Φιλοσοφια
Volume 27, 2: 2026

AGAINST MODAL ACCOUNTS OF ALGORITHMIC ROBUSTNESS

Kurt Christian B. Tubera
University of the Philippines - Diliman, Philippines

Machine learning (ML) models are increasingly used to make decisions in medicine, economics, and other high-stakes environments that demand epistemic trust. The role they play is epistemic because they produce outputs that inform us, and such outputs can either lead to true or false beliefs. Even if most algorithms tend to generate true outputs during their training phase, they are nonetheless untrustworthy because they occasionally produce false outputs when deployed in real-world environments. Philosophers have proposed algorithmic robustness as a response to this problem. Algorithmic robustness requires ML models to be stably truth-conducive, avoiding accidentally true outputs. One leading approach explains algorithmic robustness in terms of modal conditions, which require algorithm outputs to be true in nearby possible worlds for a system to be robust. This paper challenges the modal conditions for algorithmic robustness. I claim that such conditions are unnecessary for robustness. In support of this claim, I argue that the modal conditions for algorithmic robustness fail to preserve constitutive epistemic links (i.e., relations between evidence, belief, and truth) that occur in actual situations. By presenting cases wherein these links obtain without satisfying the modal conditions, I show that an algorithm can remain robust even without invoking counterfactual dependence.

Keywords: artificial intelligence, knowledge, modal conditions, State algorithmic robustness

INTRODUCTION

Machine learning (ML) models are increasingly being deployed in high-stakes environments. These models are algorithms trained on data sets to ensure reliability in producing accurate information. These environments include medicine (Obermeyer and Emanuel 2016; Bernhammou et al. 2020; Rajpurkar et al. 2017; Zech et al. 2018), economics (Zheng et al. 2020; Zheng et al. 2022), transportation or traffic engineering (Ranjan et al. 2019), and so on. In all these cases, algorithms offer promising prospects regarding human-AI collaboration because AI systems generally produce true outputs. A common factor among the various ways in which algorithms are applied to these fields is that they are optimized and implemented for decision-making. The role they play is *epistemic.* Such deployments of ML models require that our beliefs, formed by the outputs of these systems, are justified and true. Not to mention, these fields are social facets that have a crucial impact on human lives.

However, even if algorithms are deemed reliable insofar as they generate true outputs at a statistically high rate, experts still find it problematic to trust them. The main reason for this distrust is the brittleness of these algorithms. The output that they produce, albeit often true, can easily be false if small environmental tweaks are made. This may be a result of factors such as the limited data set through which the ML model was trained and the opacity of the algorithm.[1] As a solution to this problem, philosophers have suggested that algorithms should be robust. Algorithmic robustness is cashed out in terms of its stability in attaining a target with respect to multiple variations and interventions of the environment where the algorithm is situated (Freisleben and Grote 2023). If algorithms are reliable and robust, we have more reason to trust that their outputs are true and non-accidentally produced.

More recently, epistemologists have been in the business of proposing an account that better explains the concept of algorithmic robustness. At the core of their concern was an epistemological question: How do we ensure that the beliefs that we form through the outputs of ML models are justified? In addition, how do we know when we can trust AI systems to provide us with accurate information? A prominent and promising move among them is the counterfactual one, according to which we can cash out algorithmic robustness in terms of modal conditions. By this, I mean that they are speaking of algorithms as being robust just in case they produce true outputs in close or nearby possible worlds.[2] In the strategy of looking into possibilities, we have a better way of eliminating luck in instances where algorithms produce outputs. After all, an event is lucky if it fails to be true in close or nearby possible worlds.[3]

But are we correct to define a robust algorithm in modal terms? This paper challenges counterfactual accounts of algorithmic robustness, arguing that modal conditions are unnecessary for it. Since counterfactual accounts propose modal conditions as defining traits of algorithmic robustness, challenging their necessity is enough to challenge such accounts. To challenge the necessity of such conditions, I draw from the literature on constitutive epistemic links and knowledge. I argue that, in the same way that appealing to modal conditions for knowledge fails to preserve the constitutive links between variables that lead to knowledge, appealing to modal conditions for thinking about algorithmic robustness does not sustain relevant variables in the actual world. If this is correct, there are instances in which an algorithm is genuinely robust but does not satisfy the modal conditions.

The paper is structured as follows. In the next section, I will go over the recent studies on the implementation of algorithms in the different facets of human life, including the ones already mentioned above. These discussions show that algorithms are reliable but brittle tools. Such brittleness is sufficient to render AI systems untrustworthy. Hereafter, I proceed with the proposal made by some philosophers regarding algorithmic robustness, which states that for an ML model to be trustworthy, it needs to be robust. I then turn to recent discussions on the counterfactual accounts of algorithmic robustness. I clarify their merits over other non-modal accounts. Next, I go over the discussions of constitutive epistemic links, and how a modal analysis of knowledge does not preserve such links. This sets the framework for analyzing the modal accounts for algorithmic robustness. The succeeding portions of the paper are dedicated to revealing a flaw in such counterfactual accounts.

RELIABLE BUT BRITTLE ALGORITHMS

Multiple studies are showing promising results regarding the deployment of ML models in various environments. In the field of medicine, such algorithms are implemented for organizing patient files[4] and even detecting prospective and developing illnesses.[5] In economics, such algorithms are being applied through reinforcement learning (RL) frameworks, which help economists in terms of economic and policy designs. They also provide multi-level simulations to help with optimal taxation strategies.[6] In these cases, ML models are deemed reliable insofar as they typically produce true outputs. Their reliability is minimally indicative of their trustworthiness. After all, when we rely on such models to make decisions related to the environments mentioned above, we seem to be trusting them to give us the right information.[7] But is reliability enough for a trustworthy algorithm? One of the pressing concerns related to trusting ML models is their brittleness. They generate false outputs given minimal to maximal irrelevant or relevant changes to the data being analyzed. In other words, the reliability of the ML model is unstable across different environmental applications. Bjerring, Busch, and Munch (2025) call this *algorithmic brittleness*.

To further explain algorithmic brittleness, it is worth looking into the classes of concrete examples where such brittleness is performed, namely, (1) *adversarial brittleness,* (2) *overfitting brittleness,* and (3) *perspectival brittleness*. In the remainder of this section, I discuss these three classes of brittleness.

There are multiple cases in which an algorithm can be brittle because of adversarial attacks. An example of this is minimally targeted manipulations in traffic symbols that distort and lure such models into producing false outputs (Ranjan et al. 2019, 2). More concretely, Eykholt et al. (2018, 1626) showed an instance where they deliberately manipulated an ML model to interpret a "Stop" sign as a "Speed Limit 45" sign by adding minimal perturbations like black and white stickers. As a result, self-driving cars may be more prone to accidents if there are minimal environmental tweaks that occur. While similar instances may happen to humans, they are nonetheless not overtly affected by minimal perturbations. For instance, having small graffiti on road signs does not immediately lead humans to change their judgment or interpretation of such signs. So, the brittleness that occurs among ML models is of a much more problematic sort if we are to allow them to participate in high-stakes environments.

Overfitting brittleness, like adversarial brittleness, is particularly performative in cases wherein ML models qualify irrelevant features in their output production due to environmental complexities. An example of this is seen in a work by Zech et al. (2018), where the ML models utilized to diagnose pneumonia would not function well under different real-world clinical environments because the model includes irrelevant objects like hooks at the edges of the image, as part of the decisive factor for diagnosis. Similar evidence was shown by Makino et al. (2022), where an ML model used to classify skin cancer inappropriately included surgical skin marking in their diagnosis of malignant melanoma. The brittleness of overfitting occurs precisely when the ML model learns to create shortcut predictions by including irrelevant details into its predictive process.

Natural distribution shifts affect the output of the ML model, thereby revealing another sort of brittleness, namely, perspectival brittleness. This is particularly evident in the study by Beede et al. (2020), where ML models as competent as an ophthalmologist fail to accurately diagnose an illness because the environment's lighting in which the patient is being analyzed is non-ideal. What typically happens in cases of perspectival brittleness is as follows. First, input images are given to ML models, and they are then expected to detect objects within the image. Second, small shifts in the perspective of the image are applied. Often, small rotations are done to the input image, and the same process occurs for the ML model. After the small shifts, it is then found that ML models tend to be vulnerable to making erroneous judgments or outputs because of the minimal changes (Engstrom et al. 2019, 1).

All these examples show that an algorithm can be reliable but brittle. It is reliable in the sense that it has the propensity to produce true outputs, or it produces more true outputs than false ones.[8] But a brittle algorithm is not trustworthy. In environments where the stakes are high, there is an immediate unease about allowing brittle ML models to participate in the decision-making process. As a response to this problem, experts proposed that ML models should not only be reliable but also robust to minimally qualify for trustworthy AI. In the next section, I rehearse the discussion on *algorithmic robustness.*

ALGORITHMIC ROBUSTNESS

What brittleness has revealed, if anything, is that reliability is not the only epistemic property crucial for algorithms (Grote et al. 2024). Given the problem that ML models are reliable but brittle, experts endeavored to look for possible ways to sidestep such brittleness. One attractive move among experts is the need for robustness (Freisleben and Grote 2023; Vandenburgh 2023; Grote 2024). But how exactly can we understand robustness? This section will be centered on clarifying recent accounts of robustness as related to ML models. I will no longer go over the literature on robustness analysis in science insofar as it might stray far from the main point of this study. To my mind, the most recent discussions on the conceptual analysis of robustness in ML are enough for this paper.

According to Freisleben and Grote (2023, 7), "[t]he key idea behind robustness is that the robustness target remains stable *under changes* in the robustness modifier." Allow me to take time to unpack this account. Notice that Freisleben and Grote's conception of robustness is multi-placed in nature. That is, to understand robustness, we consider (1) the robustness target and (2) the robustness modifier. The robustness target, in this case, is the entity being analyzed for robustness. Robustness modifier, on the other hand, is the entity towards which the robustness target has a relationship and remains stable (Freisleben and Grote 2023, 7). As such, robustness is understood in terms of how stable the robustness target is with respect to the modifier.

In the cases mentioned above, ML models are reliable but not robust because such models, as targets, fail to be stable with respect to their modifiers. In cases where the ML model exhibits adversarial brittleness, the target fails to be stable by taking into account irrelevant changes in the modifier in the form of adversarial attacks. In cases

where the ML model exhibits overfitting brittleness, the target associates irrelevant details from the input data from non-training environments, thereby leading to an instability with respect to its goal (e.g., diagnosing pneumonia, cancer, etc.). Finally, in instances where ML models exhibit perspectival brittleness, the natural distribution shifts make the target unstable with respect to its modifier. The case in point here is that treating robustness as a multi-place concept aids in explaining why ML models are not robust in the classes of algorithmic brittleness.

Another thing to note in Freisleben and Grote's (2023) account is that they treat robustness as a causal concept. Recall that their proposed key idea behind robustness is the stable relationship between the target and the modifier "under changes" in the modifier. The word "change" in such an account is treated as a causal terminology (Freisleben and Grote 2023, 7). What this ultimately suggests is that the target and the modifier are causally linked. A robust target, for that matter, does not easily break the causal link with the modifier if irrelevant interventions happen in between.

Freisleben and Grote's account seems to be an apt characterization of what robustness in algorithms should be. But is it apt? Bjerring, Busch, and Munch (2023) do not think so. This is due to the reason that treating robustness as a causal phenomenon is restrictive, and as such, makes for an extremely narrow conception of robustness (Bjerring, Busch, and Munch 2023, 19). Consequences that arise between the target and modifier are not always causal. And so, in the causal account proposed by Freisleben and Grote, logical and functional consequences where ML models are genuinely robust would be excluded. This warrants a different account of robustness that caters to such a worry.

COUNTERFACTUAL ACCOUNTS OF ALGORITHMIC ROBUSTNESS

An alternative approach to robustness is the counterfactual or modal accounts. In such accounts, algorithmic robustness is evaluated in terms of possible worlds. In other words, modal accounts investigate how possible situations wherein things could have been different, and in turn, how those differences affect the outputs of the algorithm. The benefit of these sorts of accounts is that they have a way to eliminate lucky true outputs from algorithms. In this section, I will highlight two modal accounts, namely, (1) Vandenburgh's safety, and (2) Bjerring, Busch, and Munch's Nozickian accounts of algorithmic robustness.

Vandenburgh proposes that we should view algorithmic robustness in terms of safety. This takes its roots from Ernest Sosa and Duncan Pritchard's works on epistemic safety. Their central aim in epistemic safety is to offer a new definition of knowledge.

An epistemic subject, S, knows a proposition, $p =_{\text{df.}}$ (i) S believes p, (ii) p is true, and (iii) S's belief is *safe*.

It is crucial to note here that such conditions are offshoots of the traditional justified-true-belief (JTB) definition of knowledge. Epistemologists, since then, have made certain revisions to such a definition because Gettier successfully challenged the sufficiency of JTB.[9] Safety is one of those approaches, and it takes its move towards

dropping the justification condition, and changing it with a modal requirement, namely, safety.

Vandenburgh (2023) was mostly concerned with thinking about how ML models can be epistemically trustworthy. That is, how can we trust such models to lead us to knowledge, given the brittleness of such algorithms? In Vandenburgh's account, robustness is the answer, and it is cashed out in the following terms:

(Algorithmic Robustness^{S}). An algorithm is robust $=_{df.}$ The algorithm's outputs are safe.[10]

Ultimately, such an account is dependent on what we can call, for brevity, *algorithmic safety.*[11] Accordingly, we define algorithmic safety as follows.

Algorithmic Safety $=_{df.}$ $O(p) \Box\!\!\rightarrow p$[12]

Algorithmic safety requires that if an ML model produces a true output p through a certain process M, p would be true in nearby possible worlds. Algorithmic safety eliminates cases of accidentally produced true outputs by algorithms. It ensures that the method is reliable across nearby possible worlds. And so, algorithmic brittleness should be eliminated in safety. In a more recent work, Levin Hornischer (2026) advances the safety-based modal account of algorithmic robustness by showing that we can understand algorithmic robustness in probabilistic terms, thereby weakening the condition just enough to allow fallibility without leading to triviality.

More recently, Bjerring, Busch, and Munch (2025) offered an improved account of algorithmic robustness.[13] The route that they take towards proposing a distinct account is in the same lines as Vandenburgh, that is, through counterfactual semantics. However, their point of departure from Vandenburgh's account is that they utilized a distinctive approach in modal epistemology. Instead of sticking to the safety condition, they appealed and adapted Robert Nozick's modal conditions for knowledge. What exactly are these conditions?

Nozick (1981, 172), in line with the epistemologists who took a modal approach, proposed to define knowledge in the following way. S knows that p if, and only if,

(i) p
(ii) $B(p)$
(iii) $\sim p \Box\!\!\rightarrow \sim B(p)$ (Sensitivity Condition)
(iv) $p \Box\!\!\rightarrow B(p)$ (Adherence Condition)

Like the typical proponents of safety, Nozick's definition contains the truth of the proposition (i), and the belief in the said proposition (ii). However, Nozick's account differs from safety in that it offers counterfactual conditionals in (iii) and (iv). Condition (iii) states that had p been false, then S would not believe p. This is the sensitivity condition. Condition (iv) states that had p been true, S would believe p.

Applying both sensitivity and adherence conditions, Bjerring, Busch, and Munch (2025, 15) came up with the following definition of algorithmic robustness.

(Algorithmic Robustness^{N}). An algorithm is robust $=_{df.}$ Its outputs are sensitive and adherent.

As evident in the definition above, Bjerring, Busch, and Munch adopt a Nozickian account of algorithmic robustness where sensitivity and adherence are both necessary and sufficient conditions.

Algorithmic Sensitivity $=_{df.} \sim p \; \Box\!\!\rightarrow \sim O(p)$

A sensitive algorithm, according to such a definition, would not put forth an output if it had been false. It tasks us to track possible worlds where p is false, but the output is still p. In such instances, the output is not sensitive.

Algorithmic Adherence $=_{df.} p \; \Box\!\!\rightarrow O(p)$

An adherent algorithm would put forth an output if it had been true under minimal and irrelevant changes. It tasks us to track possible worlds where p is true, but the algorithm did not produce p as its output. In such cases, the algorithm is not adherent.

Algorithmic Robustness^{N}, according to Bjerring, Busch, and Munch (2025, 14), addresses the three classes of examples for algorithmic brittleness, namely, adversarial, overfitting, and perspectival. Cases where small and imperceptible variations occur, and the algorithm output immediately changes, are instantiations of adversarial brittleness that imply a failure to satisfy algorithmic adherence. Cases where significant variations occur, and the algorithm output does not change, are instances of overfitting, which violates algorithmic sensitivity. Finally, cases where no change occurs in the core content, but perspectival shifts occur are instantiations of perspectival brittleness that violate algorithmic adherence (Bjerring, Busch, and Munch 2025, 15). Ultimately, Algorithmic Robustness^{N} avoids the brittleness issues that ML models face.

What the foregoing discussion has shown is that thinking about algorithmic robustness in modal terms (whether it be Algorithmic Robustness^{N} or Algorithmic Robustness^{S}) has certain advantages. It potentially offers explanatory powers as to why brittleness in algorithms, although often viewed as mere potential threats, is a genuine problem for trusting ML models. Second, it offers a prospective solution to such brittleness. By alluding to modal conditions, we open the horizons for possibly minimizing such threats by eliminating cases of luck and risks. But are modal conditions necessary for algorithmic robustness? In the remainder of this paper, I analyze this question in detail. More particularly, I analyze the modal conditions that are deemed necessary and sufficient for algorithmic robustness.

CONSTITUTIVE EPISTEMIC LINKS

In the previous section, I discussed how the modal accounts of algorithmic robustness follow the modal analysis of knowledge in epistemology. The modal analysis of knowledge, as shown earlier, aims to eliminate epistemic luck. This is done by proposing conditions that promote or require stability across close possible worlds. That is, in a ceteris paribus possible world, such worlds should ensure that either if the subject believes the proposition through a specific method, then the same proposition is true (safety), if the proposition is false, then the subject would not believe the proposition (sensitivity), or if the proposition is true, then the subject would believe that proposition (adherence). However, evaluating knowledge in modal terms may overlook an important feature of actual epistemic situations. In this section, I will draw from Mark Anthony Dacela's (2019) work on constitutive epistemic links and how they affect knowledge acquisition in epistemic situations. Ultimately, Dacela contends that these links show why modal conditions are not necessary for knowledge.

Before we proceed with the idea of constitutive epistemic links, it is crucial to clarify what epistemic situations are. An epistemic situation is a set of epistemically relevant details in each situation (Dacela 2019, 107). Such epistemically relevant details include, inter alia, "the subject who believes the proposition, his or her belief, the evidence that led him or her to form his or her belief, his or her belief-forming method, the fact (or facts) that make the proposition true (or false), and the proposition the subject accepts (some cases also include defeaters)" (Dacela 2019, 107).

Dacela (2019, 106-107) argues that the elements in such situations are connected by *constitutive epistemic links.* There are two epistemic links that we observe here. First, there is the evidence-belief link. That is, a subject's evidence for a proposition is "necessarily linked to the formation of his or her belief" (Dacela 2019, 107). Second, there is also the fact-proposition link in the sense that "facts that actually obtain are necessarily linked to the truth value of a proposition" (Dacela 2019, 107).

The analysis of the epistemic links can be thought of in set-theoretical terms. That is, we can think of evidence, facts, beliefs, and propositions as sets: Let:

E = set of evidence
F = set of facts or state of affairs
$B(p)$ = the subject's belief in a proposition
p = the proposition

From these variables, an epistemic situation can be thought of as a tuple ES = ($E, F, B(p), p; R_1, R_2$). Now, given the link between the set of evidence and beliefs, we can think of their connection as follows.

R_1: $(E \rightarrow B(p))$

This signifies the link between the evidence for the belief that we hold about a proposition. In addition, we can also think about the connection between a set of facts and the truth of the proposition as follows:

R_2: $(F \models p)$

This means that the set of facts that obtain in the epistemic situation entails the proposition. In other words, the set of facts determines the truth value of the proposition.

Dacela further contends that the set of evidence available to the subject and the set of facts that make a proposition true may be related in different ways. In some cases, E and F may be equivalent (Dacela 2019, 108). When this occurs, the evidence perfectly tracks the truth-making facts. And so, whenever a subject forms a belief based on the evidential set, the proposition will be true. In other cases, E and F may be disjoint sets (Dacela 2019, 108). That is, no member of E belongs to F. In such circumstances, if a subject believes a proposition because of E, the proposition may still be false. Additionally, if F obtains, the proposition may be true, but the subject may not believe it. Finally, the sets may stand in a subset relation in that all members of E belong to F, but not all members of F belong to E, or vice versa (Dacela 2019, 109). In the former case, the belief in a proposition based on E does not guarantee truth. In the latter case, the truth of the proposition does not guarantee that the subject will form a belief. These relations follow from the constitutive links between evidence-belief and fact-truth.

What these relations show us is a fundamental feature of epistemic situations and the modal conditions. The set of evidence on which the belief is based does not always perfectly correspond to the facts. In fact, most of our beliefs are grounded in evidence that does not perfectly correspond to facts. And sometimes, that is enough for us to know. Requiring that the relationship between our evidential basis for belief and the set of facts may be an overly strict condition. If this is correct, preserving the epistemic link between evidence, belief, and fact in the actual epistemic situation is enough for knowledge. We do not have to appeal to a sort of truth-preservation condition across possible worlds. Modal analyses of knowledge, however, require such counterfactual preservation. In doing so, they risk neglecting the constitutive epistemic links that occur in the actual world. What this tells us is that modal conditions impose a very strong requirement for knowledge.

The discussion thus far shows that the constitutive links that obtain in the actual epistemic situation are more relevant to knowledge compared to a truth-preservation requirement across possible worlds. When the relevant evidential and factual relations are in place, a belief may already be justified or well-grounded even if those relations are not preserved in nearby possible worlds. According to Dacela, this shows that the modal conditions are not necessary for knowledge. This raises an important question for the present paper. If modal conditions are not necessary for knowledge, should algorithmic robustness be understood in modal terms? In the next portion of the paper, I will analyze whether algorithmic robustness is properly cashed in modal terms.

IS ALGORITHMIC ROBUSTNESS MODAL?

Since the modal conditions for algorithmic robustness are inspired by the modal analyses of knowledge in epistemology, it is plausible that the criticisms against the latter also apply to the former. Drawing on the discussion of constitutive epistemic links, this section argues that modal accounts of algorithmic robustness overlook an important feature of the actual epistemic situation. I will do this in the following steps. First, I apply Dacela's analysis of constitutive epistemic links to algorithms. I proceed to show that the modal accounts evaluate counterfactual circumstances that are irrelevant to the epistemic links that produce algorithmic outputs in the actual world. Next, I present cases in which an algorithm remains robust despite violating the modal conditions proposed by such modal accounts. I then offer a diagnosis of why this is the case. Such a diagnosis should explain that the modal conditions for robustness neglect the constitutive epistemic links that make an algorithm robust, and in some cases, requiring counterfactual truth preservation can make for an overly stringent condition for robustness. This problem reveals that counterfactual instantiations are unnecessary for robustness.

Let me begin this analysis by applying Dacela's notion of constitutive epistemic links to algorithmic systems. The first thing to observe is that we can identify analogous elements within the epistemic situation of an ML model. However, in the case of ML models, the following will be the crucial elements within an epistemic situation:

I is the set of inputs or the training data that the model relies on.
O is the output behavior of the model.
p is the proposition at hand.
F is the set of facts that obtain in the epistemic situation.

In such a setting, the epistemic situation of an algorithm is a tuple $ES_A = (I, O, p, F; R_1, R_2)$. One upshot of identifying ES_A is that we can understand the epistemic situation of an algorithm through the relationship of the elements.

Evaluating algorithmic performance requires attending to the relation between the elements in an algorithm's epistemic situation within the actual world. The framework helps us do just that. Following Dacela's work, we can investigate analogous constitutive epistemic links in ES_A. First, there is the link between input and output:

R_1: $I(p) \rightarrow O(p)$

That is, given a particular input, an algorithm generates the accurate corresponding output. Second, there is also the fact-truth link:

R_2: $F \models p$

Plainly put, this link states that the state of affairs determines whether the proposition at hand is true. These epistemic links explain how algorithms should be evaluated. The input should give rise to its outputs, and the facts determine whether the outputs are true. We can easily think of examples of ML models operating behind self-driving cars. If the input is "Stop", the model should produce the output or behavior corresponding to the input. But of course, whether the sign actually says stop depends on the relationship between the input and the facts.

If these epistemic links are correct, then an analysis of algorithmic robustness cannot neglect examining whether such links are maintained in the actual world. Now, the problem is that modal accounts evaluate robustness by considering how an algorithm produces its output behavior across nearby possible worlds. As a result, they shift attention away from actual epistemic relations between inputs, outputs, the state of affairs, and the proposition at hand. If that is so, modal accounts of algorithmic robustness risk evaluating features that are not directly relevant to algorithms. If a modal account of algorithmic robustness should succeed with its assessment of algorithms, it should evaluate algorithms according to the preservation of the links as they obtain in the actual epistemic situation.

However, modal accounts of algorithmic robustness do not evaluate such links. The conditions proposed by both Algorithmic Robustness^N and Algorithmic Robustness^S require us to assess whether the relation between outputs and truth is preserved across nearby possible worlds. But this shifts the focus from the actual links (i.e., $I(p) \rightarrow O(p)$ and $F \models p$) to counterfactual scenarios where these relations may vary. The consequence of such conditions is that modal accounts fail to preserve the relationship between the elements that explains how algorithmic outputs are produced and whether they are correct. As such, there is a risk here. Applying modal accounts risks failing to evaluate the actual performance of an algorithm, since the epistemic elements they require us to assess are non-actual. This also raises the question as to whether these modal conditions are actually necessary for algorithmic robustness. In the following, I will examine cases that support this idea. I show that there can be instances where an algorithm is robust, but it does not satisfy the conditions proposed

by the modal accounts of algorithmic robustness (i.e., safety, adherence, and sensitivity).

Let us examine a case where this is so. A good starting point is with Algorithmic Robustness^S. According to Algorithmic Robustness^S, an algorithm is robust just if it is safe. Now, consider the following case.

DoctorAI. A disease classifier named DoctorAI has the capacity to detect developing illnesses. Its training data includes a wide range of modalities for diagnosing illnesses, CT scans, blood biomarkers, and patient history. Before its deployment, it had been trained for more than 20 years on thousands of datasets simulating those from actual hospitals. It also has a function of cross-validating its outputs. As such, DoctorAI is highly reliable, and it has the propensity to produce true outputs. A year ago, DoctorAI was deployed in a town hospital. Over time, it has accumulated outputting over a million accurate diagnoses about certain types of diseases. One time, DoctorAI was able to detect a rare sort of disease based on a patient's input, let us call it *"rare lung disease."* From there, DoctorAI outputs "The patient has *rare lung disease*." The patient indeed has a rare lung disease. However, in nearby possible worlds where another patient without the disease comes in for examination, and a certain lung tissue expands just a tiny undetectable bit, the CT scan would change, and DoctorAI would still output that "the patient has *rare lung disease*." Is DoctorAI robust?

Before answering the question posed in the latter portion of DoctorAI, it is helpful to see the relevant elements in the epistemic situation. Let ES_A represent the following: (1) *I* represents the patient's input, (2) *O* represents the output behavior of the DoctorAI, (3) *p* represents the statement "The patient has rare lung disease", and (4) F represents the state of affairs in the actual world.

With the elements clearly identified, we can evaluate the epistemic situation through the epistemic links. Again, we are looking for the links between input-output (R_1) and fact-proposition (R_2). In the actual case, DoctorAI produces the output *O(p)* based on the input *I(p)*. The corresponding proposition *p* is true by virtue of the relevant state of affairs in the actual world *F*. So, the epistemic links are preserved in the actual epistemic situation, which renders its judgment correct. It is robust because there is a sort of stability between the links. This notion of stability between such links is coherent with Freisleben and Grote's account of robustness that does not appeal to counterfactual truth-preservation.

However, DoctorAI fails to be a robust algorithm because it does not satisfy the safety condition in Algorithmic Robustness^S. Recall that the safety condition for algorithms states that an algorithm is safe just if p would be true in nearby possible worlds if the model produces output *p*. Yet, the latter portion of the case of DoctorAI tells us that *p* would have easily been false if the circumstances had changed. And in those circumstances, the model will continue to produce output *p*. So, should we judge that the algorithm is no longer robust? A lesson that we can take away from this case is that DoctorAI remains stable by virtue of the relevant epistemic elements in the actual world. Even if an algorithm is modally fragile, it does not imply that the algorithm is actually brittle. If this is correct, then we have reason to believe that algorithmic safety is not necessary for algorithmic robustness.

Let us now turn to Algorithmic Robustness^{N}, according to which an algorithm is robust only if it satisfies the sensitivity condition. An algorithm is sensitive just in case, if a proposition were false, the algorithm would not produce that proposition. Keeping this in mind, consider the following case.

GrabBot. GrabBot is an automated self-driving system equipped with a machine learning classifier trained to recognize and respond to road signs. Such an ML model is meant to avoid any sort of perturbations on traffic or road signs, thereby making it almost immune to adversarial attacks. The training data set of GrabBot is pervasive in that it tends to output correct judgments regarding road signs. It seems that we can trust GrabBot in driving matters. However, unbeknownst to us, GrabBot can make mistakes when a star is drawn on a road sign. Apparently, it only fails when a star is drawn. A "Stop" sign is interpreted as a "Speed Limit 45" sign whenever a star is drawn on a road sign. Now, on an occasion where a person is in the back seat of GrabBot, the ML model recognized a stop sign. It had no stars in it, and it proceeded as normal, stopping carefully and ensuring the road was safe before proceeding.

Is the GrabBot algorithmically robust? Similar to the approach in safety, let me first make a couple of remarks regarding the case of GrabBot. The epistemic situation can be fleshed out as follows. First, the input *I* pertains to the Stop sign. Second, the output *O* pertains to the output behavior of GrabBot. This behavior is seen in GrabBot's interpretation of the Stop sign. F, of course, still pertains to the state of affairs in the actual world, while *p* is the proposition whose truth value is determined by F. In our case, let us take the proposition to be p if it reads "The sign says Stop," and q if it reads "The speed limit is 45." We think of p when we talk about the true proposition, and q if we are talking about the false proposition. Since we are looking for situations where the ML model continues to produce a proposition as its output even if the proposition is false, our focus will be on q first.

GrabBot, if we subscribe to Algorithmic Robustness^{N}, is not a robust ML model. This is precisely because the model fails to satisfy the sensitivity condition. Recall that the sensitivity condition requires that if *p* were false, the model would not produce *p* as its output. But in a possible situation where the Stop sign has a small star beside it, the GrabBot would still behave in such a manner that produces the interpretation "Speed Limit 45." So, absent further analysis, we can already disqualify GrabBot as robust since it violates sensitivity. This is only applicable if we subscribe to Algorithmic Robustness^{N}.

But is the judgment that GrabBot is not robust the correct one? Let us look at the constitutive epistemic links in the actual epistemic situation. Notice that the input-output link (R_1) is preserved in the actual epistemic situation. That is, the input *p* is correctly interpreted by the output *p*. Further, the link between the fact and the proposition (R_2) is also preserved. In the actual epistemic situation, it is true that the sign reads as a Stop sign. So, GrabBot not only exhibits reliability in its capacity to produce true outputs, but it also exhibits stability insofar as it properly preserves the link between the epistemically relevant elements in such a situation. Again, we draw a similar conclusion that we had with Algorithmic Robustness^{S}: In cases where an ML model fails to satisfy sensitivity, it can still be robust. As such, sensitivity is not required for algorithmic robustness.

Having now seen reasons why algorithmic sensitivity is unnecessary for a robust algorithm, we already have reasons for rejecting Algorithmic Robustness^{N} insofar as such an account takes the two conditions it proposes as a conjunction. However, let us give the benefit of the doubt to its second condition, namely, algorithmic adherence. Consider the following:

StockAI. A highly sophisticated AI, *StockAI*, is trained to predict quarterly economic growth. During its training phase, developers ensured that its training data set was ideal (i.e., noise-free and clean). The reason behind this is that such developers sought to create an idealized simulation of macroeconomic behavior. When StockAI was deployed in the real world, it was able to generate nearly a thousand accurate propositions about future inflation trends, investment yields, and so on. More recently, it was able to predict that "the inflation will increase by 5% in the next quarter" based on the latest data that it collected. The next quarter arrived, and the inflation did increase by 5%. Again, StockAI arrived at the proposition through predictive analysis borne out of an idealized training data. But since economic trends are extremely vulnerable to variations in the real world, small perturbations or shifts in the actual world data could have altered StockAI's prediction.

Like the analysis of the previous two cases, let me first identify the elements in the epistemic situation of StockAI. The input in this case is the data that the ML model collected.[14] Another element is the output behavior of the ML model with respect to the input. Of course, there is also the proposition *p,* which states that "The inflation will increase by 5% in the next quarter." Finally, the fact that it happened is what determined that proposition's truth value.

Is StockAI robust? The proponents of Algorithmic Robustness^{N} would say no. This is because StockAI violates the adherence condition. Recall that the adherence condition states that if a proposition were true, then the ML model would produce the proposition as its output. But in another possible epistemic situation where the proposition "The inflation will increase by 5% in the next quarter" is true, it would still be possible for the output not to predict this. Such a possibility of failure in prediction is precisely because economic trends are extremely vulnerable to variations in the actual world. This just means that the interpretation of the ML model is always going to be fallible, especially in situations like this one.

But I do not think that we can conclusively say that StockAI is not robust. After all, it preserves the relevant epistemic links in the actual epistemic situation. For instance, the input that contains the data is correctly tied to the output. The StockAI is not robust if it bases its output on elements other than the relevant inputs. In addition, the link between the fact and the proposition that the StockAI produced is also preserved. It correctly produced a proposition that corresponds to the state of affairs. The preservation of epistemic links tells us that the ML model was able to do its job correctly and robustly. Even if there is the possibility of failure, thereby leading to the violation of the algorithmic adherence condition, StockAI can still be classified as a robust algorithm. This analysis of StockAI echoes the conclusion derived from the previous cases. That is, the modal conditions are not necessary for algorithmic robustness.

DoctorAI, GrabBot, and StockAI all have something in common. In all of these cases, the algorithm was able to preserve the constitutive epistemic links in the

epistemic situation (recall the status of the tuple ES_A in each of the cases). That is to say, it was able to ensure a successful procedure in terms of producing a proposition that is justified. Alongside this, each model also violated the requirements of the modal accounts. These illustrative cases show us two things. First, our understanding of algorithmic robustness can be supplemented by the preservation of epistemic links, as it is compatible with the non-modal accounts of algorithmic robustness. In other words, the explanatory capacities of non-modal accounts of algorithmic robustness may be strengthened by bringing constitutive epistemic links into the picture. Second, preserving epistemic links do not require a sort of modal analysis. And if algorithmic robustness can be explained through epistemic links, satisfying the modal conditions is not necessary for a robust ML model.

In addition, this analysis unearths a deeper problem for any modal account of algorithmic robustness. As stated by Dacela (2019, 117) in talking about the modal analysis of knowledge, this is more than a *criterion problem*. That is, we are not merely having problems with identifying similar or close possible worlds. The problem is that shifting our attention from actual to possible epistemic situations leads us to overlook what matters in assessing relevant epistemic situations and evaluating algorithmic robustness.

In sum, the foregoing discussions support the conclusion that modal conditions are not necessary for algorithmic robustness. Again, this applies to the existing modal accounts of algorithmic robustness, namely, Algorithmic Robustness^S and Algorithmic Robustness^N. In the next section, I will anticipate possible objections to this conclusion. More particularly, I examine how the proponents of the different modal accounts of algorithmic robustness might respond to the findings above.

OBJECTIONS AND REPLIES

What I have established so far is that we have reasons to believe that modal conditions are not necessary for algorithmic robustness. This section will investigate three objections to this claim. The first objection will be about how modal accounts are consistent with fallible but robust algorithms. The second objection will focus on how maintaining reliability and preserving constitutive epistemic links do not eliminate luck. The third objection highlights how a non-modal account of algorithmic robustness is compatible with luck. Immediately after each objection, I will offer my responses.

On the Consistency of Modal Accounts with Robust Yet Fallible Algorithms

It may be argued that counterfactual accounts are consistent with robust yet fallible algorithms. A closure condition may be implemented to ensure that a modally robust algorithm is compatible with certain failures.[15] In addition, proponents of the counterfactual accounts of algorithmic robustness can think of strategies like restricting the relevant class of nearby worlds, weakening modal requirements, or adopting hybrid approaches to avoid charges of requiring infallibility.[16]

Before I reply to this objection, I first need to clarify why there is an intuition that the analysis above might lead to the idea that the counterfactual accounts of algorithmic robustness are incompatible with robust yet fallible algorithms. In Dacela's (2019, 116-117) discussion of *constitutive epistemic links,* he showed that fallible beliefs will always be unsafe. This is because the safety conditions tend to impose a stringent requirement that only allows infallible beliefs to count as knowledge. However, it may be argued that this does not necessarily apply to algorithmic robustness.

I do not dispute such an idea. It may very well be the case that we can think of a modal account of algorithmic robustness that is weak enough to admit the possibility of errors, but strong enough to eliminate a huge chunk of epistemically lucky situations. We can take Hornischer's (2026, 21-23) proposal of the probabilistic robustness grounded on an analysis using modal logic. Such an account, to my mind, counts as a counterfactual account that allows for errors. As stated by Hornischer, it is strong enough to refrain from classifying every ML model as robust, and it is weak enough to avoid triviality, allowing for potential mistakes.

However, even if modal accounts allow fallibility, they still impose counterfactual requirements that may be unnecessary. The fundamental insight being advanced in this work is that the counterfactual conditions imposed by the modal accounts are not necessary for algorithmic robustness. This is because they sidestep analyzing what matters in an algorithm's epistemic situation. An accurate assessment of the robustness of algorithms should be able to tackle these elements in an epistemic situation, and the modal conditions do not do that. So, any sort of modal account for algorithmic robustness, even if they allow fallibility, would neglect actual epistemic links if its priority is to examine counterfactual scenarios.

Reliability and Epistemic Links Cannot Rule Out Epistemic Luck

Proponents of the modal accounts of algorithmic robustness may argue that reliability alone is not enough to eliminate epistemic luck. After all, the main reason why robustness is thought to be a requirement on top of reliability is that an ML model can reliably produce true outputs while being brittle. Thinking of the preservation of the constitutive epistemic links in an algorithm's epistemic situations is like thinking about reliability and expecting the input and output to function well. But such a response is still vulnerable to epistemic luck.

In response to this objection, it must be made clear that I am not proposing a complete alternative account of algorithmic robustness. I am only identifying a necessary feature that previous analysis has missed, namely, the preservation of constitutive epistemic links in an algorithm's epistemic situation. Constitutive epistemic links provide the relevant connections. So, adding that layer of analysis is not the same as expecting the algorithm to produce true beliefs while waiting for its outputs to be aligned with its inputs. A takeaway here is that robustness should not be about eliminating modal luck, unless we want an overtly stringent condition. What matters is the presence of the relevant connection between inputs, outputs, and facts in the actual epistemic situation.

On Lucky Robust Algorithms

The proponent of the modal accounts may push the second objection further by asking if an epistemically lucky algorithm can still be robust. After all, an algorithm that preserves epistemic links and is reliable can still be lucky in producing true outputs.

Following Dacela's response to this objection in the case of the modal conditions for knowledge, the findings of this paper show that algorithmic robustness may be compatible with the epistemic luck that the proponents of the counterfactual accounts had in mind. Since they analyze epistemic luck in terms of counterfactual scenarios, it may be that they think about luck through the non-modal preservation between the model's output and the fact that determines whether the output is true. But these lucky instances have nothing to do with what happened in actual epistemic situations. If luck is cashed out in these terms, it is unproblematic to say that a robust algorithm is compatible with luck. After all, algorithms are fallible. Imposing a plausibly infallible requirement makes it overtly stringent.

CONCLUSION

So far, this paper has been an appraisal of the recently developed counterfactual or modal accounts of algorithmic robustness. Given the deployment of ML models in high-stakes environments, experts worried whether such models are worthy of trust. Algorithmic robustness, cashed out in modal terms, has been a proposed solution to this worry. In the preceding portions of this paper, I began by rehearsing this sort of algorithmic robustness. I then offered cases and reasons why modal conditions are unnecessary for algorithmic robustness. I have shown that there is a fundamental flaw in requiring modal conditions for the subject matter.

This inquiry aims to show how the study of epistemology can contribute to the minimal step of thinking about ways of engineering and possibly deploying ML models in different environments. It may be the case that we are imposing unnecessary conditions upon ML models in their performance. It is worthwhile ensuring that the requirements that we are looking for in an AI system are *bona fide* relevant necessities. Perhaps identifying the constitutive epistemic links and how we can preserve them in the actual world deserves our attention before we set conditions that task us to go beyond our world of actuality into the world of possibilities.

NOTES

1. Throughout this paper, I will follow Kathleen Creel (2020) in thinking of transparency and opacity as two sides of the same coin.
2. In this paper, I will use the terms "modal" and "counterfactual" to refer to matters related to possible worlds.
3. As such, I subscribe to a modal interpretation of luck here. It needs to be stated that there are multiple accounts of luck. Two of the leading accounts are (1) proximity and (2) proportion accounts. I will no longer be discussing these in detail because they

are off-topic relative to this paper. For a recent discussion on this matter, see Samuel Kahn's work (2025).

4. See Obermeyer and Emanuel (2016) for the applications of machine learning algorithms in clinical medicine.

5. See Bernhammou et al. (2020) and Rajpurkar et al. (2017) for some of the developments on this matter.

6. See Zheng et al. (2020) and Zheng et al. (2022) for discussions on the AI economist.

7. In this paper, I limit my discussion to the sort of trust that is epistemic in nature. That is, we are trusting ML models to give us accurate information.

8. In the literature, reliability is cashed out in two different ways; (1) frequency, and (2) propensity. Frequency-based approaches investigate reliability in terms of how frequent a true output or belief is generated as opposed to false ones. This takes its roots from the earliest works on reliabilism, especially influenced by Goldman's (1979) views. Propensity-based approaches investigate the tendency of a process to produce true outputs or beliefs. One leading proponent of this approach is William Alston (1995). Recent work in the epistemology of AI prominently applies the propensity-based approach in analyzing algorithm reliability. The works of Durán (2025a), Durán (2025b), Durán and Formanek (2018), and Durán and Jongsma (2021) are examples of recent discussions on this subject matter. This paper's points and arguments apply to both approaches.

9. See Edmund Gettier's (1963) work titled "Is Justified True Belief Knowledge?".

10. I take inspiration from Bjerring, Busch, and Munch (2025, 22) in this interpretation of Vandenburgh's account of algorithmic safety.

11. Vandenburgh's safety generally follows the line of thought of epistemic safety theorists. According to Sosa (1999, 142), a belief is safe if and only if the belief in P obtains only if it were so that p. Pritchard (2014, 156) follows this definition with his own, saying that a belief is safe if it could not have easily been false.

12. In counterfactual semantics, we can read this as follows: "If it were the case that O(p), then it would have been the case that p."

13. I say "improved" because they viewed both the safety and causal accounts as overtly narrow in being inclusive of genuine cases of robust algorithms.

14. This case is a bit different from the examples above because of two things. First, the input is in a disjoint set relation with the set of facts. That is, the set of data in the input does not contain the set of facts. Second, the ML model operates for future predictions. Nonetheless, I do not think that such a difference affects the analysis in any shape or form.

15. I am grateful to an anonymous reviewer for this insight.

16. I thank an anonymous referee for raising this concern.

REFERENCES

Alston, William. 1995. "How to think about reliability." *Philosophical Topics* 23: 1-9.

Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, OPaisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. "A human centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1-12. https://doi.org/10.1145/3313831.337671

Bernhammou, Yassir, Achchab Boujemâa, Francisco Herrera, and Siham Tabik. 2020. "Break His based breast cancer automatic diagnosis using deep learning: Taxonomy, survey, and insights." *Neurocomputing* 375: 9-24. https://doi.org/10.1016/j.neucom.2019.09.044

Bjerring, Jens Christian, Jacob Busch, and Lauritz Aastrup Munch. 2025. "A Counterfactual Account of Algorithmic Robustness." *Minds and Machines* 35: 34. https://doi.org/10.1007/s11023-025-09734-z

Creel, Kathleen A. 2020. "Transparency in complex computational systems." *Philosophy of Science* 87: 568-589. https://doi.org/10.1086/709729

Dacela, Mark Anthony. 2019. "Are modal conditions necessary for knowledge?" *Kritike* 13: 101-121.

DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. 2021. "AI for radiographic COVID-19 detection selects shortcuts over signal." *Nature Machine Intelligence* 3: 610-619.

Durán, Juan M. 2025a. "Beyond transparency: computational reliabilism as an externalist epistemology of algorithms." In *Philosophy of Science for Machine Learning,* 55-79. https://doi.org/10.1007/978-3-032-03083-2_4

Durán, Juan M. 2025b. "In defense of reliabilist epistemology of algorithms." *European Journal for Philosophy of Science* 15: 37. https://doi.org /10.1007/s13194-025-00664-2

Durán, Juan M., and Nico Formanek. 2018. "Grounds for trust: Essential epistemic opacity and computational reliabilism." *Minds and Machines* 28: 645-666. https://doi.org/10.1007/s11023-018-9481-6

Durán, Juan M., and Krystyna R. Jongsma. 2021. "Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI." *Journal of Medical Ethics* 47 (5): 329–335.

Engstrom, Logan, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. "Exploring the Landscape of Spatial Robustness." In *Proceedings of the International Conference on Machine Learning*, 1802–1811.

Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, and Dawn Song. 2018. "Robust Physical-World Attacks on Deep Learning Visual Classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1625–1634.

Freisleben, Till, and Thomas Grote. 2023. "Beyond Generalization: A Theory of Robustness in Machine Learning." *Synthese* 202: 1–28. https://doi.org/10.1007/s11229-023-04334-9

Gettier, Edmund. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23 (6): 121-123. https://doi.org/10.1093/analys/23.6.121.

Goldman, Alvin. 1979. "What is justified belief?" In *Justification and Knowledge,* edited by G. Pappas, 1–23. Dordrecht: Reidel.

Grote, Thomas. 2024. "Machine learning in healthcare and the methodological priority of epistemology over ethics." *Inquiry: A Journal of Medical Care Organization, Provision and Financing* 68: 1-30. https://doi.org/10.1080/0020174X.2024.2312207

Grote, Thomas, Konstantin Genin, and Emily Sullivan. 2024 "Reliability in machine learning." *Philosophy Compass*. https://doi.org/10.1111/phc3.12974

Hornischer, Levin. 2026. "Robustness and trustworthiness in AI: a no-go result from formal epistemology." *Synthese* 207: 22. https://doi.org/10.1007/s11229-025-05272-4

Kahn, Samuel. 2025. "Proximity Beats Proportions in Modal Accounts of Luck." *International Journal of Philosophical Studies.*

Nozick, Robert. 1981. *Philosophical Explanations.* Cambridge: Cambridge University Press.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine." *New England Journal of Medicine* 375 (13): 1216.

Pritchard, Duncan. 2014. "Knowledge Cannot Be Lucky." In *Contemporary Debates in Epistemology,* 2nd ed., edited by Matthias Steup, John Turri, and Ernest Sosa, 152–164. Chichester: Wiley-Blackwell.

Ranjan, Anurag, J. Janal, Andreas Geiger, and Michael J. Black. 2019. "Attacking Optical Flow." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2404–2413.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, et al. 2017. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning." *arXiv preprint* arXiv:1711.05225.

Sosa, Ernest. 1999. "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13: 141–153. https://www.jstor.org/stable/2676100

Vandenburgh, Jacob. 2023. "Machine Learning and Knowledge: Why Robustness Matters." *arXiv preprint* arXiv:2310.19819.

Zech, John R., Marcus A. Badgeley, Ming Liu, Andrew B. Costa, Joseph J. Titano, and Eric K. Oermann. 2018. "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study." *PLoS Medicine* 15 (11): e1002683.

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. 2022. "The AI Economist: Taxation Policy Design via Two-Level Deep Multiagent Reinforcement Learning." *Science Advances* 8 (18). https://doi.org/10.1126/sciadv.abk2607

Zheng, Stephan, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Max Gruesbeck, David C. Parkes, and Richard Socher. 2020. "The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies." *arXiv preprint* arXiv:2004.13332.

Acknowledgements

I thank the editors of this special issue and the two anonymous referees for their constructive comments, which greatly improved this manuscript. Much of my interest in the subject matter of this paper was ignited by the discussions in the AI and Epistemology Workshop held at De La Salle University, Manila. I thank Nikolaj J.L.L. Pedersen, Mark Anthony Dacela, Rosallia Domingo, Jeremiah Joven Joaquin, Jade Espuelas, John Ian Boongaling, Peter Graham, Enrique Benjamin R. Fernando III, Hazel T. Biana, Joyce Fungo, Masashi Kasaki, the organizers of the event, and the other philosophers who contributed to the discussion. I am also grateful to Gavin Nigel Chuacuco, Joshua Jose Ocon, Apollo Gueco, Lumberto Mendoza, Paul Benedict Cano, Paul Adrian Galvez, and Richmond Zachary Ona for their comments on earlier drafts. I also express my heartfelt gratitude to Mark Anthony Dacela for introducing me to his objections to the necessity of the modal conditions for knowledge. Finally, special thanks are due to Jermain Kirsten Apostol for her support throughout the development of the paper.