Synthese (2026) 207:22
https://doi.org/10.1007/s11229-025-05272-4

ORIGINAL RESEARCH

Robustness and trustworthiness in AI: a no-go result from formal epistemology

Levin Hornischer[1]

Received: 7 December 2024 / Accepted: 11 September 2025

Abstract
A major issue for the trustworthiness of modern AI-models is their lack of robustness. A notorious example is that putting a small sticker on a stop sign can cause AI-models to classify it as a speed limit sign. This is not just an engineering challenge, but also a philosophical one: we need to better understand the concepts of robustness and trustworthiness. Here, we contribute to this using methods from (formal) epistemology and prove a no-go result: No matter how these concepts are understood exactly, they cannot have four prima facie desirable properties without trivializing. To do so, we describe a modal logic to reason about the robustness of an AI-model, and then we prove that the four properties imply triviality via a novel interpretation of Fitch's lemma. We then discuss the consequences for explicating a viable notion of robustness for AI. A broader theme of the paper is to build bridges between AI and epistemology: Not only does epistemology provide novel methods for AI, but modern AI also provides many new questions and perspectives for epistemology.

Keywords Artificial intelligence · Robustness · Trustworthiness · Epistemology · Fitch's paradox · Modal logic

1 Introduction

Despite the tremendous success of artificial intelligence (AI), a major issue for its trustworthiness is the lack of robustness. Paradigmatic examples are so-called adversarial attacks: a minuscule change to the input of the AI-model—like some noise or a sticker—can yield a completely different output (Goodfellow et al., 2014; Eykholt

✉ Levin Hornischer
Levin.Hornischer@lrz.uni-muenchen.de

[1] Munich Center for Mathematical Philosophy, LMU Munich, Ludwigstr. 31, 80539 Munich, Germany

Published online: 05 January 2026

et al., 2018). For a principled response to the robustness problem, we not only need engineering solutions but also a better conceptual understanding of robustness and trustworthiness.

In this paper, we contribute to this conceptual understanding using methods from epistemology and logic. After all, these methods excel at understanding intuitive concepts precisely: famously, this has been done for belief and knowledge, and here we transfer these methods to robustness and trustworthiness. An important upshot of this kind of analysis is to identify the limits of these concepts: how much of what we would want from them is actually achievable. Knowing such upper bounds of robustness and trustworthiness is crucial in guiding both our expectations and engineering.

How are robustness and trustworthiness understood in AI? Let us start with robustness. If not left intuitive, it usually refers to various forms of immunity to so-called *distribution shifts*. They occur when an AI-model is deployed—as common in practice—on input data sampled from a different distribution than the data on which it was trained: e.g., when the model was trained on patient data from a few hospitals but is deployed at many hospitals (Koh et al., 2021). Recently, a unified conceptual analysis of robustness—that includes distribution shifts—was provided by Freiesleben and Grote (2023, p. 1) as "the relative stability of a robustness target [e.g., the deployment performance of the AI-model] with respect to specific interventions [e.g., shifting] on a modifier [e.g., the deployment distribution]" (cf. Braiek & Khomh, 2025). Examples include the following.

1. The already mentioned *adversarial attacks* can be seen as synthetic distribution shifts: the AI-model is deployed on artificially shifted inputs (Taori et al., 2020).[1]
2. But the AI-model should also be robust to *natural distribution shifts*: those arising in the real world, e.g., when the images come from different surroundings and lighting conditions than those seen during training (Taori et al., 2020).
3. *Performativity* is also a form of distribution shift: The prediction of an AI-model—say, whether a person gets a loan—supports decisions of the bank which leads some applicants to manipulate their features towards more favorable outcomes. This changes the distribution to which the AI-model is applied. Ideally, the AI-model is robust to its induced distribution shifts (Perdomo et al., 2020).
4. *Shortcut learning* occurs, e.g., if an AI-model seemingly classifies cows perfectly after training, but does so for the wrong reasons and hence fails on pictures of cows outside their usual environment. The AI-model has learned the unintended shortcut that 'grass' predicts 'cow', which fails outside the training distribution (Geirhos et al., 2020).

The concept of trustworthiness is even more tricky. An expert group set up by the European Commission requires trustworthy AI to be lawful, ethical, and robust (High-Level Expert Group on Artificial Intelligence, 2019). (For an overview of trustworthiness in AI, see Huang et al., 2020.) Here, we actually do not need more

[1] For an overview on adversarial attacks, see Serban et al. (2020) and for a formalization Dreossi et al. (2019).

detail on the notion of trustworthiness beyond the fact that it plausibly implies robustness. The reason is the following.

We are interested in the limitations of *any* reasonable precise notion of robustness and trustworthiness. So we will not presuppose a specific definition of these concepts. Instead, we will talk about them in a general language—that leaves open their exact nature—and are interested in what we can derive about robustness and trustworthiness just from some general principles (cf. axiomatic method). This is similar to how we can meaningfully talk about the concept of belief and knowledge without having to define precisely what they are. Using the methods of epistemology, we can—despite this generality—still derive our no-go result.

More precisely, the summary of the paper is as follows. In Sect. 2, we illustrate the issue of robustness and trustworthiness with a standard guiding example, namely classifying images of handwritten digits—which, as we show, easily generalizes to other AI-models, including large language models. From that, we develop, in Sect. 3, a simple formal logic to reason about the robustness and trustworthiness of AI-models. In Sect. 4, we identify four *prima facie* desirable principles about robustness and trustworthiness. To already mention these principles, they informally read as follows.

- *Factivity*: If the AI-model shows behavior φ robustly, it, in particular, shows behavior φ.
- *Robustness*: Every trustworthy behavior is robust.
- *Countermodels*: For every trustworthy behavior that is not shown on every input (i.e., is not trivial), there is some input on which the AI-model robustly does not show this behavior (so that input is a good counterexample, hence the name).
- *Moore-closure*: For every trustworthy behavior φ, it is also trustworthy that the AI-model cannot be easily tricked into showing behavior φ if it currently does not (this has to do with the Moore sentence in epistemology, hence the name).

In Sect. 5, we prove that these principles imply triviality, via a novel interpretation of Fitch's lemma. In Sect. 6, we discuss the consequences of the no-go result for the goal of explicating a viable notion of robustness. We distinguish between (1) a uniform notion of robustness (i.e., there is a fixed range of robustness that is guaranteed whenever the AI-model is applied) and (2) a non-uniform notion (i.e., every application of the AI-model has some range of robustness), as well as (3) a probabilistic notion of robustness. The first satisfies all principles and hence trivializes, while the last two do not trivialize but hence need to give up some principle (indeed, the Countermodels principle). We argue that the first is too strong and the second too weak a notion of robustness—with the third being a promising start to an intermediary notion. Section 7 explores the generality of the no-go result: we sketch how it also applies to the notion of explainability in AI as well as stability in philosophy. Section 8 highlights further open questions suggested by the analogies between AI and epistemology, which are summarized in Fig. 1. In particular, it asks how the impossibility result—like many others—may plant seeds for positive results, e.g., by moving from a qualitative to a quantitative notion of robustness. We conclude in Sect. 9.

The broader theme of the paper—following recent 'calls to action' (Buckner, 2019; 2023; Grote et al., 2024)—is to further reconnect philosophy and modern AI.

Fig. 1 Analogies between AI and epistemology that will be developed in the paper

AI	*Epistemology*
Robustness	Safety condition for knowledge
Adversarial attacks	Gettier cases
Training	Internal justification
?	External justification
Triviality	Modal collapse
AI-model can be tricked	Moore-sentence
Triviality of uniform robustness	Anti-luminosity
Non-uniform robustness	Margin-for-error principle
No-go result	Fitch's lemma
Explainability	Provability

Using the example of robustness and trustworthiness, we show how this connection is fruitful in both directions: how well-studied epistemological and logical concepts can be applied to pressing issues in AI, and how this also sheds new light on these concepts via these novel applications.

2 A case study

As a simple but standard example, we consider an AI-model that classifies input images according to which digit they depict. At the end of this section, we show that our discussion actually applies much more generally also to large language models (like ChatGPT) and models for automated decision-making.

2.1 Classification behavior of the model

We want to talk about the classification behavior of this AI-model, in order to assess its robustness. So we want to make statements of the form 'on input s, the AI-model shows behavior φ'. For example, given this picture s as input, the AI-model classifies it as depicting digit 2. Thus, we are broadly in the setting of formal verification for AI (Seshia et al., 2022; Huang et al., 2017; Albarghouthi, 2021): we want to see if the AI-model (the system) in a state s (the environment) has property φ (the specification).

In suggestive logical notation, we write $M, s \vDash \varphi$ for 'on input s, the AI-model M shows behavior φ'. We now want to describe the relation $M, s \vDash \varphi$ more precisely. Before we start, though, note that the statement '$M, s \vDash \varphi$' is one that we, as external observers, make about the behavior of the AI-model; it is not an internal statement of how the AI-model processes the input. In other words, $M, s \vDash \varphi$ concerns the externally observable behavior of the AI-model and does not say that the AI-model internally 'believes' that φ (after all, such anthropomorphizing language is very difficult to make precise).

To describe $M, s \vDash \varphi$, we recursively consider the structure of φ, starting with the basic—or atomic—behavior of the AI-model, i.e., its classification behavior. We use the atomic sentences $p_0, \dots, p_9$ corresponding to the digits $0, \dots, 9$, so we can describe the atomic behavior as:

- $M, s \vDash p_i$ iff the AI-model M classifies input image s as depicting digit i.[2]

We can then naturally extend this to complex behavior φ: like $\neg p_2 \wedge p_3$. So $M, s \vDash \neg p_2 \wedge p_3$ says that the AI-model does not classify input s to depict digit 2, but it does classify it to depict digit 3. Thus, we define

- $M, s \vDash \neg\varphi$ iff $M, s \nvDash \varphi$.[3]
- $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$.

Again note the externality: $M, s \vDash \neg\varphi$ means that the AI-model does not show behavior φ. It does not mean that the AI-model classifies the input as non-φ (whatever that means).

Importantly, we also want to talk about the robustness of the AI-model. For now, let us take robustness as resilience to adversarial attacks (item 1 in our list). We write $M, s \vDash \Box p_i$ to say that the AI-model M *robustly* classifies input s as depicting digit i, i.e., it cannot be adversarially attacked on this classification. This means that on all similar enough inputs s', the AI-model still classifies input s' as depicting digit i. Thus, we define

- $M, s \vDash \Box\varphi$ iff for all s' that are relevantly similar to s, we have $M, s' \models \varphi$.

Of course, 'relevantly similar' is not a precise notion and may depend on the context (the deployment situation, the safety-criticality, the type of adversarial attack, the model, etc.). However, a straightforward way—which is used in the verification literature—is to specify 'relevantly similar' as having a distance with respect to some fixed metric d on the inputs of less than some fixed threshold $\epsilon > 0$. For concreteness, the input space X is a subset of $\mathbb{R}^n$ where n is the number of pixels of the images, so a vector $s \in \mathbb{R}^n$ describes the color value (or grayscale) of each pixel in the image, and X consists of those vectors that represent realistic images (rather than random pixel values). A standard metric then is given by the L_2-norm, and we can set, say, $\epsilon := 0.1$. Thus, 'relevantly similar' is a binary relation on inputs, and we get a *Kripke semantics* for $\Box$. (We will discuss issues and alternatives of this approach in Sect. 6.)

Finally, it will also be convenient to say that a property holds for all inputs:

- $M, s \models \boxbox\ \varphi$ iff for all s', we have $M, s' \vDash \varphi$.

Importantly, the relation $M, s \vDash \varphi$ connects the model-internals (the micro-level) with the model-externals (the macro-level). The left side of '$\vDash$' describes the AI-model at the micro-level: it refers to the internal parameters of the model that are not visible to the end-user (e.g., the model architecture and the weight values, which, so far, we left implicit, but later also make the explicit) as well as the pixel-values of the input image. The right side of '$\vDash$' describes the AI-model at the macro-level: the behavior of the AI-model that can externally be observed by the end user, described

[2] Throughout, 'iff' abbreviates 'if and only if'.

[3] Here '$M, s \nvDash \varphi$' is shorthand for 'it is not the case that $M, s \vDash \varphi$'.

in a human-understandable language (e.g., the classification of the input in categories that are meaningful to us). Obviously, we hence want to better understand this relation: its logic, i.e., the laws governing this link between the model-internals and the model-externals, and how this connects to robustness and trustworthiness.

2.2 Robustness and trustworthiness of the AI-model

Now we have a simple language to talk about the behavior of the AI-model. With it, we have a way to express *local* robustness: $M, s \vDash \Box\varphi$ says that, for input s, the AI-model M robustly shows behavior φ. But for deploying the AI-model in practice, we want a more *global* robustness: that the model still is locally robust also on new inputs that it has not seen during training.[4] For example, to trust our AI-model in its classifications of the digit 2, we want that whenever it classifies a realistic pixel image $s \in X$ as a digit 2 (i.e., $M, s \vDash p_2$), then it does so robustly (i.e., $M, s \vDash \Box p_2$). (So we do not require robustness on any arbitrary pixel image $s \in \mathbb{R}^n$, which may be too strong, but only on the realistic ones that we consider as allowed input, i.e., the $s \in X \subseteq \mathbb{R}^n$.) This means that for every allowed input $s \in X$, we have $M, s \vDash p_2 \to \Box p_2$. In logical terminology, a sentence ψ that is true at every state of a given model is said to be *valid* in that model—denoted $M \vDash \psi$.[5] Generally speaking, then, we say a behavior φ is *(globally) robust* for the AI-model M if $M \vDash \varphi \to \Box\varphi$.

Moving to trustworthiness, when is a behavior φ of the AI-model trustworthy? For example, for $\varphi = p_2$, we want that if the AI-model classifies an input s as depicting the digit 2, we can trust that the input really depicts a 2. As mentioned, however, it is very difficult both to theoretically characterize and practically identify the set T of trustworthy behavior. In general, for φ to be trustworthy this means that, if the AI-model shows behavior φ, we can trust that behavior, but what that amounts to exactly is difficult to specify. In the case of p_2, this means trusting that the input really depicts a 2, but for more complex φ, this may be harder to say. Hence, as often done in mathematics, we will treat T as a variable (for a set of sentences) and only make some plausible assumptions about it: for example, that trustworthiness implies robustness (i.e., if $\varphi \in T$, then φ is robust). We will formulate three further such plausible assumptions (Sect. 4), and then our no-go result shows that they entail triviality. Thus, we still gain knowledge about the trustworthy behaviors T, without explicitly defining it.

2.3 A special case of the impossibility result

To already get a flavor for such an impossibility, we present a simple but far-reaching special case of our no-go result.

[4] See, e.g., Ruan et al. (2019) for a discussion of local vs. global robustness.

[5] It is valid (simpliciter)—written $\vDash \psi$—if it is valid in every model of the considered class of models. But here the model-relative notion is enough.

It adds one more assumption, namely that the input space $X \subseteq \mathbb{R}^n$ of the AI-model—i.e., the set of realistic images—is path-connected.[6] Intuitively, this means that, for any two realistic images s and s', one can continuously change the pixel values of s so as to arrive at the pixel values of s' while only producing realistic images in the process. This has some intuitive plausibility, but we will not further elaborate on it, since our main no-go result will do away with this assumption.

Theorem 1 *Assume the input space X of the AI-model M is a path-connected subset of $\mathbb{R}^n$. Let d be a metric on X, let $\epsilon > 0$, and assume that 'relevantly similar' in the interpretation of $\Box$ is explicated as $d(s, s') < \epsilon$. If trustworthiness implies robustness, then only trivial behavior is trustworthy: i.e., if $\varphi \in T$, then either φ is valid in M or $\neg\varphi$ is valid in M.*[7]

Consequently, if we buy the plausible assumptions for our running example, then no interesting behavior—i.e., behavior that is present on some inputs but absent in others—can be trustworthy! In particular, if we only consider robustness—i.e., stipulate $T := \{\varphi : M \vDash \varphi \to \Box\varphi\}$—and define (local) robustness, as usual, via a metric and a threshold, then (global) robustness trivializes.

The response to the impossibility could be that (a) despite the initial plausibility of the assumptions, we have to give up some and (b) we also need to refine the explication of robustness (not simply use a metric and a threshold). We will discuss this in Sect. 6, after discussing the main no-go result. In particular, as an alternative to the above 'uniform' robustness that uses a fixed robustness range ϵ, we develop (1) a 'non-uniform' robustness where ϵ can depend on the input and (2) a probabilistic version that 'smoothens' the sharp cut-off provided by the robustness range.

As mentioned, Theorem 1 will be a consequence of our main result, but, in Appendix A.1, we still give a direct proof. It is based on the following instructive idea. If the trustworthy behavior φ were not trivial, there is an input s where the model shows behavior φ and an input s' where it does not. By connectedness, there is a sequence of inputs $s = s_1, s_2, \dots, s_{n-1}, s_n = s'$ such that adjacent inputs are at most ϵ-far apart. Since the model shows behavior φ on s_1 and trustworthy behavior is robust, the model shows behavior φ robustly on s_1. Hence it also shows this behavior on the relevantly similar input s_2. But now we can repeat the reasoning and eventually conclude that also on s_n the model shows behavior φ, contrary to assumption.

This proof is reminiscent of the anti-luminosity argument by Williamson (2000, ch. 4) in epistemology. It argues that many mental states, like feeling cold, are not *luminous* to us, i.e., we may be in that state without being in a position to know that we are. This argument considers the mental states $s_1, \dots, s_n$ that we go through, in millisecond intervals, from feeling cold at dawn to feeling hot at noon. If feeling cold is luminous, then, in s_1, we can know that we feel cold, so, for this knowledge to be

[6] A subset X of $\mathbb{R}^n$ is path-connected if, for any $x, y \in X$, there is a path in X from x to y (i.e., a continuous function $p : [0, 1] \to X$ with $p(0) = x$ and $p(1) = y$).

[7] In fact, the theorem holds for any way of explicating relative similarity as a binary relation R (beyond the relation $d(s, s') < \epsilon$) to interpret $\Box$, as long as R is connected (i.e., for all s and s' there are $s = s_0, s_1, \dots, s_n = s'$ with $s_i R s_{i+1}$).

reliably based, we should also feel cold in the very similar state s_2. Again, we continue this reasoning to get that also in $s_n$we feel cold, contrary to assumption.

2.4 Generalizing to other AI-models

Finally, let us see how the framework developed so far also captures other forms of AI-models and robustness—not just adversarial attacks to a classifier.

Concerning other model architectures, let us consider two examples. First, if we are dealing with tabular data and automated decision-making, our inputs no longer are pixel images but feature vectors, and our AI-model outputs predictions based on the inputted features. Again, these outputs are described by atomic sentences: for example, $M, s \vDash p_{\mathrm{loan}}$ iff the AI-model M outputs that the person whose features are s should get a loan. Second, consider a large language model (LLM) predicting next words. Then s is the prompt and the outputs again are described by atomic sentences: for example, $M, s \vDash p_{\mathrm{shining}}$ iff the language model M continues input $s =$‘The sun is’ with the next word ‘shining’. As before, we can describe more complex behavior using the connectives $\neg, \wedge, \Box, \neg$. Though, in describing robustness, ‘relevantly similar’ now means, e.g., being synonymous (for inputs to the LLM) or being identical outside of protected attributes like gender or disability status (for automated-decision making). (We come back to LLMs in the probabilistic explication of robustness in Sect. 6.3.)

Concerning other notions of robustness, we could, for example, capture natural distribution shifts (item 2 in our list) by interpreting ‘relevantly similar’ as: input s' is a shifted version of input s (e.g., a picture of the same scene but with different lighting conditions). However, we get many more modeling options by noting that the states s need not be just inputs. They can include all kinds of choices in the machine learning pipeline—which is used by Freiesleben & Grote (2023) to define a general account of robustness. So, for example, a state s could, in addition to the input, consist of (1) a description of the task conceptualization, (2) the dataset in raw and in prepared form, (3) the choice of model architecture, hyperparameters, and training algorithm, and (4) the data distribution on which the model is deployed. Thus, we could capture performativity 3, for example, by saying s is relevantly similar to s' iff the deployment distribution of s' is the one obtained from that of s after using the AI-model for a certain amount of time. As an intermediate example, one can consider states as pairs $s = (w, x)$ consisting of a set of weights w of the AI-model and a given input x. Then robustness could require that the model not only shows the same behavior on similar inputs but also on similar weights. Thus, we would require that we get the same behavior had we trained the model sampling the training data in a different order.

In sum, we obtain a very general framework to describe robustness and trustworthiness of AI-models. We now turn to analyzing this framework more formally.

3 A simple logic of robustness

Generalizing from the preceding examples, we have a language built from atomic sentences using the operators $\neg, \wedge, \Box, \boxed{\scriptstyle\Box}$. Intuitively, the sentences φ describe externally observable behavior in a human-understandable language. The sentences are true or false at the internal states s of the AI-model M—where a state at least includes the input to the AI-model but can also include, e.g., the weights of the AI-model. We write $M, s \vDash \varphi$ iff, in state s, the model M shows behavior φ. So the relation $\vDash$ links the internal states to the external behavior. We write $M \vDash \varphi$ (and say φ is valid in *M*) if behavior φ is shown in every state, i.e., φ is a behavioral law of *M*. However, there is no agreed-upon formal explication of $\vDash$. In particular, we do not yet have a formal semantics for the robustness operator $\Box$.

Thus, we are in a similar situation that modal logicians were in before the advent of Kripke semantics (and other semantics, like the topological semantics). They did not have a formal semantics for the necessity operator $\Box$, which describes when a sentence φ is necessarily true (at a possible world s). Instead, they looked at the *logic*—or the governing *laws*—of sentences involving the necessity operator. They axiomatically described the laws via a derivability relation $\vdash$, where $\vdash \varphi$ means that φ is logically provable—or 'derivable'—from the chosen axioms.[8] The relation $\vdash \varphi$ is meant to track the still to be formally defined semantic validity $\vDash \varphi$ (truth at every possible world in every model). Although purely syntactic, $\vdash$ still captures much of the (unknown) semantics and it still can be philosophically assessed by discussing the plausibility of its axioms.

Hence we now also play the logicians' trick: We do not define the semantic relation $M, s \vDash \varphi$ explicitly, but rather make assumptions about the behavioral laws. So a derivability relation $\vdash$ then describes those behavioral laws that we can derive from these assumptions. Just like in modal logic, we also distinguish two types of assumptions: the logical axioms (which we describe below) and the 'contentful' assumptions specifically about robustness and trustworthiness (which we describe in Sect. 4).[9] Thus, every AI-model with a choice of formal semantics for $\Box$ (e.g., Kripke semantics) gives rise to a derivability relation $\vdash$, which describes the behavioral laws. The logical axioms governing the derivability relation describe a minimal logic of robustness, and the contentful assumptions strengthen this logic. Even if the behavioral laws do not specify whether a particular behavior φ is shown in a particular state of the AI-model, they still describe many important aspects of the AI-model. In particular, to say that the AI-model is robust concerning its classification of the digit 2 is to say that $\vdash p_2 \to \Box p_2$.

In the remainder of this section, we formally define our notion of language and the logical axioms on the behavioral laws, i.e., the derivability relation. (For a reference on modal logic, see Blackburn et al., 2001.)

[8] Typically, $\vdash$ is just a unary relation because $\varphi \vdash \psi$ (i.e., ψ is derivable from φ) is equivalent to $\vdash \varphi \to \psi$.

[9] In modal logic, the logical axioms include (at least for normal modal logics), e.g., the K-axiom $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$; while the 'contentful' axioms, that are specific to different interpretation of $\Box$, are, e.g., the T-axiom $\Box\varphi \to \varphi$, which holds when $\Box$ is interpreted as knowledge but not if it is interpreted as belief.

As already indicated, the *language* is defined as follows. We choose a nonempty (finite or countably infinite) set of atomic sentences At. We recursively build complex sentences: if φ and ψ are sentences, so are $\neg\varphi$, $\varphi \wedge \psi$, $\Box\varphi$, $\boxdot\, \varphi$. We write $\mathcal{L}$ for the set of all sentences. We define the usual abbreviations (for some $p \in \mathsf{At}$):

$$\begin{array}{rlrl} \varphi \vee \psi & := \neg(\neg\varphi \wedge \neg\psi) & \bot & := p \wedge \neg p \\ \varphi \rightarrow \psi & := \neg\varphi \vee \psi & \Diamond\varphi & := \neg\Box\neg\varphi \\ \varphi \leftrightarrow \psi & := (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi) & \Diamonddot\,\varphi & := \neg\boxdot\ \neg\varphi. \end{array}$$

We use the variables $p, q, r, \ldots$ to range over atomic sentences and $\varphi, \psi, \chi, \ldots$ to range over sentences. For our purposes, a propositional language is enough: We do not need further syntactic structure on the atomic properties like first-order quantification. The language $\mathcal{L}$ is a well-known bimodal language (Shehtman, 1999).

A *derivability relation* (or *logic*) $\vdash$ is a set of sentences (i.e., a subset of $\mathcal{L}$) satisfying the axioms A1–A4 below. Here we write $\vdash \varphi$ iff sentence φ is in the set $\vdash$.

A1 If φ is a truth-functional tautology,[10] then $\vdash \varphi$ (*classical logic*)
If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$, then $\vdash \psi$ (*modus ponens*)
A2 $\vdash \Box(\varphi \wedge \psi) \rightarrow (\Box\varphi \wedge \Box\psi)$ (*distributivity for* $\Box$)
If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \Box\varphi \leftrightarrow \Box\psi$ (*congruence for* $\Box$)
A3 If $\vdash \varphi$, then $\vdash \boxdot\,\varphi$ (*necessitation for* $\boxdot$)
If $\vdash \varphi \rightarrow \psi$, then $\vdash \Diamonddot\,\varphi \rightarrow \Diamonddot\,\psi$ (*semi-congruence for* $\Diamonddot$)
A4 $\nvdash \bot$ (*consistency*)

In words, axiom A1 says that we follow the scientific standard of basing our reasoning on classical logic.[11] For example, if φ is a tautological behavior, it is shown at every state in every model. (After deriving the no-go result, we discuss in Sect. 8—as a way out—the idea of weakening the underlying logic to some non-classical logic; but, for now, we go with the standard choice of logic.) Axiom A2 says that $\Box$ distributes over $\wedge$, and that provably equivalent sentences can be substituted inside a $\Box$-context. Axiom A3 says that $\boxdot$ satisfies necessitation (if φ is provable, so is $\boxdot\,\varphi$) and that also the 'implication-version' of the substitution of provably equivalents holds. Finally, axiom A4 says that the logic is consistent. Note that a typical assumption for modal operators is that they are normal, i.e., validate the K-axiom ($\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$) and necessitation. For $\Box$, this implies the axiom A2, and for $\boxdot$, this implies axiom A3.

[10] More precisely, there is a sentence χ with atomic sentences $p_1, \ldots, p_n$ and there are sentences $\psi_1, \ldots, \psi_n$ such that φ is the result of replacing each occurrence of p_i with ψ_i (for all $i = 1, \ldots, n$) and χ only contains Boolean connectives (i.e., $\neg$, $\wedge$) and is a classical tautology (i.e., true under any classical valuation).

[11] Here, we do not assume uniform substitution because, if $\vdash$ describes the sentences that are valid in a model with a fixed interpretation of the atoms, we might have, e.g., $\vdash p \rightarrow \Box p$ but not $\vdash q \rightarrow \Box q$, so we cannot substitute q for p (cf. Williamson, 1999, p. 129).

4 Principles about robustness and trustworthiness

To recall, given any AI-model (or a class thereof) and interpretation of robustness, we denote the behavioral laws φ, i.e., those behaviors that are always shown, as $\vdash \varphi$. In the preceding section, we collected some logical axioms about $\vdash$. In this section, we collect some further ‘contentful’ principles about robustness and trustworthiness. We write T for the set of behaviors on which the AI-model can be trusted, but since we do not know exactly which behaviors belong to T, we only formulate assumptions about T that seem plausible regardless of what its elements are.

4.1 Factivity

The first principle does not yet involve T but simply requires that robustness is factive: If the AI-model shows behavior φ robustly, then it in particular shows behavior φ.

P1 $\forall \varphi \in \mathcal{L} : \ \vdash \Box\varphi \to \varphi$ (*Factivity*)

On the intuitive reading of $\Box$ as ‘for all relevantly similar states’, this axiom is highly plausible: whatever ‘relevantly similar’ means exactly, a state arguably should be relevantly similar to itself, so if it satisfies $\Box\varphi$ it also satisfies φ.[12]

So what could be objections to Factivity? In Sect. 6.3, we will consider one explication of robustness which fails Factivity. This explication reads $\Box\varphi$ as ‘with high probability also similar states show behavior φ’. This explication fails Factivity since high probability does not imply truth. However, even then, we will argue that the following weaker form of Factivity still is very plausible:

P1′ $\forall \varphi \in T : \ \vdash \Box\Diamond\varphi \to \Diamond\varphi$ and $\vdash \Box\varphi \to \Diamond\varphi$ (*Weak-factivity*)

Factivity indeed implies Weak-factivity: The quantification is now only over sentences in *T*; the first conjunct is Factivity but restricted to sentences with a leading diamond; and the second conjunct is implied by Factivity, since Factivity implies $\vdash (\Box\varphi \to \varphi) \wedge (\varphi \to \Diamond\varphi)$.

Our impossibility result will only need Weak-factivity and not full Factivity. So the upshot is that, even if, for most explications of robustness, Factivity is valid, our impossibility also applies to explications that only satisfy Weak-factivity, like the probabilistic explication.

4.2 Robustness

We have already encountered the next principle: that trustworthiness should imply robustness—as the expert group requires (High-Level Expert Group on Artificial Intelligence, 2019).

P2 $\forall \varphi \in T : \ \vdash \varphi \to \Box\varphi$ (*Robustness*)

[12] Indeed, the axiom $\Box\varphi \to \varphi$ is known as the T-axiom in modal logic and is validated by Kripke frames whose accessibility relation is reflexive.

In words: If the AI-model shows trustworthy behavior on an input (or, more generally, in a state), then it shows this behavior robustly, i.e., also on relevantly similar input (or states). In particular, to trust the AI-model in its behavior, it should not be adversarially attackable. Interestingly, related ideas are discussed in epistemology. To to put this principle into philosophical perspective, we sketch this analogy in the remainder of this subsection. See Hornischer (2021, ch. 8), Vandenburgh (2023), and Grote et al. (2024) for such connections between knowledge and trustworthiness.[13]

The basic idea of the analogy is that behavior is to trustworthy behavior what belief is to knowledge. This may well be an analogy and not a similarity (Gentner, 1983): i.e., behavior may be dissimilar to belief (if one does not anthropomorphize), but the relations between behavior and trustworthy behavior on the one hand might be similar to the relations between belief and knowledge on the other hand. If the AI-model classifies a given image as depicting the digit 2, then whatever it takes to render this behavior trustworthy is similar to whatever it takes to turn a belief into knowledge. Let us consider two examples of how this idea may help to transfer insights from epistemology to AI, specifically concerning robustness. The focus here is on illustrating such a transfer, not on working it out in detail.

First, the analogy suggests that principle P2 corresponds to the so-called safety condition for knowledge. This condition identifies a feature of knowledge that mere belief need not have. It says that knowledge requires safety in the sense that, if we know that p is the case, we still would not be wrong in similar cases.[14] This is illustrated by Russell's famous stopped clock (Russell, 1948, 170f.): We look at a clock with the intention of coming to know what time it is. Incidentally, at this very moment, the clock shows the right time, but, unbeknownst to us, it actually has stopped exactly 24 hours ago. Intuitively, we would not consider the true belief about the current time that we obtained in this way to be knowledge. And, indeed, in the very similar case where we had looked at the clock just one minute earlier, we would have been wrong, so the safety condition is violated.[15] There is much discussion on how to formulate the safety condition precisely, but it usually runs along the following lines: If a knows p based on a method m in situation s then, in all situations s' similar to s, if a believes p in s' based on m, then p is true at s'. So, if safety needs to be added to belief to get to knowledge, how is this relation transferred—via the analogy—to behavior of AI-models? Plausibly, robustness needs to be added to behavior (among others) to get to trustworthy behavior—just like principle P2 requires. Surely this needs a more thorough discussion, but it suggests a potentially fruitful application of the extensive philosophical discussion of safety conditions to trustworthiness in AI.

Second, the analogy suggests that trustworthiness is as elusive a concept as knowledge. Just as it is notoriously hard (if at all possible) to specify what else is required to turn (justified and true) belief into knowledge, it plausibly is equally hard to specify what else is required to turn (well-trained and correct) behavior of an AI-model into a trustwor-

[13] Whether such modal conditions for knowledge are compatible with best practices in statistics is discussed, e.g., by Mayo-Wilson (2018) and Fletcher & Mayo-Wilson (2024), which we will pick up again in Sect. 8.

[14] See, e.g., Williamson (2000), Sosa (1999) and, for an overview, Ichikawa & Steup (2018) and Rabinowitz (2024). For other 'stability-requiring' notions in epistemology, see Rott (2004).

[15] Closely related to the safety condition is the idea that knowledge should not involve epistemic luck: in the example, we were just lucky that our belief turned out to be correct (Pritchard, 2005).

thy behavior. In other words, the analogy casts doubt on the prospect of a fully satisfying conceptual analysis of trustworthiness into crisp and operationalizable sufficient and necessary conditions. To make this point, the analogical counterpart to the infamous Gettier examples (which obstruct an analysis of knowledge) may be adversarial attacks. Consider a deep neural network that correctly classifies a camera input as containing a stop sign but that can be adversarially attacked. Then this classification behavior not only is correct (the analog of being true) but it also is well-trained (the analog of justified), since the classification is produced after a long training process. Yet, the classification behavior of the AI-model is not trustworthy, since it can be adversarially attacked. So the AI-model has the analog of justified true belief but not the analog of knowledge—just like a Gettier case. Again, this needs further discussion. For example, the above kind of justification is 'internalistic', and an attempt to explain away Gettier examples is to demand that justification should instead be externalistic (Ichikawa & Steup, 2018). It would be interesting to consider what the analogical counterpart is for AI.

4.3 Countermodels

The next principle is a minimal falsifiability requirement:

P3 $\forall \varphi \in T : \ \vdash \neg\varphi \to \Diamond\Box\neg\varphi$ (*Countermodels*)

In words: For any trustworthy behavior φ, if, on some input (or state), the AI-model does not show behavior φ, then there also should be some input (or state) on which it robustly does not show behavior φ. This input (or state) thus is a good *countermodel* to showing behavior φ—hence the name.

For example, it may be that the AI-model does not classify an input s as depicting the digit 7, because that input is on the decision boundary between a 1 and a 7. By slightly enhancing the characteristics of a 7, we get a similar input s' which the AI-model classifies as 7. So it is *not* the case that the AI-model robustly does not show behavior p_7. However, all that the principle requires is that if p_7 is trustworthy, then there is *some* input on which the AI-model robustly does not show behavior p_7. This input may be, for example, a clearly written 2 or even a completely black image. The AI-model will robustly not classify them as a 7—otherwise p_7 arguably was not trustworthy in the first place.

The bigger the state space, the easier it is to satisfy principle P3. For example, let us consider states as pairs $s = (M, x)$ where M is an AI-model that can take x as input. Then, to satisfy principle P3, we only need to find some model that robustly does not show behavior φ on some input. Conversely, if the conclusion of the no-go result is that this principle must be violated, then it does not just say that the present model cannot robustly show some behavior, but, in fact, no model can (cf., e.g., Bastounis et al., 2024).

4.4 Moore-closure

The last principle requires a certain closure property of the trustworthy behaviors.

P4 $\forall \varphi \in T \exists \psi \in T : \ \vdash \psi \leftrightarrow \varphi \vee \Box\neg\varphi$ (*Moore-closure*)

In words: If φ is a trustworthy behavior, so is $\varphi \vee \Box\neg\varphi$, up to logical equivalence. Before discussing its plausibility, let us explain the name.

Moore sentences are known from epistemology (e.g., van Benthem, 2004). They are sentences of the form 'p, but I do not know it' or, as a formula, $p \wedge \neg\Box p$. They are important because they can be true but one can never know them: otherwise one would know p and (know that) one does not know p. (They play an important role in Fitch's paradox, as we will see in Sect. 5). Instead of knowledge, we interpret $\Box$ here as robustness, but the formulas are unchanged: Given a sentence φ, we define the *Moore sentence* and the *dual Moore sentence* of φ, respectively, as

$$\mathsf{M}\varphi := \varphi \wedge \neg\Box\varphi \qquad\qquad \mathsf{W}\varphi := \varphi \vee \Box\neg\varphi,$$

where W is dual to M in the usual sense that $\mathsf{W}\varphi$ is equivalent to $\neg\mathsf{M}\neg\varphi$.

Here we only consider W. Intuitively, $M, s \vDash \mathsf{W}\varphi$ says that, on input (or state) s, either the AI-model M already shows behavior φ or else it robustly does not show behavior φ. In other words, $M, s \vDash \mathsf{W}\varphi$ means that, on input (or state) s, the AI-model cannot be *φ-tricked*: it cannot be tricked into showing behavior φ (if it currently does not show this behavior) by providing a very similar input (or state). Thus, $\mathsf{W}\varphi$ describes behavior of the AI-model that we naturally are interested in when assessing the trustworthiness of the model: after all, if the model can be tricked into classifying an input as, say, depicting digit 2, it intuitively did not have high (objective) certainty when it did not classify the input as depicting digit 2.[16]

Now, the principle demands: if behavior φ is trustworthy, so is $\mathsf{W}\varphi$, i.e., it is trustworthy that the AI-model cannot be φ-tricked. This seems plausible: As just discussed, for behavior to be trustworthy, we should have some certainty, reason, or evidence that the AI-model cannot easily be tricked into it, and this justification hence also renders $\mathsf{W}\varphi$ trustworthy. We also show formally that this principle is satisfied on natural explications of robustness (Sect. 6).

5 The no-go result

Now we can state the no-go result (Sect. 5.1) and discuss the idea of the proof, which utilizes Fitch's lemma (Sect. 5.2). The formal proof is in Appendix A.2.

5.1 Statement of the result

Let us summarize the framework that we now have in place to state the no-go result. We have a language $\mathcal{L}$ to describe the externally observable behavior of AI-models, including their robustness. This behavior is realized in the internal states of the AI-models. The derivability relation $\vdash$ describes the behavioral laws: those behaviors that are shown in every state in the considered AI-models. The set T contains those sentence describing trustworthy behavior. We neither know $\vdash$ nor T exactly, so we

[16] For an overview of uncertainty quantification in deep learning, see, e.g., Abdar et al. (2021), Hüllermeier & Waegeman (2021) and for a connection with robustness Gustafsson et al. (2020).

have collected four prima facie plausible principles about them, in addition to the logical axioms A1–A4.

The no-go result now provides new insights about robustness and trustworthiness, as encoded in $\vdash$ and T: namely, these four principles imply triviality.

Theorem 2 *For any derivability relation $\vdash$ (describing the behavioral laws) and set of sentences T (describing the trustworthy behavior), the principles P1–P4 imply triviality, i.e., $\forall \varphi \in T : \; \vdash \Diamond\varphi \leftrightarrow \Box\varphi$. This still holds when replacing P1 by P1′.*

Thus, whatever the behavioral laws are ($\vdash$) and whatever behavior is trustworthy (T), at least one of the prima facie plausible principles has to go, unless the trustworthy behavior is trivial. Here triviality means that, for any state of a considered AI-model, as soon as it shows the trustworthy behavior φ on *some* similar state, it actually shows this behavior on *all* similar states. Intuitively, then, there are no *decision boundaries* for trustworthy behavior φ, i.e., there are no states in the considered AI-models with close-by φ-states on one side and close-by $\neg\varphi$-states on the other side. It is in this sense that trustworthy behavior is trivial. Thus, Theorem 2 is a *modal collapse* argument: trustworthy sentences are modally trivial (cf. the modal collapse argument of Quine, 1960, 191f.).

In Sect. 6, we will discuss the consequences of the no-go result and whether triviality should be accepted or some of the principles should be given up. But first we discuss, in the next subsection, the idea of the proof.

5.2 Proof idea: reinterpretation of Fitch's paradox

The proof relies on a reinterpretation of Fitch's paradox in terms of robustness. Here we (1) recap Fitch's paradox, (2) state our reinterpretation, and (3) describe the idea of the proof. The formal proof is in Sect. A.2.

(1). Fitch's paradox is also known as the Church–Fitch paradox or the paradox of knowability. Nowadays, it is stated as follows: If all truths can be known ($\forall p : p \rightarrow \Diamond Kp$), then all truths are already known ($\forall p : p \rightarrow Kp$). The contrapositive implication was first published by Fitch (1963) acknowledging an anonymous referee who, much later, was discovered to be Alonzo Church. (See Brogaard & Salerno, 2019 for a brief history.) Even though Fitch worked in the context of value concepts, the implication is mostly interpreted in the above form as an objection to verificationalism, i.e., the view that all truths are knowable (Brogaard & Salerno, 2019).

In more precise words, the *formal* implication

$$\forall p : p \rightarrow \Diamond Kp \;\; \Rightarrow \;\; \forall p : p \rightarrow Kp,$$

which is also called Fitch's lemma, can be established with fairly minimal proof-theoretic assumptions on the modal operators $\Diamond$ and K. It gains philosophical meaning by *interpreting* the quantifier as ranging over all declarative statements and the modal operators as describing metaphysical possibility and knowledge, respectively.

(2). Here we wish to provide another interpretation in terms of robustness: First, we make explicit the set T of sentences that we quantify over. In the original version, this was the set of all (declarative) sentences. Thus, we can account for the worry that not all sentences are subject to verificationalism but only those of a certain logical form (cf. Dummett, 2001). Second, and most importantly, we suggest a different interpretation of the modal operators. We (re-) interpret the metaphysical possibility operator $\Diamond$ as the global possibility operator $\Diamond\mkern-11mu\diamond$ and we (re-) interpret the knowledge operator K as the robustness operator $\Box$.

For example, while, in the original interpretation, the formula $\forall p : p \to \Diamond Kp$ expressed the knowability principle, in our reinterpretation, the formula $\forall p : p \to \Diamond\mkern-11mu\diamond \Box p$ expresses the idea that all behavior can be robust. Here, we will not assume this principle in full generality but only in the restricted form of principle P3: i.e., we do not assume it for all p, but only for negations of trustworthy behaviors, i.e., sentences of the form $\neg\varphi$ with $\varphi \in T$.

(3). Now, the idea of the proof is as follows. Given a trustworthy behavior $\varphi \in T$, we need to show $\vdash \Diamond\varphi \leftrightarrow \Box\varphi$. By the Robustness principle, we have $\vdash \varphi \to \Box\varphi$. By Factivity, we have $\vdash \Box\varphi \to \varphi$. And an easy argument shows that Factivity also implies $\varphi \to \Diamond\varphi$ (apply the contrapositive version of Factivity to $\neg\varphi$). So it remains to show $\vdash \Diamond\varphi \to \varphi$. Toward a contradiction, assume this fails. So there must be some state s with $s \vDash \Diamond\varphi$ but $s \nvDash \varphi$, i.e., $s \models \neg\Box\neg\varphi \wedge \neg\varphi$. Equivalently, $s \vDash \neg\psi$, where $\psi := \varphi \vee \Box\neg\varphi$. But by the Moore-closure principle, we have $\psi \in T$ (up to equivalence). So the Countermodels principle applies and we get $\vdash \neg\psi \to \Diamond\mkern-11mu\diamond \Box\neg\psi$. Since $s \vDash \neg\psi$, we also have $s \vDash \Diamond\mkern-11mu\diamond \Box\neg\psi$. We now argue that this cannot be. Indeed, it implies that there is a (potentially different) state s' with $s' \vDash \Box\neg\psi$. Recall that $\neg\psi$ is equivalent to $\neg\varphi \wedge \neg\Box\neg\varphi$. Further, robustly showing behavior $\varphi_1 \wedge \varphi_2$ implies robustly showing behavior φ_1 and robustly showing behavior φ_2. Hence s' robustly shows behavior $\neg\varphi$ and s' robustly shows behavior $\neg\Box\neg\varphi$. By Factivity, s' also shows behavior $\neg\Box\neg\varphi$. Hence, writing $\chi := \Box\neg\varphi$, we have $s' \vDash \chi$ and $s' \vDash \neg\chi$, which is the desired contradiction. As mentioned, the fully formal proof is in Appendix A.2, which also shows why in this proof we actually only need Weak-factivity.

6 Consequences of the no-go result

The no-go result provides a very general limitation for robustness and its interplay with trustworthiness: no matter how we understand these concepts, we can never satisfy all the principles from Sect. 4 without trivializing. As with any no-go result, this prompts the question: do we need to bite the bullet of triviality or should we give up one of the initially plausible assumptions? In particular, what does this mean for the goal of formalizing the concepts of robustness and trustworthiness, so that they can be verified for AI-models?

We start this discussion in Sect. 6.1 by considering the concrete explication of robustness via a metric and a threshold, as in the case study from Sect. 2. We find that all principles are satisfied (thus corroborating the principles), so we need to accept the triviality in this case (and we also discuss the severity of the triviality). This is exactly the special case that we saw as Theorem 1.

Looking for ways to avoid the triviality, we are lead to an important conceptual distinction in the explication of robustness, which we explore in Sect. 6.2. The just mentioned explication is *uniform* in the sense that it has a fixed robustness range ϵ that should work for all states of the AI-model. But we can also consider *non-uniform* explications where we only require for each state some robustness range. This distinction between uniform and non-uniform robustness matches exactly the two most common semantics for modal logic: Kripke semantics and topological semantics, respectively. We find that, on this weaker non-uniform robustness, we can avoid triviality. Hence, by the no-go result, we also have to give up at least one principle: indeed, we show that the Countermodels principle is violated.

In Sect. 6.3, we explore a third, probabilistic explication of robustness, which is a 'smoothened' version of uniform robustness and which also can avoid triviality.

We discuss the three explications of robustness in Sect. 6.4.

6.1 Uniform robustness (Kripke semantics)

As seen in Sect. 2, the most straightforward formalization of robustness is via a metric d on the input space X of the AI-model M and a threshold ϵ. Thus,

- $M, s \vDash p$ iff the AI-model shows basic behavior p on input s (classifying it as having label p, next word prediction p, etc.)
- $M, s \vDash \neg\varphi$ iff $M, s \nvDash \varphi$
- $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$
- $M, s \vDash \Box\varphi$ iff, for all s', if $d(s, s') < \epsilon$, then $M, s' \vDash \varphi$
- $M, s \vDash \boxbox\, \varphi$ iff, for all s', $M, s' \vDash \varphi$.

With this explication, we have a concrete setting to discuss the consequences of the no-go result. For the purpose of this discussion, let us stipulate trustworthy behavior to be robust behavior, i.e., $T := \{\varphi : M \vDash \varphi \to \Box\varphi\}$. In other words, for now, we put to the side other conditions for trustworthiness than robustness. With these choices, we can ask which principles are satisfied.

It is not difficult to show, that, in this setting, not only are all logical axioms satisfied (defining $\vdash \varphi$ by $M \vDash \varphi$) but also all the principles P1–P4—the details are in Appendix A.3. This corroborates the principles: they come out true on the most straightforward explication of robustness. However, the flip side is that the no-go result implies triviality: for all $\varphi \in T$, we have $M \vDash \Diamond\varphi \leftrightarrow \Box\varphi$.

Here is what triviality means in this setting. We can partition the input space X into equivalence classes, where two inputs s and s' are considered equivalent iff there is a sequence of inputs $s_1, \ldots, s_n \in X$ such that $s = s_1$, $s' = s_n$ and $d(s_i, s_{i+1}) < \epsilon$ for all $1 \leq i < n$. So we cluster together inputs that are connected by similarity. Triviality then means that, for trustworthy behavior $\varphi \in T$ and any given cluster, either the AI-model shows behavior φ on all inputs of the cluster or the AI-model does not show behavior φ on all inputs of the cluster. In particular, this entails Theorem 1 as a special case: if we additionally assume the input space to be path-connected, then the entire input space is one single cluster, so trustworthy behavior either is shown on every input or on no input.

But is triviality really a bad thing? Arguably yes: The severity comes from the fact that the triviality applies to *all* AI-systems. Sure, there might be *some* AI-systems where the triviality is not an issue and, in fact, welcome. For example, take a binary classifier that is well-trained on an input space (aka data manifold) in which the positive class and the negative class are nicely separated with respect to the metric that we use to explicate robustness. So the positive class and the negative class form two clusters in the input space and the AI-model has a trivial classification behavior in each cluster. The problem is that there also are binary classification tasks where such a clean separation is impossible. In the input space, there might be many data points that have, within ϵ-distance, both positively and negatively labeled data points. In these cases, the classification behavior can never be robust, because, by the no-go result, it would have to be trivial on a cluster.

This prompts us to look for ways to avoid triviality, so we should weaken our notion of robustness—as it then becomes easier to obtain nontrivial robust behavior. The problem with the preceding explication of robustness that has lead to triviality is that we insist on a robustness range of some fixed $\epsilon > 0$, even when we are close to the decision boundary. So an obvious weakening is to conceive of robustness as more gradual: when we are close to the decision boundary, we expect a smaller robustness range than when we are far away. Thus, while there might not be a *uniform* robustness range ϵ that works for all inputs s ($\exists\forall$), we can at least *non-uniformly* require that for every input s there is some robustness range ϵ ($\forall\exists$). This move also occurs in *margin for error principles* in epistemology and vagueness (see, e.g., Williamson, 1994, sec. 8.3 or Égré, 2015, sec. 2).

Consequently, we only fix a metric d upfront and change the clause for robustness to:

(∗) $M, s \vDash \Box\varphi$ iff there is $\epsilon > 0$ such that, for all s', if $d(s, s') < \epsilon$, then $M, s' \vDash \varphi$.

Thus, on the uniform explication of robustness, 'relevantly similar' is an absolute and all-or-nothing matter: either two states s and s' are similar, i.e., $d(s, s') < \epsilon$, or they are not. But on the non-uniform explication, it is a relative and gradual matter: two states s and s' are similar to degree ϵ if $d(s, s') < \epsilon$; and non-uniform robustness requires robustness for some degree of similarity.[17]

6.2 Non-uniform robustness (topological semantics)

We saw an important conceptual distinction for robustness: uniform robustness requires one fixed robustness range $\epsilon > 0$ that works for all states of the AI-model,

[17] Since the uniform robustness parameter ϵ acts as a quantitative measure of the trustworthiness of the AI-model, we also may want a weakened substitute. Following Ruan et al. (2019), we can take this to be the expected value, across the different inputs s, for the maximal robustness range ϵ that we can choose for the input s.

while non-uniform robustness only requires for each state some robustness range $\epsilon > 0$. Uniform robustness satisfies all principles but hence, by the no-go result, trivializes. We will now see that the weaker non-uniform robustness indeed can deliver nontrivial behavior that still is robust in that weaker, non-uniform sense.

Before we get there, we first observe that non-uniform robustness has a crucial advantage over uniform robustness: not only does it quantify away the threshold ϵ, it also is independent of the metric (provided the metric comes from a norm, as described shortly). This is important because, in practice, various metrics are used (e.g., L_0, L_1, L_2, or Wasserstein), and it generally is regarded as an open problem which is the right notion of distance (see, e.g., Freiesleben, 2022, sec. 5.2, for discussion). Typically, the metrics on the input space $X \subseteq \mathbb{R}^n$ are given by a norm, and it is a basic mathematical fact that all these norms are equivalent in the sense of determining the same topology, namely the usual Euclidean topology on $\mathbb{R}^n$. Hence clause $(*)$ for non-uniform robustness can equivalently be expressed as:

$(**)$ $M, s \vDash \Box\varphi$ iff there is an open set $U \subseteq X$ with $s \in U$ such that, for all $s' \in U$, we have $M, s' \vDash \varphi$.

This is exactly the topological semantics for modal logic (e.g. van Benthem & Bezhanishvili, 2007).

Now here is how we can avoid triviality on this explication. Let us again stipulate, for the purpose of this discussion, trustworthy behavior to be robust behavior, i.e., $T := \{\varphi : M \vDash \varphi \to \Box\varphi\}$. Now consider an open set $U \subseteq X$ that is not closed (e.g., an open interval $(a, b) \subseteq \mathbb{R}$ in the case of dimension $n{=}1$). We can regard U as a behavior p by defining: $M, s \vDash p$ iff $s \in U$. Then p is a nontrivial trustworthy behavior, i.e., $p \in T$ and $M \nvDash \Diamond p \leftrightarrow \Box p$ (see Appendix A.4 for a proof). The common topological spaces have plenty of open set that are not closed, so we get plenty of robust behaviors that are not trivial, as desired.

Since non-uniform robustness is built on the topological semantics, it satisfies all the logical axioms, so our no-go results applies. Since we avoid triviality, at least one of the principles P1–P4 must be violated. Which one? It is not difficult to show that Factivity, Robustness, and Moore-closure are satisfied (again, see Appendix A.4). So the Countermodels principle must fail: there must be $\psi \in T$ such that $M \nvDash \neg\psi \to \Diamond\Box\neg\psi$. In fact, ψ can be chosen as $\mathsf{W}\varphi$ for any nontrivial behavior $\varphi \in T$. Recall that $\mathsf{W}\varphi$ says that the AI-model cannot be φ-tricked. So we have (proof in Appendix A.4):

> For every nontrivial robust behavior φ (open and not closed), the AI-model can be φ-tricked ($\mathsf{W}\varphi$ is false on some inputs) but for generic inputs it cannot ($\mathsf{W}\varphi$ is generically true, i.e., an open and dense set).

Indeed, on the non-uniform explication of robustness, there is no reason to expect the Countermodels principle: Violating $\mathsf{W}\varphi$ amounts to being on the decision boundary for φ, which (literally) is the topological boundary of the set of states satisfying φ (proof in Appendix A.4). If φ is nontrivial robust (open and not closed), this boundary is nonempty, so $\mathsf{W}\varphi$ can be violated. However, $\mathsf{W}\varphi$ can never be robustly violated, since it would require a nonempty open subset of the boundary of φ, which is topologically impossible since φ is open.[18]

In sum, on the non-uniform explication of robustness, we can have nontrivial robust behavior and we should give up the Countermodel principle. In that sense, non-uniform robustness fares better than the too strong uniform robustness, but on the other hand one might argue that non-uniform robustness is too weak a notion of robustness, which we discuss for the remained of this subsection.

The argument goes as follows. We said that norm-based metrics are usually used to define distances in connection with robustness, and all these metrics induce the Euclidean topology on the input space. So this is the natural topology on the input space when discussing non-uniform robustness. But this renders too many AI-models robust: Consider a binary classifier M that takes pixel images $s \in \mathbb{R}^n$ and outputs 'yes' or 'no' depending on whether the image depicts a stop sign. Say this classifier has a typical neural network architecture; so it realizes a function $N : \mathbb{R}^n \to [0, 1]$ which has been trained to output the probability $N(s)$ with which it takes the image s to depict a stop sign, and the classifier outputs 'yes' once this probability exceeds, say 0.9. So $[\![p_{\text{yes}}]\!] = \{s \in \mathbb{R}^n : M, s \vDash p_{\text{yes}}\}$ is the set of pictures the classifier takes to depict a stop sign. Since typical neural networks are continuous functions with respect to the Euclidean topology, $[\![p_{\text{yes}}]\!] = N^{-1}((0.9, 1])$ is open. But this now means that, according to non-uniform robustness, any such classifier automatically is robust in classifying stop signs, regardless of how it was trained! So non-uniform robustness does not track the intuitive notion of robustness according to which many such classifiers are not robust.

What to make of this argument? Initially, one might take issue with the fact that the robustness range ϵ can get arbitrarily small. After all, a robustness range below computer precision is practically useless. So one might want to require a minimal robustness range. However, then one is back with uniform robustness and its problems resurface. As another option, then, one might consider other topologies than the Euclidean one, giving up the idea that the topology comes from some norm-based metric. After all, the appropriate topology that captures similarity between images according to human cognition may be different from the 'low-level' Euclidean topology on the pixel values. Presumably, not every open set of the latter topology is open in the former topology, so the preceding argument that non-uniform robustness is too weak does not go through anymore.[19] Yet further options are to use other formalisms to explicate robustness. Among the broadly topological options are neighborhood semantics (Pacuit, 2017) or the d-semantics (van Benthem & Bezhanishvili, 2007, sec. 3.1). One could also consider the robustness analysis of Moggi et al. (2018, thm. A.2) using domain theory (which is used in Zhou et al., 2023).

Yet another option is to consider a conceptually different explication of robustness in terms of probability, which we explore next.

[18] However, this changes if we add the requirement that trustworthy behavior should not only be open but in fact regular open (a set is regular open if it is the interior of its closure). Then Countermodels holds but Moore-closure fails.

[19] Arguably, the topology capturing a human notion of similarity 'supervenes on' the Euclidean topology, i.e., is coarser: if an input s is in an open set U of the human-similarity topology, there should be some small threshold $\epsilon > 0$ such that all inputs s' that are at least ϵ-much low-level similar to s also are high-level similar to s—hence U also is Euclidean open.

6.3 Probabilistic robustness

Instead of a metric/uniform or topological/non-uniform explication of robustness, we can also consider a *probabilistic* explication (as already mentioned in Sect. 4.1).[20] Let us consider two motivating examples.

First, we could attempt to 'smoothen' the all-or-nothing nature of the uniform explication of robustness. There, to recall, we chose a metric d and a range $\epsilon > 0$ and defined $s \vDash \Box\varphi$ iff, for all s', if $d(s, s') < \epsilon$, then $s' \vDash \varphi$. The range ϵ provides a sharp cut-off point: states s' within the range are considered, states outside are not. To smoothen this cut-off point, we can instead choose a probability measure M_s (e.g., a multivariate normal distribution centered at s) such that states s' with small distance to s have high M_s-probability and states s' with large distance to s have low M_s-probability. Thus, we do not exclude states s' outside of the ϵ-range, we only give them lower probability. Then, for some threshold $\tau \in [0, 1]$—e.g., τ=0.7—, we define

(†) $M, s \vDash \Box\varphi$ iff $M_s(\{s' : M, s' \vDash \varphi\}) > \tau$.

Thus, we do not 'sharply' require that all states within range ϵ have property φ, but rather we 'smoothly' require that most states do. More precisely, when we sample states in such a way that similar states are more likely, then the probability that φ is satisfied is high, i.e., above τ. (Also cf. the probabilistic explication of robustness mentioned in footnote 17.)

Second, another example for a probabilistic notion of robustness comes from AI systems that generate output probabilistically. Specifically, consider LLMs. While they deterministically compute next-token probabilities, the generation of the output is still probabilistically done via a decoding method (e.g., *Top-K* or *Top-P* sampling). Thus, we can conceive of a state s as a pair (x,y) where x is a prompt and y is a generated output, and we have a probability measure M (determined by the decoding method) where $M(y|x)$ describes how likely output y is given prompt x. Say, we consider the prompt $x_0 =$ 'Tell me an interesting fact about Paris' and the LLM generates $y_0 =$ 'Paris is home to the Eiffel Tower'. We want to make sure that the LLM did not just happen to mention 'Paris' in the output but does so robustly. So we want that the probability of mentioning 'Paris' is high. Formally, we consider the property p of states $s = (x, y)$ that 'Paris' is mentioned in y. We define s having p robustly, i.e., $s \vDash \Box p$ as $M(\{y : (x, y) \vDash p\}|x) > \tau$, for some fixed threshold τ.

What these examples have in common is that we have a state space X(i.e., the states that the AI model can be in), and for each state $s \in X$, we have a probability measure M_s. In the first example, M_s describes the notion of 'similarity to s', and in the second example, where $s = (x, y)$ is a prompt-output pair, M_s is the probability measure that assigns a state (x', y') the probability $M(y'|x')$ if $x' = x$ and otherwise the probability 0. Formally, this is exactly the structure of a *Markov process*, i.e., M is a function $M : X \to \Delta(X)$ where X is a nonempty finite set and $\Delta(X)$ is the set of probability measures on X(Desharnais et al., 2002; Kozen et al., 2013).

[20] We thank an anonymous reviewer for urging us to say more on this.

(For brevity, we restrict us here to finite state spaces, so we do not have to introduce the concept of measurability.) We write M_s for $M(s)$. Our interpretation of robustness, then, is a fragment of Markovian logic (e.g., Kozen et al., 2013; Fagin & Halpern, 1994; Aumann, 1999; Heifetz & Mongin, 2001 and references therein). Given $\tau \in [0,1]$ and a Markov process $M : X \to \Delta(X)$, we start with an interpretation I of the atomic sentences, i.e., I maps atomic sentence p to the set of states $I(p) \subseteq X$ where p is satisfied. For instance, in our second example, $I(p)$ is the set of those states $s = (x,y)$ where output y contains 'Paris'. Then we define:

- $M, s \vDash p$ iff $s \in I(p)$
- $M, s \vDash \neg\varphi$ iff $M, s \nvDash \varphi$
- $M, s \vDash \varphi \wedge \psi$ iff $M, s \vDash \varphi$ and $M, s \vDash \psi$
- $M, s \vDash \Box\varphi$ iff $M_s(\llbracket\varphi\rrbracket) > \tau$, where $\llbracket\varphi\rrbracket := \{s \in X : M, s \vDash \varphi\}$
- $M, s \vDash \boxbox\, \varphi$ iff, for all s', $M, s' \vDash \varphi$.

As a result, $M, s \vDash \Diamond\varphi \Leftrightarrow M_s(\llbracket\varphi\rrbracket) \geq 1 - \tau$ (see Appendix A.5 for a proof).

The resulting logic is not a normal modal logic because the K-axiom is not satisfied (see Appendix A.5 for a counterexample). Importantly, though, it still satisfies axioms A1–A4 on a derivability relation, when understanding $\vdash$ as validity in M for a fixed interpretation I (and also when quantifying over all interpretations, see Appendix A.5).

Moreover, as already indicated in Sect. 4.1, this explication of robustness violates Factivity (i.e., $\vdash \Box\varphi \to \varphi$): just because φ has high probability does not mean it is true. However, we now argue that Weak-factivity P1$'$ is plausible. First, regarding $\Box\varphi \to \Diamond\,\varphi$, this is valid under the plausible requirement on the threshold that $\tau \geq \frac{1}{2}$ (proof in Appendix A.5). Second, regarding $\Box\Diamond\,\varphi \to \Diamond\,\varphi$, let us say that Markov process M is τ-*coherent* if, for all $A \subseteq X$ and $x \in X$,

$$\text{if } M_x(A) > \tau, \text{ then } M_x\big(\{y \in X : M_y(A) \leq \tau\}\big) \leq \tau.$$

If M is τ-coherent, then it validates $\Box\Diamond\,\varphi \to \Diamond\,\varphi$ for all φ and all interpretations I (proof in Appendix A.5). Not every Markov process is τ-coherent; however, a sufficient condition is that, for all $s \in X$, we have $M_s(\{s\}) \geq 1 - \tau$, i.e., each states assigns a non-negligible probability to itself (see Appendix A.5 for a proof). This makes sense on a similarity interpretation: if M_s describes degrees of similarity to s, then s should get a non-negligible degree of similarity to itself. In fact, τ-coherence in general seems plausible for our case: if state x considers states y likely that do not consider event A likely, then x itself should not consider A likely. In other words, if state x considers A likely, then x should not consider states with the opposite opinion likely.

The other principles have the same motivation as before: Whatever the choice T of trustworthy behavior, P2 says that trustworthy behavior should be robust, P3 says that every nontrivial trustworthy behavior is robustly not shown in some state, and P4 says that if φ is trustworthy, so is $\mathsf{W}\varphi = \varphi \vee \Box\neg\varphi$, i.e., the claim that the AI-model cannot be tricked into φ. Thus, our theorem applied to this specific explication of robustness yields the following as the 'smoothened' version of Theorem 1.

Corollary 3 *Let $\tau \in [\frac{1}{2}, 1]$, let $M : X \to \Delta(X)$ be a Markov process, and let I be an interpretation of the atomic sentences. Write $\vdash \varphi$ if $M, s \vDash \varphi$ for all $s \in X$. Then, for every set T of sentences, the principles P1', P2, P3, and P4 imply triviality: $\forall \varphi \in T : \vdash \Box\varphi \leftrightarrow \Diamond\varphi$, i.e., there is no $\varphi \in T$ and $s \in X$ with $1 - \tau \leq M_s([\![\varphi]\!]) \leq \tau$.*

Here we make essential use of the fact that we do not need full Factivity but only Weak-factivity to get our impossibility. The reason is that Fitch's lemma actually does not need full Factivity but only the axiom $\Box\neg\Box\varphi \to \neg\Box\varphi$, known as negative infallibility. (On a doxastic reading of $\Box$, this says that one's beliefs about what one does not believe are infallible.) Analogously, on our reinterpretation of Fitch's lemma, we also do not need full Factivity but only $\forall \varphi \in T : \vdash \Box\Diamond\varphi \to \Diamond\varphi$. For a discussion of Fitch's paradox beyond factive operators, see San (2020).

So, for Markov processes, the four principles imply triviality. But is there even a way to avoid triviality, when giving up on one of the principles? Yes, here is an example. Let $\tau = \frac{2}{3}$ and consider the two states $X = \{a, b\}$ with the Markov process M defined by $M_a(\{a\}) = \frac{3}{4}$, $M_a(\{b\}) = \frac{1}{4}$, $M_b(\{a\}) = \frac{1}{2}$, and $M_b(\{b\}) = \frac{1}{2}$. Consider an interpretation with $I(p) := \{a\}$. Let $T := \{p\}$. Then we do not have triviality, since $1 - \tau \leq M_b([\![p]\!]) \leq \tau$. Moreover, principles P1$'$, P2, and P4 are satisfied, and principle P3 indeed fails (proof in Appendix A.5).

6.4 Discussion: uniform vs non-uniform vs probabilistic

We started this section with the 'what to give up?' question: whether, in light of the no-go result, we should accept triviality or rather give up some of the principles. We made the discussion concrete by considering three suggestive explications of robustness: uniform, non-uniform, and probabilistic robustness. Here we summarize these explications and discuss what they suggest for the 'what to give up?' question.

Uniform robustness satisfies all principles and thus provides some support for the principles. By the no-go result, it trivializes, which is problematic in all but the simplest AI tasks. So rather than taking uniform robustness to suggest that we should accept triviality, the more plausible conclusion is that it is too strong a notion of robustness. So we should look for weaker notions of robustness that allow for nontriviality.

Thus, we have considered non-uniform robustness. It can easily afford nontrivial robust behavior, but, by the no-go result, it has to give up one of the principles, namely the Countermodels principle. Hence this explication fares better and tentatively suggests rather giving up a principle than accepting triviality. However, we also provided an argument that it is too weak a notion of robustness, because, for the natural choice of the Euclidean topology, every classifier turns out to be robust.

As an intermediate notion we considered probabilistic robustness. As a 'smoothened' version of uniform robustness, it can avoid triviality (in that sense it is weaker than uniform robustness). It also does not render every classifier robust (in that sense it is stronger than non-uniform robustness): If an input s gets classified positively, it could still be, e.g., if $s$is close to the decision boundary, that M_s assigns considerable probability to inputs beyond the decision boundary, so the classification is not robust on the probabilistic explication. However, whether probabilistic robustness is a satisfying explication still needs further discussion: we come back to this as an open question in Sect. 8.

7 Outlook: further applications

As an outlook, we highlight in this section the generality of the no-go result. Because we only used a derivability relation $\vdash$ with very minimal assumptions and treated T as just a variable for a set of sentences, we can give the no-go result many other interpretations besides robustness.

As discussed in Hornischer (2021, ch. 8), a general interpretation of $\Box$ is as stability: $M, s \vDash \Box\varphi$ means that the (abstract) state s *stably* has the (abstract) property φ in the considered model M, i.e., states that are relevantly similar to s still have property φ. Thus, the stable properties are those φ that, if true, are stably true (i.e., $\varphi \to \Box\varphi$ is valid in the model). This interpretation can then be applied to many instances of stability discussed in the philosophical literature. Some of the examples treated in Hornischer (2021, ch. 8) are: stability under errors of measurement, stability of belief, stability as invariance under transformations, and significance in mathematical modeling. The no-go result then provides limitations for each of those interpretations.

Moreover, also within AI, we can give $\Box$ a different interpretation. One is explainability (Adadi & Berrada, 2018; Miller, 2019). Let us explore this in the remainder of this section. Now $M, s \vDash \Box\varphi$ means that we can explain why the AI-model M shows behavior φ on input s. As before, we now want to know: what is the explainable behavior, i.e., the behavior that, if observed, we can explain? In other words, what can we say about $T = \{\varphi : M \vDash \varphi \to \Box\varphi\}$? Aiming for an answer with our no-go result, note that principle P2 is satisfied by construction, and principle P1 requires correctness of explanation: if we can explain why the AI-model shows behavior φ on input s, the model in particular should show behavior φ (i.e., the explanation should be *faithful*; see, e.g., Jacovi & Goldberg, 2020). To see principle P4, first note that we arguably have $\vdash \Box\varphi \to \Box\Box\varphi$ (which is known as the transitivity axiom in modal logic): if we can explain why φ, then we can also explain why we can explain why φ because we can point to our explanation of φ. With this, it is straightforward to show that if $\varphi \in T$, i.e., $\vdash \varphi \to \Box\varphi$, then also $\vdash \mathsf{W}\varphi \to \Box\mathsf{W}\varphi$, i.e., $\mathsf{W}\varphi \in T$. Moreover, we would expect nontriviality: For at least some $\varphi \in T$, if we cannot explain why the AI-model does not show behavior φ on input s, this does not mean that we can explain why it shows behavior φ on input s, so $\Diamond\varphi \to \Box\varphi$ should not be valid.[21] As the axioms also have some plausibility, our no-go result suggests that principle P3 must fail: There must be explainable behavior φ and an input s such that the AI-model does not show behavior φ on input s, but on no input can we explain why the AI-model does

[21] This is familiar from the interpretation of $\Box$ as (informal) provability. There we also do not have excluded middle: just because we cannot prove that something is false does not mean we can prove it is true. (Formal provability behaves somewhat differently as it, e.g., violates Factivity $\Box\varphi \to \varphi$, see Boolos (1993).)

not show behavior φ. More concisely: the absence of some explainable behavior is inexplicable!

To give a formal semantics for explainability, we might turn to truthmaker semantics (Fine, 2017). Given our AI-model M classifying handwritten digits, we allow as states pairs $s = (x, x')$ where x' is a part of the input image x (e.g., the 9 pixels in the middle of image x). Then we interpret $M, s \vDash \varphi$ as: x' is the part of the input x that was most relevant for the model showing behavior φ. This is an example of a feature attribution method in interpretable machine learning, and other explainability methods could be considered as well. The point is that this aligns well with the 'exactness' of truthmaker semantics. A state makes true a sentence in truthmaker semantics iff it contains exactly as much information to make the sentence true—no more, no less. Similarly, a feature attribution explanation should contain exactly the information to explain the behavior. Accordingly, the clause for conjunction should read: $M, s \vDash \varphi \wedge \psi$ iff s is the fusion of two parts s' and s'' such that $M, s' \vDash \varphi$ and $M, s'' \vDash \psi$. Thus, it would be interesting to compare the present no-go result with impossibility results concerning feature attribution (Bilodeau et al., 2024).

8 Open questions

We have seen that we can fruitfully use tools from formal epistemology—and, in particular, modal logic—to better understand robustness and trustworthiness. This framework also suggests many open questions for further research, as we will see now.

Bridges between AI and epistemology. Along the way, we have seen many connections between AI and epistemology, which we summarize in Fig. 1. Future research should work out those connections because they are—as mentioned—useful in both directions. It should also discuss to what extent modal conditions for knowledge—and, analogously, for trustworthiness—are compatible with best practices in statistics (Grote et al., 2024; Mayo-Wilson, 2018; Fletcher & Mayo-Wilson, 2024).[22] This is particularly important for a further evaluation of the probabilistic explication of robustness—to which we will come below.

One important such connection is to consider solutions to Fitch's paradox of knowability that have been suggested in philosophy and see if they translate to providing nontrivial explications of robustness and trustworthiness.[23] For an overview of such solutions, see Brogaard & Salerno (2019). For example, Edgington (1985) suggests to replace the knowability principle 'all truths can be known' with the principle that 'all actual truths are possibly known to be actual truths', which does not lead to paradox. Thus, the *actual* situation where the truth holds can be different from the *counterfactual* situation in which the knowledge about this actual truth is obtained. This has been criticized by Williamson (1987), but defended, e.g., by Rückert (2009) who employs a careful distinction between the indicative and subjunctive mood (akin

[22] We are grateful to an anonymous reviewer for raising this point.

[23] Many thanks to an anonymous reviewer for suggesting this and the references to Wansing (2002) and Rückert (2009) below.

to the actuality operator); or by Schlöder (2021) who employs possible 'courses of inquiry'. Yet another suggested solution is to change the underlying logic: So far we followed the scientific standard of using classical logic, but some solutions suggest moving to, e.g., relevant logic. We pick this up below, when discussing open questions concerning logic.

Moreover, one might take inspiration from robustness analysis in science (Weisberg, 2006; Woodward, 2006; Schupbach, 2018; Fletcher, 2020) and see how this may be transferred to machine learning. In essence, robustness analysis is a method to determine if a model makes trustworthy predictions by checking if different models of the same phenomenon make the same prediction. One way of transferring this to machine learning would be to train different models of the same phenomenon by using the same training dataset but varying, say, the seeds and hyperparameters. This is familiar from *ensemble learning* and is used, e.g., in Wortsman et al. (2022) to improve accuracy and robustness in fine-tuning large pre-trained models.

Furthermore, when it comes to further applications of the no-go result beyond robustness, we outlined another application to explainability. We may similarly consider further concepts like bias or fairness (e.g., express modally the various definitions of fairness (Shmueli et al., 2023) and their joint inconsistency (Chouldechova, 2017; Kleinberg et al., 2017)).

Further exploring explications. We have seen three explications of robustness: metric/uniform, topological/non-uniform, and probabilistic. While the no-go result provides an argument against the metric explication, we already suggested, in Sect. 6.2, ways to further refine the topological explication. Moreover, the probabilistic explication should be explored further. For example, turning probabilistic quantitative statements into all-or-nothing qualitative ones has well-known problems. Notoriously, using a fixed threshold (less than 1) faces the lottery paradox, which threatens the accepted qualitative statements to not be closed under conjunction (see Genin & Huber, 2022 for background and discussion). More generally, there are issues relating approaches to evidence, justification, and knowledge in statistics with those in epistemology (Fletcher & Mayo-Wilson, 2024), in particular when it comes to epistemic closure, i.e., the *K*-axiom (Mayo-Wilson, 2018)—which we did not assume, though, in our probabilistic explication. Working these out—generally, but also in the context of robustness in AI—is important, as stressed by Grote et al. (2024, p. 5). One approach that, at least, avoids the former problem of the lottery paradox is the stability theory of belief (Leitgeb, 2017). It requires, roughly, the statement's high probability to be stable under updating with further relevant information. Applying this to AI-models thus suggests a probabilistic explication of robustness that is worth working out (see Hornischer, 2021, ch. 8 for an initial discussion).

Logic. There also are various interesting questions about how to refine the modal logic that we use to describe the AI-model.

First, we already mentioned that some suggested solutions to Fitch's paradox of knowability change the underlying logic (see Brogaard & Salerno, 2019, sec. 3 for an overview). In particular, Wansing (2002) suggests a paraconsistent constructive relevant modal epistemic logic, and there is a recent surge of interest in combining relevant logic with epistemic logic (see, e.g., Bílková et al., 2016; Punčochář et al., 2023, Sedlár & Vigiani, 2024, Standefer & Mares, 2025, to mention but a few). Also, Hornischer & Berto (2025) provide an interpretation of relevant logic using dynamical systems, including neural networks. One might also interpret the conditional as expressing qualitative laws about the behavior of the AI-model governed by non-monotonic logic (Leitgeb, 2005). Finally, we can consider further ways of changing the underlying classical logic to a hyperintensional logic (Berto & Nolan, 2021): namely, the already mentioned truthmaker semantics as a potential formalization for explainability.

Second, we could also add a counterfactual ($\rightsquigarrow$) to describe counterfactual explanations like 'if you earned 10k more, you would have gotten the loan' (Wachter et al., 2017). Then $M, s \vDash \varphi \rightsquigarrow \psi$ would mean something along the lines of: for the inputs s' closest to s where φ holds, also ψ holds.[24] For an explication of this idea, see Hudetz & Crawford (2022). Determining a sound and complete axiomatization of such counterfactuals provides the logic of counterfactual explanations for the AI-model, which helps to assess its philosophical plausibility.

Third, one can also add operators known from dynamic epistemic logics (Baltag & Renne, 2016) to describe the learning dynamics of the AI-model and how this changes its behavior. For example, $s \vDash [p_2]\varphi$ could mean something along the lines of: after learning that input s depicts the digit 2, the AI-model shows behavior φ on input s. See, e.g., Baltag et al. (2019a, b) and Schultz Kisby et al. (2024). Similarly, we could also use the modal μ-calculus, so we can speak about the behavior of the AI-model after training on a dataset, i.e., after having reached a fixed point of the just mentioned learning process.

Quantitative version. Often, impossibility results spur *possibility* results by analyzing what still is possible. Is this the case here, too? Specifically, can we get possibilities when moving from a qualitative to a quantitative description of robustness? For example, rather than just qualitatively describing whether the AI-model behaves robustly on a given input, one might quantitatively analyze the robustness range ϵ or the probability of robust behavior (also cf. Freiesleben & Grote, 2023).

9 Conclusion

We have seen that the robustness and trustworthiness of AI-models can profitably be investigated using modal logic. We formulated four prima facie plausible principles about the notions of robustness and trustworthiness (no matter how they are understood precisely). However, with a reinterpretation of Fitch's lemma, we have proven that these principles imply triviality. This has consequences for formalizing the concepts of robustness and trustworthiness, which is necessary for verifying them for AI-models. Specifically, a uniform notion of robustness is too strong, while a non-uniform notion is too weak—with a probabilistic notion of robustness being a promising intermediary. We also noted the generality of the no-go result: it also applies to explainability in AI and stability in philosophy. As indicated by the open questions, there is much potential for a fruitful interaction between AI, on the one hand, and logic and formal epistemology, on the other hand.

[24] Here it again matters what counts as close: which metric we use, whether we consider only realistic or also adversarial inputs, etc. (Freiesleben, 2022).

A Appendix

A.1 Proof of Theorem 1

Proof of Theorem 1 If $\varphi \in T$ were not trivial, there are inputs s and s' with $M, s \vDash \varphi$ and $M, s' \nvDash \varphi$. Since the input space is path-connected, there is a path from $s$to s'. By dividing the path into sufficiently small steps, we get a sequence of points on the path $s = s_0, s_1, \ldots, s_N = s'$ with $d(s_i, s_{i+1}) < \epsilon$.[25] Since $M, s_0 \vDash \varphi$ and trustworthiness implies robustness, also $M, s_0 \vDash \Box\varphi$, so, since s_1 is similar to s_0, also $M, s_1 \vDash \varphi$. But now we can apply the same reasoning to s_1 and get that $M, s_2 \vDash \varphi$. And so on, until eventually $M, s_N \vDash \varphi$, contradicting $M, s' \nvDash \varphi$. □

A.2 Proof of Theorem 2

In its reinterpreted form, Fitch's lemma states: If all false *T*-sentences are robustly false somewhere, they, in fact, are already robustly false:

Lemma 4 *(Fitch's lemma). Let $\vdash$ be a derivability relation and T a set of sentences satisfying P1′ and P4.*[26] *Then (1) implies (2) where*

1. $\forall \varphi \in T: \ \vdash \neg\varphi \to \Diamond\hspace{-0.72em}\diamond\,\Box\neg\varphi$
2. $\forall \varphi \in T: \ \vdash \neg\varphi \to \Box\neg\varphi.$

The key idea for the proof is to substitute the (dual) Moore sentence into the assumption via a proof by contradiction (see Sect. 5.2). The formal proof is a slight adaption of the standard proof of Fitch's lemma (e.g van Benthem, 2004; Brogaard & Salerno, 2019). and goes as follows.

Proof Assume (1) and let $\varphi \in T$. To show $\vdash \neg\varphi \to \Box\neg\varphi$, we show 'by contradiction' that $\vdash \neg(\neg\varphi \to \Box\neg\varphi) \to \bot$ (using that $(\neg\chi \to \bot) \to \chi$ is a truth-functional tautology). We formally show this by a chain of conditionals. First, qua truth-functional tautology,

$$\vdash \neg(\neg\varphi \to \Box\neg\varphi) \to \neg\varphi \wedge \neg\Box\neg\varphi. \tag{1}$$

Next, by P4, let $\psi \in T$ with $\vdash \psi \leftrightarrow \mathsf{W}\varphi$. Hence, by using the appropriate truth-functional tautologies, $\vdash$ proves $\neg\psi \leftrightarrow \neg\mathsf{W}\varphi$ and $\neg\mathsf{W}\varphi \leftrightarrow \neg\varphi \wedge \neg\Box\neg\varphi$, so

[25] To be precise: Let $p : [0, 1] \to X$ be the path with X the input space and $p(0) = s$ and $p(1) = s'$. Since p is a continuous function from a compact metric space into a metric space, it is uniformly continuous (Heine–Cantor theorem). So there is $\delta > 0$ such that, for all $t, t' \in [0, 1]$, if $|t - t'| < \delta$, then $d(p(t), p(t')) < \epsilon$. Let $N \geq 1$ be big enough such that $\frac{1}{N} < \delta$ (which exists by the Archimedean property). For $i = 0, \ldots, N$, set $t_i := \frac{i}{N}$ and $s_i := p(t_i)$. Then $s_0 = p(t_0) = p(0) = s$ and $s_n = p(t_n) = p(1) = s'$ and, for $i \in \{0, \ldots, N-1\}$, $d(s_i, s_{i+1}) = d(p(t_i), p(t_{i+1})) < \epsilon$ since $|t_i - t_{i+1}| = |\frac{i+1}{N} - \frac{i}{N}| = \frac{1}{N} < \delta$.

[26] In fact, only the first conjunct of P1′ is needed, namely $\forall \varphi \in T : \ \vdash \Box\Diamond\varphi \to \Diamond\varphi$.

$$\vdash \neg\varphi \wedge \neg\Box\neg\varphi \leftrightarrow \neg\psi. \tag{2}$$

Since $\psi \in T$, we have, by assumption 1, that

$$\vdash \neg\psi \rightarrow ◈\,\Box\neg\psi. \tag{3}$$

By congruence for $\Box$ and (semi-) congruence for ◈, Equation 2 yields

$$\vdash ◈\,\Box\big(\neg\psi\big) \leftrightarrow ◈\,\Box\big(\neg\varphi \wedge \neg\Box\neg\varphi\big). \tag{4}$$

By distributivity for $\Box$ (i.e., $\vdash \Box(\varphi \wedge \psi) \rightarrow (\Box\varphi \wedge \Box\psi)$) and semi-congruence for ◈,

$$\vdash ◈\,\Box\big(\neg\varphi \wedge \neg\Box\neg\varphi\big) \rightarrow ◈\,\big(\Box\neg\varphi \wedge \Box\neg\Box\neg\varphi\big). \tag{5}$$

By P1$'$, we have $\vdash \Box\neg\Box\neg\varphi \rightarrow \neg\Box\neg\varphi$. By the truth-functional tautology $(\varphi \rightarrow \psi) \rightarrow (\chi \wedge \varphi \rightarrow \chi \wedge \psi)$ and semi-congruence for ◈, we have

$$\vdash ◈\,\big(\Box\neg\varphi \wedge \Box\neg\Box\neg\varphi\big) \rightarrow ◈\,\big(\Box\neg\varphi \wedge \neg\Box\neg\varphi\big). \tag{6}$$

By the truth-functional tautology $(\varphi \wedge \neg\varphi) \rightarrow \bot$ and semi-congruence for ◈,

$$\vdash ◈\,\big(\Box\neg\varphi \wedge \neg\Box\neg\varphi\big) \rightarrow ◈\,\bot. \tag{7}$$

Finally, since $\vdash \neg\bot$ (a truth-functional tautology), necessitation implies $\vdash ⧈\,\neg\bot$, so $\vdash \neg◈\,\bot$, so, by the truth-functional tautology $\neg\varphi \rightarrow (\varphi \rightarrow \bot)$, we have

$$\vdash ◈\,\bot \rightarrow \bot. \tag{8}$$

Now we can chain all these conditionals together[27] to get $\vdash \neg(\neg\varphi \rightarrow \Box\neg\varphi) \rightarrow \bot$, as desired. □

Now we can easily prove the no-go result.

Proof of Theorem 2 By P1$'$ and P4, Lemma 4 applies, so P3 implies, for any $\varphi \in T$, that $\vdash \neg\varphi \rightarrow \Box\neg\varphi$, i.e., $\vdash \Diamond\varphi \rightarrow \varphi$. By P2, also $\vdash \varphi \rightarrow \Box\varphi$. By P1$'$, also $\vdash \Box\varphi \rightarrow \Diamond\varphi$. So $\forall\varphi \in T:\ \vdash \Diamond\varphi \leftrightarrow \Box\varphi$, as needed. Since P1 implies P1$'$, this reasoning also goes through when using P1 instead of P1$'$. □

A.3 Proofs in Section 6.1

We show that, in the setting of uniform robustness described in Sect. 6.1, the logical axioms A1–A4 and the principles P1–P4 are satisfied.

[27] More precisely, if $\vdash \varphi \rightarrow \psi$ and $\vdash \psi \rightarrow \chi$, use the truth-functional tautology $(\varphi \rightarrow \psi) \rightarrow \big((\psi \rightarrow \chi) \rightarrow (\varphi \rightarrow \chi)\big)$ and modus ponens to obtain $\vdash \varphi \rightarrow \chi$.

- Axiom A1: if φ is a truth-functional tautology, then $M, s \vDash \varphi$ reduces to a tautological truth-condition in the metalanguage, which hence is satisfied. Moreover, if $\vdash \varphi$ and $\vdash \varphi \to \psi$, then φ is true at all states and any state which makes φ true also makes ψ true, so all states make ψ true, so $\vdash \psi$.
- Axiom A2 and A3: In this setting, $\Box$ (resp., $\boxed{\scriptstyle\Box}$) has the usual Kripke semantics with the binary relation sRs' iff $d(s, s') < \epsilon$ (resp., sRs' holds always). So these are normal modal operators and hence satisfy A2 and A3.
- Axiom A4: We assume that there is at least one state s, so $s \nvDash \bot$, so $\nvdash \bot$.
- Factivity: We have $M \vDash \Box\varphi \to \varphi$ because similarity is reflexive: if $s \vDash \Box\varphi$, then, since $d(s, s) = 0 < \epsilon$, also $s \vDash \varphi$.
- Robustness: If $\varphi \in T$, then, by definition of T, $M \vDash \varphi \to \Box\varphi$.
- Countermodels: If $\varphi \in T$ and $s \vDash \neg\varphi$ but there is no s' with $s' \vDash \Box\neg\varphi$, then, in particular, $s \nvDash \Box\neg\varphi$. So there is s' with $d(s, s') < \epsilon$ and $s' \nvDash \neg\varphi$, i.e., $s' \vDash \varphi$. Since $\varphi \in T$, we have $M \vDash \varphi \to \Box\varphi$, so $s' \vDash \Box\varphi$. Since $d(s', s) = d(s, s') < \epsilon$, have $s \vDash \varphi$, contradiction.
- Moore-closure: If $\varphi \in T$ but $\varphi \vee \Box\neg\varphi \notin T$, then there is s with $s \vDash \varphi \vee \Box\neg\varphi$ but $s \nvDash \Box\big(\varphi \vee \Box\neg\varphi\big)$. By the latter, there is s' with $d(s, s') < \epsilon$ and $s' \nvDash \varphi \vee \Box\neg\varphi$. Since $s' \nvDash \Box\neg\varphi$, there is s'' with $d(s', s'') < \epsilon$ and $s'' \vDash \varphi$. Since $M \vDash \varphi \to \Box\varphi$, also $s'' \vDash \Box\varphi$. Since $d(s'', s') = d(s', s'') < \epsilon$, also $s' \vDash \varphi$, contradiction.

A.4 Proofs in Section 6.2

We work in the setting of Sect. 6.2, proving the claims made there:

- Claim: $p \in T$ and $M \not\models \Diamond p \leftrightarrow \Box p$.

  Proof: First, it is not difficult to show: $\varphi \in T$ (i.e., φ is robust) iff $[\![\varphi]\!] := \{s \in X : M, s \vDash \varphi\}$ is an open subset of the input space X. Hence $p \in T$. Second, it is a basic fact of the topological semantics that $\Box$ is the topological interior operator and $\Diamond$ is the topological closure operator. Hence we cannot have $M \models \Diamond p \leftrightarrow \Box p$, because otherwise the closure of U, i.e., $[\![\Diamond p]\!]$ is identical to the interior of U, i.e., $[\![\Box p]\!]$, which implies that U is closed (and also open).

- Claim: Axioms A1–A4 as well as Factivity, Robustness, and Moore-closure are satisfied.

  Proof: For the axioms we reason as in Appendix A.3, except for $\Box$, which is readily seen to be a normal modal operator on the topological interpretation. Factivity is satisfied since $[\![\Box\varphi]\!]$ is the topological interior of $[\![\varphi]\!]$and hence a subset of it. Robustness is satisfied by the stipulation $T := \{\varphi : M \vDash \varphi \to \Box\varphi\}$. Moore-closure is satisfied because, if $\varphi \in T$, then $[\![\varphi]\!]$ is open, so also $[\![\varphi]\!] \cup [\![\Box\neg\varphi]\!]$ is open, and hence $\varphi \vee \Box\neg\varphi \in T$.

- Claim: For every nontrivial robust behavior φ (open and not closed), the AI-model can be φ-tricked ($\mathsf{W}\varphi$ is false on some inputs) but for generic inputs it cannot

($\mathsf{W}\varphi$ is generically true, i.e., an open and dense set)

Proof: Failure of the Countermodels principle means that there is an input s where $\neg\mathsf{W}\varphi$ is true (hence $\mathsf{W}\varphi$ is false on some input) but $\Diamond\!\!\!\!\diamond\,\Box\neg\mathsf{W}\varphi$ is false. The latter means that $\Box\neg\mathsf{W}\varphi$ is false everywhere, i.e., $\neg\Box\neg\mathsf{W}\varphi$ is true everywhere, so $[\![\Diamond\mathsf{W}\varphi]\!] = X$. Hence the closure of $[\![\mathsf{W}\varphi]\!]$ is the whole space, so $\mathsf{W}\varphi$ is dense. Since $\mathsf{W}\varphi \in T$, it also is open.

- Claim: Violating $\mathsf{W}\varphi$ amounts to being in the topological boundary of $[\![\varphi]\!]$.

  Proof: Note that $[\![\neg\mathsf{W}\varphi]\!] = [\![\neg\varphi \wedge \neg\Box\neg\varphi]\!] = [\![\neg\varphi]\!] \cap [\![\Diamond\varphi]\!]$. Since $[\![\varphi]\!]$ is open and $\Diamond$ is topological closure, $[\![\neg\mathsf{W}\varphi]\!]$ is the closure of $[\![\varphi]\!]$ minus the interior of $[\![\varphi]\!]$, which, by definition, is the topological boundary of $[\![\varphi]\!]$.

A.5 Proofs in Section 6.3

In this subsection, let $\tau \in [0, 1]$, let $M : X \to \Delta(X)$ be a Markov process, and let I be an interpretation. We prove the claims made in Sect. 6.3.

- Claim: $M, s \vDash \Diamond\varphi \Leftrightarrow M_s([\![\varphi]\!]) \geq 1 - \tau$.

  Proof: $M, s \vDash \Diamond\varphi$ iff $M_s([\![\neg\varphi]\!]) \not> \tau$ iff $1 - M_s([\![\varphi]\!]) \leq \tau$ iff $M_s([\![\varphi]\!]) \geq 1 - \tau$.

- Claim: The K-axiom is not valid for the Markov process interpretation. (The counterexample can also be chosen to be τ-coherent.)

  Proof: Let $\tau = \frac{2}{3}$. Consider the three states $X = \{a, b, c\}$ and the Markov process M which maps $x \in X$ to the measure M_x defined by $M_x(\{x\}) = \frac{1}{2} \geq 1 - \tau$and $M_x(\{y\}) = \frac{1}{4}$ for $y \in X \setminus \{x\}$. (The sufficient condition mentioned in the claim below shows that this Markov process is τ-coherent.) Consider the interpretation I with $I(p) = \{a, b\}$ and $I(q) = \{a\}$. So $[\![p \to q]\!] = \{c, a\}$. Then $a \vDash \Box(p \to q)$ since $M_a([\![p \to q]\!]) = \frac{1}{4} + \frac{1}{2} > \tau$, and $a \vDash \Box p$ since $M_a([\![p]\!]) = \frac{1}{2} + \frac{1}{4} > \tau$, but $a \nvDash \Box q$ since $M_a([\![q]\!]) = \frac{1}{2} \not> \tau$.

- Claim: Axioms A1–A4 are satisfied when reading $\vdash$ as $M, s \vDash \varphi$ for all $s \in X$. This implies that this also holds when we define $\vdash$ by additionally quantifying over all interpretations on M.

  Proof: Regarding A1, we reason as in Appendix A.3. Regarding A2, we have, for any $s \in X$, that $M_s([\![\varphi \wedge \psi]\!]) \leq M_s([\![\varphi]\!])$, and similarly for ψ, so if $M_s([\![\varphi \wedge \psi]\!]) > \tau$, then $M_s([\![\varphi]\!]) > \tau$ and $M_s([\![\psi]\!]) > \tau$, hence $\vdash \Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$. Moreover, if $\vdash \varphi \leftrightarrow \psi$, then $[\![\varphi]\!] = [\![\psi]\!]$, so, for all $s \in X$, we have $M_s([\![\varphi]\!]) > \tau$ iff $M_s([\![\psi]\!]) > \tau$, hence $\vdash \Box\varphi \leftrightarrow \Box\psi$. Regarding A3, we use the fact that $\boxbox$ is a normal modal operator. Regarding A4, let, since X is nonempty, $s \in X$, so $s \nvDash \bot$, so $\nvdash \bot$.

- Claim: If $\tau \geq \frac{1}{2}$, then, for any $\varphi \in \mathcal{L}$ and $s \in X$, we have $M, s \vDash \Box\varphi \to \Diamond\varphi$.

  Proof: If $M_s(\llbracket\varphi\rrbracket) > \tau$, then, since $\tau \geq \frac{1}{2} \geq 1 - \tau$, we have $M_s(\llbracket\varphi\rrbracket) \geq 1 - \tau$.

- Claim: If M is τ-coherent, then, for any $\varphi \in \mathcal{L}$ and $s \in X$, we have $M, s \vDash \Box\Diamond\varphi \to \Diamond\varphi$.

  Proof: By contraposition, assume $s \nvDash \Diamond\varphi$ and show $M, s \nvDash \Box\Diamond\varphi$. By assumption, $s \vDash \Box\neg\varphi$. Write $A := \llbracket\neg\varphi\rrbracket$. Then $M_s(A) > \tau$, so, since M is τ-coherent, $M_s\big(\{y \in X : M_y(A) \leq \tau\}\big) \leq \tau$. Note that

  $$\{y \in X : M_y(A) \leq \tau\} = \{y \in X : y \nvDash \Box\neg\varphi\} = \llbracket\Diamond\varphi\rrbracket .$$

  So we have $M_s\big(\,\llbracket\Diamond\varphi\rrbracket\,\big) \leq \tau$. Hence $s \nvDash \Box\Diamond\varphi$, as needed.

- Claim: There are Markov processes that are not τ-coherent.

  Proof: Let $\tau = \frac{2}{3}$. Consider the two states $X = \{a, b\}$ and the Markov process M defined by $M_a(a) = \frac{1}{4}$ and $M_a(b) = \frac{3}{4}$, and $M_b(a) = \frac{3}{4}$ and $M_b(b) = \frac{1}{4}$. Consider $A := \{b\} \subseteq X$ and $x := a$. Then $M_a(A) = \frac{3}{4} > \tau$, but $M_a(\{y : M_y(A) \leq \tau\}) = M_a(\{b\}) = \frac{3}{4} \nleq \tau$.

- Claim: Assume that, for all $s \in X$, we have $M_s(\{s\}) \geq 1 - \tau$. Then M is τ-coherent.

  Proof: Let $A \subseteq X$ and $s \in X$. Assume $M_s(A) > \tau$. Then

  $$M_s\big(\{y \in X : M_y(A) \leq \tau\}\big) \leq M_s\big(X \setminus \{s\}\big) = 1 - M_s(\{s\}) \leq \tau,$$

  as needed.

- For the Markov process M defined in Sect. 6.3, principles P1′, P2, and P4 are satisfied but principle P3 fails.

  Proof: Recall that $\tau = \frac{2}{3}$ and $M_a(\{a\}) = \frac{3}{4}$, $M_a(\{b\}) = \frac{1}{4}$, $M_b(\{a\}) = \frac{1}{2}$, and $M_b(\{b\}) = \frac{1}{2}$, with $I(p) = \{a\}$ and $T = \{p\}$. Since $\tau \geq \frac{1}{2}$ and M satisfies the sufficient condition for τ-coherence, principle P1′ is satisfied. For P2, note that if $s \vDash p$, then $s{=}a$, so $M_s(\llbracket p\rrbracket) = \frac{3}{4} > \tau$, so $s \vDash \Box p$. Moreover, note that $\llbracket\Box\neg p\rrbracket = \emptyset$, because if $x \vDash \Box\neg p$, then $M_x(\{b\}) > \tau$, but $M_a(\{b\}) = \frac{1}{4} \not> \tau$ and $M_b(\{b\}) = \frac{1}{2} \not> \tau$. Hence P3 fails, since $b \vDash \neg p$ but $b \nvDash \text{◈}\Box\neg p$. Finally, principle P4 holds, since, for $\varphi \in T$, take $\psi := \varphi = p \in T$, then $\vdash \psi \leftrightarrow \mathsf{W}\varphi$, since $\llbracket\mathsf{W}\varphi\rrbracket = \llbracket p\rrbracket \cup \llbracket\Box\neg p\rrbracket = \llbracket p\rrbracket = \llbracket\psi\rrbracket$.

Acknowledgements For inspiring discussions and comments, I am grateful to Samson Abramsky, Johan van Benthem, Franz Berto, Arianna Betti, Luca Incurvati, Michiel van Lambalgen, Hannes Leitgeb, Aybüke Özgün, Robert van Rooij, Sonja Smets, and the audiences of the MLC seminar and the Logic of Conceivability Seminar at the University of Amsterdam as well as the 2025 ‘Epistemology and Theory of Machine Learning' workshop at LMU Munich. This paper grew out of Chapter 8 of my PhD thesis. I would also like to thank the anonymous referees of this journal for very helpful feedback.

Funding Open Access funding enabled and organized by Projekt DEAL. Initial parts of this work were done within the project 'Foundations of Analogical Thinking' (Project No. 322–20–017) of the research program 'PhDs in the Humanities', financed by the Dutch Research Council (NWO). Later parts of the work were funded by the Bavarian Ministry for Digital Affairs.

Data availability Not applicable.

Declarations

Competing interests None.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Albarghouthi, A. (2021). Introduction to neural network verification. *Foundations and Trends® in Programming Languages*, *7*(1–2), 1–157. https://doi.org/10.1561/2500000051

Aumann, R. J. (1999). Interactive epistemology ii: Probability. *International Journal of Game Theory*, *28*(3), 301–314. https://doi.org/10.1007/s001820050112

Baltag, A., Gierasimczuk, N., Özgün, A., Vargas Sandoval, A. L., & Smets, S. (2019a). A dynamic logic for learning theory. *Journal of Logical and Algebraic Methods in Programming*, *109*, 100485. https://doi.org/10.1016/j.jlamp.2019.100485

Baltag, A., Li, D., & Pedersen, M. Y. (2019b). On the right path: A modal logic for supervised learning. In P. Blackburn, E. Lorini, & M. Guo (Eds.), *Logic, Rationality, and Interaction (LORI 2019). Lecture notes in computer science* (Vol. 11813, pp. 1–14). Springer. https://doi.org/10.1007/978-3-662-60292-8_1

Baltag, A., & Renne, B. (2016). Dynamic epistemic logic. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Winter 2016 edn. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/dynamic-epistemic/

Bastounis, A., Campodonico, P., Schaar, M., Adcock, B., & Hansen, A. C. (2024). On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of 'i don't know'. *arXiv*, 2408.02357.

Berto, F., & Nolan, D. (2021). Hyperintensionality. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Summer 2021 edn. Metaphysics Research Lab, Stanford University.

Bílková, M., Majer, O., & Peliš, M. (2016). Epistemic logics for sceptical agents. *Journal of Logic and Computation*, *26*(6), 1815–1841. https://doi.org/10.1093/logcom/exv009

Bilodeau, B., Jaques, N., Koh, P. W., & Kim, B. (2024). Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, *121*(2), 2304406120.

Blackburn, P., Rijke, M., & Venema, Y. (2001). *Modal logic. Cambridge tracts in theoretical computer science*. Cambridge University Press.

Boolos, G. (1993). *The logic of provability*. Cambridge University Press.

Braiek, H. B., Khomh, F. (2025). Machine learning robustness: A primer. In M. Lorenzi, M. A. Zuluaga (Eds.), *Trustworthy AI in medical imaging* (1st edn., pp. 1–43). Elsevier.

Brogaard, B., & Salerno, J. (2019). Fitch's paradox of knowability. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Fall 2019 edn. Metaphysics Research Lab, Stanford University.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), 12625. https://doi.org/10.1111/phc3.12625

Buckner, C. J. (2023). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press. https://doi.org/10.1093/oso/9780197653302.001.0001

Chouldechova, A. (2017). Fair prediction with disparate impact: A studyof bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Desharnais, J., Edalat, A., & Panangaden, P. (2002). Bisimulation for labelled markov processes. *Information and Computation*, *179*(2), 163–193. https://doi.org/10.1006/inco.2001.2962

Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A., & Seshia, S. A. (2019). A formalization of robustness for deep neural networks. arXiv:1903.10033.

Dummett, M. (2001). Victor's error. *Analysis*, *61*(1), 1–2. https://doi.org/10.1093/analys/61.1.1

Edgington, D. (1985). The paradox of knowability. *Mind*, *94*(376), 557–568. https://doi.org/10.1093/mind/XCIV.376.557

Égré, P. (2015). Vagueness: Why do we believe in tolerance? *Journal of Philosophical Logic*, *44*(6), 663–679.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1625–1634).

Fagin, R., & Halpern, J. Y. (1994). Reasoning about knowledge and probability. *Journal of the ACM (JACM)*, *41*(2), 340–367.

Fine, K. (2017). *Truthmaker semantics* (pp. 556–577). Wiley-Blackwell. Chap. 22.

Fitch, F. B. (1963). A logical analysis of some value concepts. *The Journal of Symbolic Logic*, *28*(2), 135–142.

Fletcher, S. C. (2020). The principle of stability. *Philosophers' Imprint*, *20*(3), 1–22.

Fletcher, S. C., & Mayo-Wilson, C. (2024). Evidence in classical statistics. In M. Lasonen-Aarnio & C. Littlejohn (Eds.), *Routledge Handbook of the philosophy of evidence* (pp. 515–527). Routledge.

Freiesleben, T. (2022). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, *32*, 77–109. https://doi.org/10.1007/s11023-021-09580-9

Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, *202*, 109. https://doi.org/10.1007/s11229-023-04334-9

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Genin, K., & Huber, F. (2022). Formal representations of belief. In E. N. Zalta & U. Nodelman (Eds.), *The stanford encyclopedia of philosophy*, Fall 2022 edn. Metaphysics Research Lab, Stanford University, ???.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples) arXiv:1412.6572.

Grote, T., Genin, K., & Sullivan, E. (2024). Reliability in machine learning. *Philosophy Compass*, *19*(5). https://doi.org/10.1111/phc3.12974

Gustafsson, F. K., Danelljan, M., & Schon, T. B. (2020). Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 318–319.

Heifetz, A., & Mongin, P. (2001). Probability logic for type spaces. *Games and Economic Behavior*, *35*(1–2), 31–53. https://doi.org/10.1006/game.1999.0788

High-Level Expert Group on Artificial Intelligence. (2019, April). Ethics guidelines for trustworthy AI. *Technical report*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hornischer, L. (2021). *Dynamical systems via domains: Toward a unified foundation of symbolic and non-symbolic computation.* PhD thesis, University of Amsterdam, Institute for Logic, Language and Computation. https://www.illc.uva.nl/Research/Publications/Dissertations/DS-2021-10.text.pdf. Accessed 1 Sept 2025.

Hornischer, L., & Berto, F. (2025). *The logic of dynamical systems is relevant.* (p. 012). Mind. https://doi.org/10.1093/mind/fzaf012

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, *37*, 100270.

Huang, X., Kwiatkowska, M., Wang, S., & Wu, M. (2017). Safety verification of deep neural networks. In R. Majumdar & V. Kunčak (Eds.), *Computer aided verification (CAV 2017). Lecture notes in computer science* (Vol. 10426). Springer. https://doi.org/10.1007/978-3-319-63387-9_1

Hudetz, L., & Crawford, N. (2022). Variation semantics: When counterfactuals in explanations of algorithmic decisions are true. http://philsci-archive.pitt.edu/20626/

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Ichikawa, J. J., & Steup, M. (2018). The analysis of knowledge. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Summer 2018 edn. Metaphysics Research Lab, Stanford University.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th innovations in theoretical computer science conference (ITCS 2017). Leibniz International proceedings in informatics (LIPIcs)* (Vol. 67, pp. 43–14323). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J.Kundaje, A., … Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning* (Vol. 139, pp. 5637–5664). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/koh21a.html

Kozen, D., Larsen, K. G., Mardare, R., & Panangaden, P. (2013). Stone duality for markov processes. In *2013 28th Annual ACM/IEEE Symposium on Logic in Computer Science* (pp. 321–330). IEEE. https://doi.org/10.1109/LICS.2013.38

Leitgeb, H. (2005). Interpreted dynamical systems and qualitative laws: From neural networks to evolutionary systems. *Synthese*, *146*(1), 189–202.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press.

Mayo-Wilson, C. (2018). Epistemic closure in science. *The Philosophical Review*, *127*(1), 73–114.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*. https://doi.org/10.1016/j.artint.2018.07.007

Moggi, E., Farjudian, A., Duracz, A., & Taha, W. (2018). Safe & robust reachability analysis of hybrid systems. *Theoretical Computer Science*, *747*, 75–99. https://doi.org/10.1016/j.tcs.2018.06.020

Pacuit, E. (2017). *Neighborhood semantics for modal logic*. Springer. https://doi.org/10.1007/978-3-319-67149-9

Perdomo, J., Zrnic, T., Mendler-Dünner, C., & Hardt, M. (2020). Performative prediction. In H. D. Iii & A. Singh (Eds.), *Proceedings of the 37th International conference on machine learning* (Vol. 119, pp. 7599–7609). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v119/perdomo20a.html

Pessach, D., Shmueli, E. (2023). Algorithmic fairness. In L. Rokach, O. Maimon, E. Shmueli (Eds.), *Machine Learning for Data Science Handbook* (pp. 867–886). Springer. https://doi.org/10.1007/978-3-031-24628-9_37

Pritchard, D. (2005). *Epistemic luck*. Oxford University Press.

Punčochář, V., Sedlár, I., & Tedder, A. (2023). Relevant epistemic logic with public announcements and common knowledge. *Journal of Logic and Computation*, *33*(2), 436–461. https://doi.org/10.1093/logcom/exac100

Quine, W. V. O. (1960). *Word and object*. Cambridge, Mass: MIT Press.

Rabinowitz, D. (2024). The safety condition for knowledge. *The internet encyclopedia of philosophy*. https://www.iep.utm.edu/safety-c/

Rott, H. (2004). Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, *61*, 469–493.

Ruan, W., Wu, M., Sun, Y., Huang, X., Kroening, D., & Kwiatkowska, M. (2019). Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19* (pp. 5944–5952). https://doi.org/10.24963/ijcai.2019/824

Rückert, H. (2009). A solution to fitch' paradox of knowability. In S. Rahman, J. Symons, D. M. Gabbay, & J. P. Bendegem (Eds.), *Logic, epistemology, and the Unity of science. Logic, epistemology, and the Unity of science* (Vol. 1, pp. 351–380). Springer. https://doi.org/10.1007/978-1-4020-2808-3_18

Russell, B. (1948). *Human knowledge: Its scope and its limits*. Allen & Unwin.

San, W. K. (2020). Fitch's paradox and level-bridging principles. *The Journal of Philosophy*, *117*(1), 5–29. https://doi.org/10.5840/jphil202011711

Schlöder, J. J. (2021). Counterfactual knowability revisited. *Synthese*, *198*, 1123–1137. https://doi.org/10.1007/s11229-019-02087-y

Schultz Kisby, C., Blanco, S. A., & Moss, L. S. (2024). What do hebbian learners learn? reduction axioms for iterated hebbian learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 14894–14901.

Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *British Journal for the Philosophy of Science*, *69*(1), 275–300. https://doi.org/10.1093/bjps/axw008

Sedlár, I., & Vigiani, P. (2024). Epistemic logics for relevant reasoners. *Journal of Philosophical Logic*, *53*(5), 1383–1411. https://doi.org/10.1007/s10992-024-09770-7

Serban, A., Poll, E., & Visser, J. (2020). Adversarial examples on object recognition: A comprehensive survey. *ACM Computing Surveys*, *53*(3), 1–38. https://doi.org/10.1145/3398394

Seshia, S. A., Sadigh, D., & Sastry, S. S. (2022). Toward verified artificial intelligence. *Communications of the ACM*, *65*(7), 46–55.

Shehtman, V. (1999). "Everywhere" and "Here". *Journal of Applied Non-Classical Logics*, *9*(2–3), 369–379.

Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141.

Standefer, S., & Mares, E. (2025). Symmetry and completeness in relevant epistemic logic. *Journal of Philosophical Logic*, *54*(2), 429–450. https://doi.org/10.1007/s10992-025-09791-w

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 18583–18599). https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

van Benthem, J. (2004). What one may come to know. *Analysis*, *64*(2), 95–105.

van Benthem, J., & Bezhanishvili, G. (2007). Modal logics of space. In M. Aiello, I. Pratt-Hartmann, & J. van Benthem (Eds.), *Handbook of spatial logics* (pp. 217–298). Springer. https://doi.org/10.1007/978-1-4020-5587-4_5

Vandenburgh, J. (2023). Machine learning and knowledge: Why robustness matters. https://arxiv.org/abs/2310.19819

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, *31*, 841.

Wansing, H. (2002). Diamonds are a philosopher's best friends. *Journal of Philosophical Logic*, *31*(6), 591–612. https://doi.org/10.1023/A:1021256513220

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science*, *73*(5), 730–742. https://doi.org/10.1086/518628

Williamson, T. (1987). On the paradox of knowability. *Mind*, *96*(382), 256–261. https://doi.org/10.1093/mind/XCVI.382.256

Williamson, T. (1994). *Vagueness*. Routledge.

Williamson, T. (1999). On the structure of higher-order vagueness. *Mind*, *108*(429), 127–143.

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology*, *13*(2), 219–240. https://doi.org/10.1080/13501780600733376

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., & Schmidt, L. (2022). Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 23965–23998). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v162/wortsman22a.html

Zhou, C., Shaikh, R. A., Li, Y., & Farjudian, A. (2023). A domain-theoretic framework for robustness analysis of neural networks. *Mathematical Structures in Computer Science*, 1–38. https://doi.org/10.1017/S0960129523000142