FACING AI: THE EPISTEMOLOGICAL DIMENSION OF TRUST

Lei NIU

ABSTRACT: How do we understand trust in human experts? The total evidence view suggests that the beliefs of experts provide additional reasons for beliefs. Correspondingly, one should combine the beliefs of experts with one's own beliefs and at least give some epistemic weight to one's own evidence. On the contrary, the preemption view suggests that unequivocal deference to experts can be reliable and rational. Because AI is predictably becoming more accurate and reliable, it makes sense to ask how to trust AI and whether trust in AI is similar to or different from trust in human experts. By comparing human experts with AI and reflecting the debate between the total evidence view and the preemption view, this paper explores the epistemological dimension of trust in AI.

KEYWORDS: epistemic experts, evidence preemption, undercutting defeater, rebutting defeater

1. Introduction

Suppose that you are visiting a place where you have visited many years before, and you have a vague impression about its traffic routes. A passerby who is a local resident coincidentally provides you with a suggestion that conflicts with your current judgement. In such a case, how do you form a rational belief? According to the total evidence view (TEV)[1], you should take all evidence that is available and relevant into consideration and weigh them against one another. For instance, without evidence indicating the reliability of the passerby, trusting oneself and at least giving some epistemic weight to one's own evidence seem to be more likely to get right.

Assume further that there is a platform service, and it provides a suggestion that conflicts with the passerby. In this situation, you do not necessarily consider all the evidence available and aggregate their credentials. The main reason is that the suggestion of platform service is authoritative. Both your own evidence and the evidence of the passerby can be reasonably unusable. Deferring to the suggestion of platform service is the idea of the preemption view (PV).[2]

[1] Defenders include, for instance, Kelly (2010); Lackey (2018).

[2] Defenders include, for instance, Raz (1988); Zagzebski (2012); Grundmann (2020, 2021).

Both the total evidence view and the preemption view agree that laypeople should fundamentally rely on experts. Two views are compatible when laypeople agree with experts. If there is a consensus among one's memory, passerby and platform service, preempting the information of platform service and considering all information available do not make a great difference in terms of forming a final belief.

The total evidence view and the preemption view can come into conflict, particularly when the opinions of laypeople differ from those of professionals. Their dispute is whether one needs to make use of one's own first-order and domain-dependent evidence.[3] The total evidence view suggests that laypeople should compare the experts' evidence with their own evidence, either making an aggregation or simply using their own first-order evidence to identify experts. If the evidence of experts is different and doubtful, laypeople should distrust experts and give significant epistemic weight to their own evidence.

An intractable problem for laypeople is that it is often difficult for them to identify reliable experts and precisely evaluate and aggregate the evidence of experts. In light of this, the preemption view prohibits the use of one's own first-order evidence. Because laypeople are, by definition, incompetent compared with experts, their evidence is highly likely insufficient and even misleading. Any use of their own evidence can easily lead to judgements that deviate from experts' judgements. Given the fact that experts are more likely to be right than laypeople in a domain-specific way, the preemption view suggests that laypeople should not use their own first-order evidence and unequivocally defer to experts.

Since there is no perfect way for laypeople to identify experts unless they are experts themselves, both the total evidence view and the preemption view have problems. In particular, when the opinions of laypeople differ from those of professionals, the total evidence view can easily lead to distrust of experts, while the preemption view can easily lead to mistrust of experts.

The existing debates between the total evidence view and the preemption view mainly focus on the epistemic relationship between human experts and laypeople. However, it seems that the emergence of artificial intelligence (AI) is used as a source of information. AI is generally understood as the technology that allows computers and machines to maximise the chance of achieving defined goals. AI technologies have changed the way people gather and analyse information, and people now ask questions and seek advice from AI. This raises questions: how to

[3] Both the total evidence view and preemption view agree that one can rationally use domain-independent evidence to check the reliability of experts.

trust AI and whether trust in AI is similar to or different from trust in human experts?

To clarify, the paper does not engage in defining AI but focuses on the epistemic features of AI and the epistemic tasks AI can fulfil. In particular, the paper focuses on epistemic tasks that AI can outperform the epistemic tasks of human believers.[4] In addition, this paper does not argue that AI can be ontologically genuine believers and have beliefs. My focus is on the rational beliefs of human believers in the face of AI, and I use the outputs and judgements of AI interchangeably. Further, for the reason of space, this paper exclusively focuses on the epistemological dimension of trust in AI, although there are other significant dimensions, such as the moral dimension.[5]

By contrasting the total evidence view and the preemption view in the contexts of AI, this paper offers a novel perspective for assessing how humans can rationally incorporate or defer to AI-generated outputs. It seems that the outputs of AI are naturally in favour of the total evidence view, which suggests that one should combine different perspectives and at least give some epistemic weight to one's own evidence. However, unlike human experts who are able to disclose their evidence and arguments, AI systems that are trained rather than programmed are opaque, making it difficult to directly interpret the inputs of AI as first-order evidence and combine it with one's own evidence. As a result, deferring to the outputs of AI becomes an option. The question is then whether there is a rational dimension of trusting AI, which implies that giving up the use of human believers' first-order evidence can be epistemically rational. In this paper, I aim to explore the epistemological dimension of trust in AI, and I argue that there is a space for the preemption view for trusting AI, although AI plays a different role with human experts.

The plan of this paper is as follows: Section 2 will introduce the debate between the total evidence view and the preemption view regarding human experts. Based on this background, section 3 will flesh out the main features of AI and highlight its epistemological implications. I conclude in section 4.

[4] There exist lots of discussions about how AI can outperform the epistemic tasks of human experts. See, for instance, Grote and Berens (2020); Alvarado (2023).

[5] The moral or ethical dimension can be seen, for instance, in Mittelstadt et al. (2016).

2. Facing Human Experts

2.1 The Total Evidence View

The total evidence view arises from the topic of peer disagreement. To form a rational and justified belief, one has to consider not only first-order evidence but also higher-order evidence. One of the higher-order evidence is the disagreements of others, including one's peers and superiors. Consider the following example:

> A scientist provides a strategy to solve a mathematical problem, and the strategy is in fact sound. He is highly confident about his appropriate response to the first-order evidence. Meanwhile, his peer, who is competent and has the same first-order evidence, checks the strategy but disbelieves it. When acquiring the belief of one's peer, what is the rational attitude toward the strategy?

A disagreement with one's peer will influence one's rational attitude. Many philosophers believe that the rational attitude in this phase depends not on one's first-order evidence that supports the strategy but on the higher-order evidence, i.e., peer disagreement. Regarding the example, one needs to split the difference between one's own credence and the peer's credence.[6] That is, the rational attitude toward the strategy is to suspend judgement. However, the total evidence view claims that one should believe what the total evidence supports.[7] What is the total evidence?

(1) First-order evidence: original body of evidence that supports the strategy

(2) Higher-order evidence: peer disagreement

Even if the disagreement of a peer can temper one's confidence, a single peer disagreement is plausibly less strong. When one is highly confident and believes that one responds appropriately to the first-order evidence, assigning at least some epistemic weight to one's first-order evidence can lead to belief in one's strategy rather than suspension of judgement. A rational belief, in this sense, depends on how one weighs higher-order evidence and its relation with first-order evidence. The focus and the core idea here, according to the total evidence view, is that one should give at least some epistemic weight to one's own first-order evidence.

Things will be different when one encounters epistemic experts and superiors. In Grundmann's words (2021: 138), "epistemic superiority is the product of two independent factors: the available body of evidence and one's reasoning competences." Compared with laypeople, an epistemic expert is someone who possesses (1) an extensive body of domain-relative evidence and (2) highly

[6] See, for instance, Feldman (2006).

[7] See, for instance, Kelly (2010).

competent reasoning capacity. The epistemic superiority is a good reason to believe that experts are more likely than laypeople to get it right.

The issue of rational belief engendered controversy once one's disagreers are treated as epistemic experts. From our daily experience, we heavily rely on epistemic superiors and epistemic experts. According to the total evidence view, the beliefs of experts should be treated as an additional source of reason, and the weight of one's own evidence should be reduced, but it should never be reduced to zero. In the face of experts, the total evidence is:

(1) First-order evidence: original body of evidence; experts' evidence

(2) Higher-order evidence: expert disagreement

According to my interpretation, the total evidence view suggests that laypeople should consider both first-order evidence and higher-order evidence, and at least give some epistemic weight to their own first-order evidence. Unlike the cases of epistemic peers, splitting the difference between experts and laypeople, or assigning more epistemic weight to the beliefs of experts is unmotivated, because there is no precise way to guide how to assign epistemic weight to the beliefs of experts. This is the case particularly when we focus on fine-grained doxastic attitudes that assign credence to propositions. Put simply, assigning 0.5 credence requires suspension of judgement. While assigning greater than 0.5 requires belief with corresponding confidence, the opposite requires disbelief. Any use of one's own evidence can make a difference. That is, one's final belief will be different from the experts' belief and one's original belief.

Once you treat someone as an epistemic expert, it naturally follows that the beliefs of the epistemic experts in a specific domain are more likely to get right. Thus, deferring to experts is the most reliable strategy. In Raz's words (1988: 68):

> [W]e can expect that in the cases in which I endorse the authority's judgment my rate of mistakes declines and equals that of the authority. In the cases in which even now I contradict the authority's judgment that rate of my mistakes remains unchanged, i.e. greater than that of the authority. This shows that only by allowing the authority's judgment to pre-empt mine altogether will I succeed in improving my performance and bringing it to the level of authority."

Although one may accept that experts are more likely to be right than laypeople, the difficulty of assigning precise epistemic weight to experts' beliefs does not necessarily challenge the total evidence view. As a response, the total evidence view can suggest that the use of laypeople's first-order evidence is an important epistemic resource to judge the reliability of experts' expertise. For example, Lackey (2018: 238) proposed:

> follow the advice of an authority, except when one is certain that the authority is wrong; follow the advice of an authority, except when one knows that the authority is wrong; follow the advice of an authority, except when what the authority says is highly doubtful […] something would strike one as highly doubtful only against the background of one's other relevant information.

It may not be expected that all laypeople become experts, but that laypeople are able to judge some pieces of experts' evidence. Laypeople should compare their evidence with that of experts and learn from experts. In this sense, trusting themselves based on their own first-order evidence is a tangible way to trust experts and to develop understanding. On the other hand, allowing the use of laypeople's first-order evidence is a viable way to deliver experts' beliefs. When experts communicate to the general public, they should show their first-order evidence and try to teach the general public to understand and evaluate at least some of their evidence.

However, recall that the total evidence view conflicts with the preemption view, particularly when laypeople disagree with experts. This implies that laypeople will disagree with the evidence of experts, or they are unable to recognise the evidence of experts. The problem is that the more complex the issue, the more likely it is that the laypeople's evidence will be wrong and the more difficult it will be for laypeople to understand experts. Ultimately, it is more likely that laypeople will distrust experts and form a false belief.

In my view, although distrusting experts can be a serious problem, it can be costly to coerce or convince laypeople to give up the use of their first-order evidence, particularly when dishonest experts have already fuelled and driven distrust. In order to build a trust relationship and deliver knowledge and true beliefs, the total evidence view that recommends using laypeople's first-order evidence may not be the most effective but at least viable. After all, there is no perfect way to identify trustworthy experts unless laypeople become experts.

2.2 The Preemption View

The preemption view clearly conflicts with the use of first-order evidence. Defenders of the preemption view can indicate that, as a matter of fact, many issues are complex, and thus laypeople's evidence can be misleading, and what at first and even second glance on experts's conclusions appears simply outrageous to laypeople. In light of this, when laypeople's evidence and judgements conflict with those of experts, laypeople will distrust experts. The use of their own evidence can make it deviant from experts' beliefs, resulting in undesirable epistemic outcomes. By noticing that laypeople are epistemically incompetent in a domain-specific way and

their first-order evidence can be misleading, the preemption view states that laypeople's first-order evidence is unusable once epistemic experts are recognised. How can laypeople identify experts who are more reliable? In my view, there are two kinds of higher-order and domain-independent evidence that can be available to laypeople.

(1) Domain-independent evidence about experts

(2) Domain-independent evidence about laypeople themselves

There is domain-independent evidence about experts, including both positive and negative indicators.[8] Reputation, reliable track records, and publications are positive indicators about experts, while dishonesty, irresponsibility, and a conflict of interest are negative indicators. In addition, there is domain-independent evidence about laypeople themselves. It is easier for laypeople to identify their own epistemic abilities in certain areas than to identify experts. Laypeople can recognise that they are completely incompetent in some domains. If so, giving zero epistemic weight to their own first-order evidence is a desirable option.

Once laypeople's domain-independent evidence is a positive indicator for experts, considering first-order evidence can lead to judgements that deviate from experts. The use of their first-order evidence is not only instrumentally irrational but also epistemically irrational. Preemption, in this sense, has also been supported by some epistemic norms to jettison one's own first-order evidence. One of the most important epistemic norms is about higher-order undercutting defeaters, which was proposed by Grundmann.[9]

Higher-order defeaters are distinguished from undercutting defeaters and rebutting defeaters.[10] Undercutting defeaters provide reasons to challenge one's evidence, while rebutting defeaters provide reasons to support the opposite conclusion.[11] Suppose that you believe that it is now 5 o'clock by your watch. The evidence that your watch is malfunctioning directly undercuts your evidence about the time. The undercutting defeater can rebut a belief, but it does not necessarily indicate that your belief about the time is false, because it is possible that it is now 5 o'clock. A rebutting defeater is a new piece of evidence, which indicates that it is

[8] Grundmann (2025) has extensively explored the positive and negative indicators of experts.

[9] For discussions, see, for instance, Grundmann (2021).

[10] The distinction between rebutting defeaters and undercutting defeaters was discussed by Pollock (1974); Grundmann (2021).

[11] Of course, when a rebutting defeater works, the evidence challenges one's conclusion and must somewhat challenge one's own evidence. Here, a rebutting defeater is a new piece of evidence that directly challenges one's conclusion.

now 6 o'clock. Unlike undercutting defeaters and rebutting defeaters, higher-order defeaters have a retrospective effect, which illustrates that the belief was originally irrational.[12] For example, suffering from a mental disorder is a higher-order evidence about the unreliable status of checking your watch. Put differently, a higher-order defeater is a new piece of higher-order evidence that removes the justification of the use of one's first-order evidence.

What is the epistemic role of human experts? In Grundmann's view, unlike peer disagreements, the testimony of experts can be both first-order evidence that supports a proposition and higher-order evidence that indicates the quality of their reasoning process. When experts disagree with laypeople, experts' beliefs are both a higher-order defeater and a special case of undercutting defeater.[13] For one thing, regarding reasoning skills, there are good reasons to believe that experts are better than laypeople in a domain-specific way. In light of this, when laypeople use their own domain-specific evidence, theoretically, they have an inferior reasoning process. The recognition of experts provides higher-order evidence that the belief-forming process of laypeople is unreliable. For another, the beliefs of experts provide undercutting defeaters for laypeople's domain-specific evidence. Because there are good reasons to believe that experts have sufficiently more domain-specific evidence than laypeople, in the face of experts, laypeople's original evidence is inadequate or even misleading and then fails to sufficiently support their conclusions. In other words, when experts disagree with laypeople, experts should have considered most evidence laypeople have, and they are more likely than laypeople to respond appropriately to the same evidence. In light of this, the beliefs of recognised experts play a role of higher-order undercutting defeater and preempt the first-order evidence of laypeople, either for the use of identifying experts or forming a final belief.

When the opinions of laypeople differ from those of experts, it is worth noting that both kinds of domain-independent evidence can be fallible and misleading, and the preemption view can easily lead to mistrusting experts. In contrast, the total evidence view can easily lead to the distrust problem, resulting in a judgement deviating from epistemic experts. Because there is no perfect way to identify experts for laypeople, the total evidence view and the preemption view provide two different imperfect solutions, and distrusting and mistrusting reliable experts are the price of epistemic incompetence.

---

[12] For the discussion about the retrospective effect of higher-order defeater, see Lasonen-Aarnio (2014); DiPaolo (2018).

[13] See Grundmann (2021: 143). Although some works try to distinguish different kinds of defeaters, there exists a kind of rich background evidence having different defeating forces.

3. Facing AI

In recent years, it has been reported that AI is increasingly outperforming human experts.[14] It then makes sense to provide an analysis of AI and compare it with human experts. In this paper, instead of providing a definition of AI, this paper focuses on the main epistemic tasks AI can perform. AI, in this paper, is understood as the technology that allows computers and machines to maximise the chance of achieving defined epistemic goals. Why is there a need to trust AI? Although there are many cases where AI-powered applications could complete certain cognitive tasks for us, many applications of AI are used as tools or supplements for human believers. For example, AI can be used to reply to emails. But still, we might use AI to respond to some emails and respond to others ourselves. The use of AI does not raise a controversy about trust, because few people will believe that AI can be epistemic superiors to human believers in email replying, at least in the current stage. The question of trust arises when AI is comparable to the best in a given epistemic task and there are possible conflicting judgements between human believers and AI. In the following, the discussions of trust focus on AI with epistemic superiors profiles.

We have two main factors for identifying human experts or epistemic superiors, that is, evidence and reasoning competence. Compared with human believers, AI can store and analyse data more powerfully. Big data is a term that refers to massive and complex data that traditional methods cannot process.[15] In addition, AI is used to not only augment human intellectual capacities but also create new areas that human believers are sufficiently incompetent in. When dealing with massive amounts of data, AI enables the delegation of challenging pattern identification, learning, and other activities to computer-based approaches. Given specific tasks, AI can excel at processing data and bring a unique set of qualities, and AI has been increasingly applied to replace important decision-making of human believers, such as medical diagnosis, allocation of jobs, and financial services.[16]

[14] For instance, in medical diagnosis or treatment recommendations, AI has demonstrated epistemic superiority over human experts. Grote and Berens (2020).

[15] For empirical research on the processing of big data, see, for instance, Molas and Nowak (2021: 4).

[16] Economic imperatives motivate the design and development of AI technologies. In order to make profits, AI is designed to greatly improve efficiency and reduce costs. The replacement of human work with AI can largely increase efficiency and lower labour costs. In addition, political imperatives, such as geopolitical competition, shape and drive the trend toward replacement. For empirical evidence, see, for instance, Deranty and Corbin (2022).

AI technologies are designed to perform epistemic tasks that people cannot or do not do properly.[17] To have epistemic impacts on the beliefs of human believers, the targets need not only be human experts or authority. For example, the outputs of microscopes or calculators can function as an epistemic defeater. Instead of arguing that we can classify AI as cognitive agents or epistemic experts, it is plausible that AI can have a superior or expert profile with respect to some epistemic tasks. As a result, it makes sense to ask how to trust in AI and whether trust in AI is similar to or different from trust in human experts.

What is so special about AI? In addition to the epistemic capacity, a remarkable feature of AI lies in its opacity. While human experts can disclose their evidence and rationalise their conclusions, the evidence and reasoning process of AI are not fully transparent and explainable. Also, a very significant worry is whether the issue of opacity raised in the paper is specific to AI. Can't we just treat AI systems like calculators and other devices that we know are reliable?

Consider the calculator for example. A calculator is highly reliable, and it operates on arithmetic algorithms. Its calculation is based on understandable mathematical rules and thus not inherently opaque to users. In addition, users can identify the reliability of the calculator by the use of their own calculation. With respect to calculation, human believers are not sufficiently incompetent and should not be treated as laypeople. Even for complicated calculations, human believers do not necessarily give up their own evidence and conclusion. Nevertheless, AI systems can be different from calculators. Instead of operating by pre-specified designs and rules, AI can mimic the brain's style of learning and be trained by itself. The reasoning process of AI runs independently of human control, and the opacity of AI becomes a problem for trust.

In response, the core task for the developments of AI is to alleviate the opacity of AI and achieve accountability. Explainable AI (XAI) methods are used to alleviate the opacity of AI.[18] For instance, pedagogical explanation is used to provide an explanation of how AI can collect data and what factors can influence most to the outputs; demographic explanation is used to provide statistics on outcomes that are similarly classified; case-based explanation is used to provide representative examples and counterexamples. In addition, to achieve AI accountability, it is necessary to establish constraints on the exercise of powers, shared norms, and

[17] Discussions about how AI technology is designed, developed, and deployed to achieve epistemic purpose see, for instance, Alvarado (2023).

[18] Different explanations can be seen also in Zerilli et al. (2018); Binns et al. (2018).

sanctions on AI designers and decision-makers.[19] Both strategies try to provide higher-order evidence about AI's reasoning capacity, interpretability, and accuracy, but it is still difficult to directly translate the inputs of AI to the first-order evidence of human believers.[20]

The rest of this paper attempts to provide an analysis of how opacity challenges the total evidence view and the preemption view of trust and provide an analysis of the epistemic role AI can play in the context of trust.

3.1 The Total Evidence View

The total evidence view suggests that one should make use of one's own first-order evidence in forming beliefs. However, the opacity inherent in AI presents significant obstacles to effectively using one's own first-order evidence. Let us consider some examples:

> **Driving Accidents Prediction**:[21] Individuals can attempt to predict the possibility of their driving accidents. The typical evidence includes road conditions and weather conditions. The input of AI includes your age, gender, number of trips taken at night, level of adherence to speed limit, miles per month, traffic patterns, and so on.
>
> **Weather Forecast**: Suppose a layperson with respect to meteorology is aware of the forecast of ChatGPT saying the weather is expected to be mostly cloudy throughout the day, and there is a very high chance of precipitation two hours later. The inputs of ChatGPT include temperature, wind speed, and weather simulation models. The layperson knows from experience that ChatGPT is highly reliable but not infallible and can see that the sky is blue and lacking in clouds.
>
> **Skin Cancer Diagnosis**:[22] A plethora of high-profile scientific publications has been reporting about AI outperforming clinicians in skin cancer diagnosis. By performing a clinical history and interview, conducting a physical exam, and performing diagnostic testing, clinicians conclude that a patient has disease X. However, AI represents its output in risk score and indicates disease Y. When making decisions, clinicians either stick to their own opinion or defer to the AI's output.

With respect to the total evidence view, there are two questions. First, whether the outputs of AI provide additional reasons for beliefs? In the cases of peer disagreement, disagreers are required to split the difference with their peers because peers have the equal probability of getting it right. As for AI, there is no precise way

[19] For AI accountability, see, for instance, Johnson (2021).

[20] Similar discussions about explainable AI methods are also seen in Fleisher (2022).

[21] Zerilli et al. (2018)

[22] Grote and Berens (2020).

to split probabilities, and then there is the trouble with combining and aggregating different perspectives.

The second question about the total evidence view is whether laypeople can translate the inputs of AI as first-order evidence or identify the reliability and accuracy of AI by the use of laypeople's first-order evidence. In both driving accident prediction, weather forecast, and skin cancer diagnosis examples, the inputs of AI are abstract and different from those of human believers (e.g., real-time road conditions, the blue sky, clinical history and interview). Some models can operate with hand-labelled inputs, while other models can make use of any inputs they define as appropriate (unsupervised learning).[23] In addition, the layperson cannot directly learn the reasoning process of AI, because human believers and AI have different cognitive systems and reasoning abilities. The inputs of AI can be processed in a complex way that is hard to inspect for human believers. In light of this, it is difficult to directly interpret the inputs of AI as first-order evidence and connect them to conclusions. Giving more epistemic weight to the layperson's evidence will lead to distrust and ignore the outputs of AI. On the contrary, giving epistemic weight to outputs of AI will lead to a judgement that giving up the use of one's own evidence. In other words, when there are reliable track records of AI, the use of layperson's evidence is either blind or inaccurate. The sense of diversity and opacity makes it troublesome for the total evidence view of trust.

3.2 The Preemption View: Higher-Order undercutting Defeater

Let us now consider the preemption view of trusting AI. The preemption view prohibits aggregating laypeople's first-order evidence with AI and the identification of AI's reliability by laypeople's first-order evidence. In other words, according to the preemption view, trusting AI implies that ignoring one's own first-order evidence can be epistemically rational. The question is whether opacity is a challenge for the preemption view.

With reference to AI, the undercutting defeater argument is not convincing, especially when we take the AI's features of opacity into consideration. The opacity of AI indicates that laypeople have difficulties translating the inputs of AI to first-order evidence, and they are unable to compare their evidence and learn the reasoning process of AI. It is not the case that AI has considered the evidence available to the laypeople and responded more appropriately to the same evidence. Because the evidence and reasoning skills of AI are entirely different from those of

[23] For some discussions about unsupervised learning, see, for instance, Schermer (2011); Van Otterlo (2013).

human believers, the outputs of AI do not directly remove the justification of the first-order evidence of human believers. In light of this, the outputs of AI cannot function as undercutting defeaters, and human beings cannot let AI's inputs preempt their own.

In addition, the opacity of AI weakens or even eliminates the force of the higher-order defeater. Although AI becomes increasingly reliable, it seems that the opacity of AI poses a challenge to the explainability and epistemic superiority of AI.[24] Because the comparison of epistemic competence is not possible, the epistemic superiority of AI can be questioned, particularly when human experts have expertise to some extent. When the judgements of human experts conflict with the outputs of AI, it is questionable to indicate that the beliefs of human experts are results of a flawed process.

3.3 The Preemption View: Higher-Order Rebutting Defeater

Because AI systems are increasingly embedded in various aspects of life and the track records of AI become increasingly reliable, it seems that completely ignoring the outputs of AI is not epistemically rational. The outputs of AI can introduce epistemic defeaters that turn justified beliefs into unjustified beliefs. The question is then: What kinds of AI and what kinds of epistemic defeaters can support the preemption view for trusting AI?

As I illustrated above, the outputs of AI do not directly challenge the first-order evidence of human believers and their process of evaluating their evidence. Nevertheless, the outputs of AI can create domains or propositions in which human believers are sufficiently incompetent and then provide the retrospective effect that one's beliefs were never rational to start out with. In other words, the outputs of AI can be evidence about the unreliable epistemic status of human believers and function as higher-order defeaters. To illustrate, it will be helpful to consider the weather forecast example and cancer diagnosis example again. AI diagnosis can integrate vast amounts of real-time data and detect patterns that are imperceptible to human experts. The weather models can predict seasonal or climate patterns over

[24] With respect to the preemption view for trusting human experts, there are proponents of the track record argument, including, for instance, Raz (1988); Zagzebski (2012). The idea is that trusting AI can produce the most reliable result. In order to promote the best track record, one should give zero weight to one's own evidence in a domain-specific way. In this sense, the track record of AI can provide a strong instrumental reason for trusting AI and is in favour of the preemption view. Instead of arguing that the track records of AI can provide sufficient reasons to ignore laypeople's first-order evidence, here we can accept that the reliable track records are the foundation of trusting AI.

months or years. In light of this, the use of AI can create domains where human believers are sufficiently incompetent in a domain-specific way. In addition, the opacity of AI is a challenge for the translation from the inputs to first-order evidence of human believer, but it does not necessarily pose a challenge for its epistemic superiority. Put differently, as for specific epistemic tasks, although there can be epistemic reasons to question the transparency of AI, there is no epistemic reason to believe that human believers can form more reliable beliefs than AI, such as processing big data or predicting the weather up to years.

In addition, unlike typical higher-order defeaters, such as peer disagreements, mind-distorting drugs, and biases, the outputs of AI can change whether one's original evidence supports their beliefs and then have strong rebutting force. In summary, there are two central features of the new defeaters. First, the outputs of AI can provide reasons to support new conclusions. Second, the outputs of AI can provide reasons to undermine the justification of drawing conclusions by human believers. Because the higher-order defeater has rebutting force, I will name it as a higher-order rebutting defeater. We can turn these considerations into the following argument:

(1) I am justified to believe that there are reliable track records of the outputs of S (AI system) about a proposition p in domain D. (AI reliability)

(2) I don't know what the inputs of S are and how S functions. (Opacity)

(3) I am justified to believe that I have a very low degree of reliability on p. (Human incompetence)

(4) When my epistemic goal is to form true beliefs regarding p, in the face of the outputs of S regarding p, either I ignore it, or combine it, or defer to it.

(5) If (1) is true, I cannot be justified to ignore the outputs of S regarding p.

(6) If (2) and (3) are true, I cannot be justified to combine the outputs of S regarding p.

(7) Therefore, I am justified to defer to the outputs of S. (from 4, 5, 6)

As trusting human experts, there is no perfect way to trust AI. The argument tries to narrow down the possible epistemic responses to the outputs of AI and eliminate irrational alternatives, leaving the deference to AI as the most likely rational response. Premise (1) is the basic foundation for considering the outputs of AI. When there is no reliable track record, human believers ought to epistemically ignore the outputs of AI. This avoids the overgeneralisation of trusting AI in the context where its track record is less clear. Premise (2) is the key feature of AI that gives rise to the trust question. (3) is used to narrow the scope of trustworthy AI and suggest that not only the reliability of AI but also the normatively binding force

provide the reasons for trusting AI. The examples above tend to show that human believers have a very low degree of reliability on speed and scale of information processing, pattern recognition, prediction, and multitasking. The epistemic superiority of AI is not grounded in the opacity of AI but in the epistemic incompetence of human believers.

Premise (4) articulates the possible epistemic responses to the outputs of AI when people have to draw conclusions. In particular, the epistemic goal must be stipulated to form true beliefs. When people have no interest regarding the proposition or aim to avoid false beliefs, the justified epistemic responses are different.[25] (5) and (6) argue that two of the three options are not epistemically justified. The epistemic justification does not assume specific theories of knowledge and justified beliefs. In general, they will predict that it is not possible to form a justified belief when ignoring a defeater. Notably, it is epistemically irrational to ignore and dismiss the outputs with demonstrated reliability and epistemic superiority. In addition, the discussion of the total evidence view suggests that combining the inputs and outputs is not feasible.[26] Laypeople cannot aggregate their own first-order evidence with those of AI and split the difference to some extent. In the face of the outputs of AI, deferring to it is most likely the rational option.

In my view, this conditional application provides a picture of how the belief of AI functions as a higher-order rebutting defeater that challenges the justification of drawing conclusions by human believers in a domain-specific way. When the outputs of AI can create domains where human believers are sufficiently incompetent, and laypeople have to draw conclusions, their reasoning process and conclusions can be defeated and preempted. In light of this, the preemption view can provide an explanation for the rational dimension of trusting AI. As an implication of the preemption view, the development and application of AI should ensure AI's reliability by undergoing rigorous testing and building extensive track records.

[25] The epistemic goal can conflict with other considerations. For instance, the deference to AI can lead to moral problems, such as a misattribution of responsibility (Elish 2019; Constantinescu *et al.* 2022). I shall leave a detailed discussion for the comparison of different considerations here.

[26] There are discussions about hybrid agents that emerge from the interactions between humans and technology, such as Ferrario et al. (2024). However, the hybrid agents are about the distribution of cognitive labours rather than the combination and aggregation of first-order evidence of human believers and the inputs of AI. The case is that both human experts and the artefact in question have different epistemic advantages, and a superiority of AI is not strictly provided. In such a case, trusting in AI can be significantly challenged, and human believers cannot form justified beliefs simply based on the outputs of AI.

4. Conclusion

In this paper, I have explored the epistemological dimension of trust in AI by reflecting on the debate between the total evidence view and the preemption view. With respect to human experts, both the total evidence view and the preemption view provide imperfect solutions to identify and trust experts. With respect to the application of AI, I have argued that the total evidence view is not viable. It is not preferable to either aggregate laypeople's first-order evidence with that of AI or assess the reliability of AI based on the comparison of them. In contrast, the preemption view can be in favour of trusting AI. Once the superiority of AI can be recognised, giving zero epistemic weight to laypeople's first-order evidence becomes an option. In addition, unlike the beliefs of human experts that play a role in higher-order undercutting defeaters, the outputs of AI can provide us with higher-order rebutting defeaters.[27]

References

Alvarado, Ramón. 2023. "AI as an Epistemic Technology". *Science and Engineering Ethics,* 29(5): 1–30.

Binns, Reuben; Van Kleek, Max; Veale, Michael; Lyngs, Ulrik; Zhao, Jun; Shadbolt, Nigel. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 377: 1–14.

Constantin, Jan and Grundmann, Thomas. 2020. "Epistemic Authority: Preemption through Source Sensitive Defeat". *Synthese* 197 (9): 4109–4130.

Constantinescu, Mihaela; Vică, Constantin; Uszkai, Radu & Voinea, Cristina. 2022. "Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors". *Philosophy and Technology* 35 (2): 1–26.

Deranty, Jean-Philippe and Corbin, Thomas. 2024. "Artificial Intelligence and Work: A Critical Review of Recent Research from the Social Sciences". *AI and Society*: 1–17.

DiPaolo, Joshua. 2018. "Higher-Order Defeat Is Object-Independent". *Pacific Philosophical Quarterly* 99 (2): 248–269.

[27] **Acknowledgements:** I am grateful to Thomas Grundmann, Paul Silva, and the anonymous reviewers for their valuable comments on earlier versions of this paper. I also want to thank the China Scholarship Council for supporting my current research.

Elish, Madeleine Clare. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction". *Engaging Science, Technology, and Society* 5 (2019): 40–60.

Feldman, Richard. 2006. "Epistemological Puzzles about Disagreement". In *Epistemology Futures*, edited by Stephen Hetherington, 216–236. Oxford University Press.

Ferrario, Andrea; Facchini, Alessandro and Termine, Alberto. 2024. "Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems". *Minds and Machines* 34 (3): 1–27.

Fleisher, Will. 2022. "Understanding, Idealization, and Explainable AI". *Episteme* 19 (4): 534–560.

Grote, Thomas and Berens, Philipp. 2020. "On the Ethics of Algorithmic Decision-Making in Healthcare". *Journal of Medical Ethics* 46 (3): 205–211.

Grundmann, Thomas. 2021. "Preemptive Authority: The Challenge From Outrageous Expert Judgments". *Episteme* 18 (3): 407–427.

Grundmann, Thomas. 2021. "Facing Epistemic Authorities: Where Democratic Ideals and Critical Thinking Mislead Cognition". In *The Epistemology of Fake News*, edited by Sven Bernecker, Amy K. Flowerree and Thomas Grundman, 134–154. Oxford: Oxford University Press.

Grundmann, Thomas. 2025. "Experts: What Are They and How Can Laypeople Identify Them?". In *Oxford Handbook of Social Epistemology*, edited by Jennifer Lackey and Aidan McGlynn, 85–106. Oxford University Press.

Johnson, Deborah G. 2021. "Algorithmic Accountability in the Making". *Social Philosophy and Policy* 38 (2): 111–127.

Kelly, Thomas. 2010. "Peer Disagreement and Higher Order Evidence". In *Social Epistemology: Essential Readings*, edited by Alvin I. Goldman and Dennis Whitcomb, 183–217. Oxford University Press.

Lackey, Jennifer. 2018. "Experts and Peer Disagreement". In *Knowledge, Belief, and God: New Insights in Religious Epistemology*, edited by Matthew A. Benton, John Hawthorne & Dani Rabinowitz, 228–245. Oxford: Oxford University Press.

Lasonen-Aarnio, Maria. 2014. "Higher-Order Evidence and the Limits of Defeat". *Philosophy and Phenomenological Research* 88 (2): 314–345.

Mittelstadt, Brent; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra and Floridi, Luciano. 2016. "The Ethics of Algorithms: Mapping the Debate". *Big Data and Society* 3 (2).

Molas, Gabriel and Nowak, Etienne. 2021. "Advances in Emerging Memory Technologies: From Data Storage to Artificial Intelligence". *Applied Sciences* 11, no. 23: 11254.

O'Leary, Daniel E. 2013. "Artificial Intelligence and Big Data". *IEEE intelligent systems*, *28*(2): 96–99.

Raz, Joseph. 1988. *The Morality of Freedom*. Oxford: Oxford University Press.

Schermer, Bart W. 2011. "The Limits of Privacy in Automated Profiling and Data Mining". Computer Law & Security Review 27(1): 45–52.

Van Otterlo, Martijn. 2013. "A Machine Learning View on Profiling". In *Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology*, edited by Hildebrandt M, de Vries K., 41–64. Abingdon: Routledge.

Zagzebski, Linda Trinkaus. 2012. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press.

Zerilli, John; Knott, Alistair; Maclaurin, James; Gavaghan, Colin. 2018. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?". *Philosophy and Technology* 32 (4): 661–683