direct and indirect harms of automation, the harms of diminished epistemic agency in the workplace may be distributed in more or less equitable ways. By approaching the problem of automation from a higher level, in which the issue is cast as one of how trust is distributed within an ecosystem of workers, we can attend to the problems of employment, income, and epistemic agency simultaneously. This conception recommends adopting design approaches which center workers' epistemic agency, and are sensitive to the ways in which trust is distributed. Ultimately, we suggest that designing AI systems with the goal of adversarial collaboration in mind better serves this vision than merely designing those systems with simple non-adversarial collaboration in mind. This is because adversarial collaboration recognizes the epistemic agency of expert workers and builds not just a continued role, but continued agency, into the desired use plan for AI. Importantly, this approach avoids the pitfalls associated with a narrow "focus on application that automate jobs…" while still harnessing the power of AI to make human decision-making and industry more reliable, more trustworthy, and more efficient (Acemoglu, 2021). As a result, adversarial collaboration represents a design strategy for "AI for good" (Acemoglu, 2021).

While we have identified two areas of concern which can arise within the ecosystem of trust in which AI systems compete (employment and epistemic harms), more work should be done to identify further areas of concern and to address complications that emerge as a result of competition for users' trust. In addition, more work should be done to identify domains in which these harms are already occurring and where they might be occurring at inequitable rates on the basis of identity-prejudicial biases. Finally, adversarial collaboration may come with its own unique set of obligations to end users. We expect that there will be cases in which the decision maker has an obligation to divulge the counter-evidence or alternative explanations offered by the adversarial AI. There are likely a variety of factors which might make disclosures of this kind more pressing. For instance, certain medical decisions with high stakes for patient health outcomes may require more transparency about the nature of the adversarial collaboration which lead to a particular diagnosis or treatment plan where low-stakes decision making may not. Further work

should be done to identify the factors which might influence how transparency is achieved for end users within an adversarially collaborative framework.

Funding: This research is funded by the SUNY-IBM AI Research Alliance under grant number AI2102.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Acemoglu, Daron. (2021). *Redesigning AI: Work, Democracy, and Justice in the Age of Automation*. Cambridge: MIT Press.

Afroogh, S. (2022). A Probabilistic Theory of Trust Concerning Artificial Intelligence: Can Intelligent Robots Trust Humans? *AI and Ethics*, 2022. https://doi.org/10.1007/s43681-022-00174-4

Afroogh, S., Esmalian, A., Donaldson, J.P., Mostafavi, A. (2021). Empathic Design in Engineering Education and Practice: An Approach for Achieving Inclusive and Effective Community Resilience. *Sustainability*, 13(7). https://doi.org/10.3390/su13074060

Akkara, J. D., & Kuriakose, A. (2020). Commentary: Artificial Intelligence for Everything: Can We Trust It? In *Indian Journal of Ophthalmology* (Vol. 68, Issue 7, pp. 1346–1347). Wolters Kluwer Medknow Publications. https://doi.org/10.4103/ijo.IJO_216_20

Attenberg, J., Ipeirotis, P., Provost, F. (2015). Beat the Machine: Challenging Humans to Find a Predictive Model's "Unknown Unknowns". Journal of Data and Information Quality 6(1), 1-17. https://doi.org/10.1145/2700832

Beane, Matt. (2022). Today's Robotic Surgery Turns Surgical Trainees into Spectators. *IEEE Spectrum*. https://spectrum.ieee.org/robotic-surgery#toggle-gdpr.

Bohannon, J. (2015). Fears of an AI Pioneer. *Science* 349(6245), 252. https://www.science.org/doi/full/10.1126/science.349.6245.252?cookieSet=1

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chen, Y. N. K., & Wen, C. H. R. (2021). Impacts of Attitudes Toward Government and Corporations on Public Trust in Artificial Intelligence. *Communication Studies*, *72*(1), 115–131. https://doi.org/10.1080/10510974.2020.1807380

Claridge, T. (2020). Trust and Trustworthiness: An Aspect of the Relational Dimension of Social Capital. *Institute for Social Capital*. https://www.socialcapitalresearch.com/trust-and-trustworthiness/

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

IWPR. (2019). Women, Automation, and the Future of Work. *Institute for Women's Policy Research* #CW476: 1—84. https://iwpr.org/wp-content/uploads/2020/08/C476_Automation-and-Future-of-Work.pdf

Jackson, S., & Panteli, N. (2021). A Multi-level Analysis of Mistrust/Trust Formation in Algorithmic Grading. International Federation for Information Processing, 12896 LNCS, 737–743. https://doi.org/10.1007/978-3-030-85447-8_61

Kahneman, D. (2003). Experiences of Collaborative Research. *American Psychologist* 58(9), 723-720. https://psycnet.apa.org/doi/10.1037/0003-066X.58.9.723

Kauff, M., Anslinger, J., Christ, O., Niemann, M., Geierhos, M., & Huster, L. (2021). Ethnic and Gender-based Prejudice Towards Medical Doctors? The Relationship Between Physicians' Ethnicity, Gender, and Ratings on a Physician Rating Website. *The Journal of Social Psychology*. https://doi.org/10.1080/00224545.2021.1927944

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., Umutlu, L. (2022). Combining the Strengths of Radiologists and AI for Breast Cancer Screening: A Retrospective Analysis. *The Lancet Digital Health* 4(7), E507-E519.

Muller, V.C. (2020). Ethics of Artificial Intelligence and Robotics. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-ai/?fbclid=IwAR3zBI5BYERCGCdEBZhAvLHExNJhPUJA9SYkvwteRUdmXBgB3ILfUk6y81o

Clark, Cory J., and Philip E. Tetlock. "Adversarial collaboration: The next science reform." *Ideological and political bias in psychology: Nature, scope, and solutions*. Cham: Springer International Publishing, 2023. 905-927.

Petersen, B.K., Chowhan, J., Cooke, G.B., Gosine, R., Warrian, P.J. (2022) Automation and the Future of Work: An Intersectional Study of the Role of Human Capital, Income, Gender and Visible Minority Status. Economic and Industrial Democracy. https://doi.org/10.1177%2F0143831X221088301

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Spiegelhalter, D. (2020). Should We Trust Algorithms? *Harvard Data Science Review*, *2*(1), 1–12. https://doi.org/10.1162/99608f92.cb91a35a

Stefano, V.D. (2019). 'Negotiating the Algorithm': Automation, Artificial Intelligence and Labour Protection. *Comparative Labor Law & Policy Journal*, 41(1). https://dx.doi.org/10.2139/ssrn.3178233

Tschopp, M. (2019, July 18). *Artificial Intelligence: Is it Worth the Risk?* SCIP. https://www.scip.ch/en/?labs.20190718

Wallace, B.C., & Paul, M.J. (2016). " Jerk " or " Judgmental " ? Patient Perceptions of Male versus Female Physicians in *Online Reviews*