Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 333–345
doi:10.1017/S0963180122000688

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Learning to Live with Strange Error: Beyond Trustworthiness in Artificial Intelligence Ethics

Charles Rathkopf[1]* and Bert Heinrichs[1,2]

[1]INM-7, Forschungszentrum Jülich GmbH, Jülich, Germany
[2]The Institute for Science and Ethics (IWE) The University of Bonn Bonner Talweg 57, 53113, Germany
*Corresponding author. Email: c.rathkopf@fz-juelich.de

Abstract

Position papers on artificial intelligence (AI) ethics are often framed as attempts to work out technical and regulatory strategies for attaining what is commonly called *trustworthy AI*. In such papers, the technical and regulatory strategies are frequently analyzed in detail, but the concept of trustworthy AI is not. As a result, it remains unclear. This paper lays out a variety of possible interpretations of the concept and concludes that none of them is appropriate. The central problem is that, by framing the ethics of AI in terms of trustworthiness, we reinforce unjustified anthropocentric assumptions that stand in the way of clear analysis. Furthermore, even if we insist on a purely epistemic interpretation of the concept, according to which trustworthiness just means measurable reliability, it turns out that the analysis will, nevertheless, suffer from a subtle form of anthropocentrism. The paper goes on to develop the concept of strange error, which serves both to sharpen the initial diagnosis of the inadequacy of trustworthy AI and to articulate the novel epistemological situation created by the use of AI. The paper concludes with a discussion of how strange error puts pressure on standard practices of assessing moral culpability, particularly in the context of medicine.

Keywords: AI ethics; trustworthy AI; anthropocentrism; strange error; reliability; moral culpability

Trustworthy AI

The term *trustworthy AI* is typically used to describe a particularly valuable technological goal. A common thought is that, in order to use artificial intelligence (AI) responsibly, we must deploy only those systems that have been deemed trustworthy. This paper gives reasons to think that trustworthy AI is not the best concept around which to organize our thinking about AI ethics. It shows that, depending on which concept of trust one has in mind, the goal of prioritizing trustworthy AI may be either irresponsible or incoherent. On one interpretation of trust, the proposal that we ought to cultivate trust in AI systems would be irresponsible because it would promote a negligent attitude toward safety. On another interpretation, the proposal that we cultivate trust in AI systems would be incoherent because it relies on the false anthropocentric assumption that AI models have sociocognitive capacities of the sort involved in interpersonal relationships. This dilemma is developed in the "Troubles with Trustworthiness" section.

In response, one might object that dilemma misinterprets the intended meaning of the phrase "trustworthy AI." In the context of AI ethics, *trustworthy* just means "confers epistemic justification." Moreover, insofar as epistemic justification for a belief derives from the judgment of an AI model, the degree of justification can be estimated with quantitative performance metrics, which are themselves free of anthropocentric bias. In the "A Non-Anthropocentric Version of Trust?" section, this objection serves as a jumping off point from which to explore the complex topic of epistemic risk in AI models. The upshot of the discussion is that the justification of AI judgments is importantly different from the

justification of human expert judgments. Reliability is a central factor in both cases, but to presume that reliability in AI models carries the same epistemic weight as it does in the case of human judgment is to ignore what this paper dubs the *strange error* associated with machine learning (ML) classification problems.

Finally, the paper argues that the existence of strange error implies that we will encounter risks that no one can be expected to have anticipated. From this, it follows that there can be harm for which no one can be blamed. The existence of strange errors gives us reason to think that, in the domain of ML judgments, standard practices of assessing culpability will be inadequate, unless and until we develop a general theory of ML artifacts that will enable us to actively mitigate strange error and associated risks. In light of this current inadequacy, the paper argues that a morally appropriate response to AI error must include schemes for restitution that do not depend on the identification of a blameworthy party. This conclusion has particular resonance in the context of medical AI because medical AI systems are particularly disposed to cause harm.

Troubles with Trustworthiness

The concept of trustworthy AI has become influential in part by virtue of its prominent role in many position papers published by large institutions, including tech companies, think tanks, and the European Union. The document that is perhaps most responsible for pushing the concept into the limelight is entitled "Ethics guidelines for trustworthy AI."[1] It is the central position paper on AI ethics published by the High-Level Expert Group of the European Union (HLEG). That document contains many valuable insights and deserves to be widely read. Its value derives largely from the more concrete policy recommendations it makes, many of which can be evaluated independently of the conceptual framework in which they are embedded.

On the conceptual side, the document is somewhat underdeveloped. No definition of trustworthiness is offered. There is also no attempt to explain why trustworthiness is an appropriate concept around which to organize efforts to develop an ethics of AI. Instead, the bulk of the document is concerned with identifying which technical and regulatory strategies are likely to promote the goal of trustworthiness, the meaning of which is presumed to be intuitively accessible. This emphasis on practical goals is understandable, and at least partially defensible. The document is not, after all, meant to be read as a philosophical treatise. However, the lack of attention to conceptual precision does have an influence on the reasoning that underlies the practical recommendations. As the paper argues below, this reasoning fails to take sufficient account of how different ML classification judgments are from human ones, and thereby fails to recognize an important sense in which the epistemic justification associated with ML classification can be weaker than it seems.

The conceptual difficulties underlying that conclusion are quite subtle. Other difficulties are more obvious, and one of them needs to be addressed immediately. It is the fact that the HLEG document is persistently ambiguous about the kind of thing the HLEG regards as a candidate for trustworthiness.

> "Trust in the development, deployment and use of AI systems concerns not only the technology's inherent properties, but also the qualities of the socio-technical systems involving AI applications."[2]

Is it the ML model itself that we should hope to trust? Or is it the broader social system that generates, deploys, and regulates the ML model? The quote above is just one example in which it seems that the authors of the HLEG document suggest that we trust both. They also use the concept of trust univocally, tacitly assuming that trusting a social system for managing a technology is the same thing as trusting the technology itself. However, this assumption is nearly as doubtful as the assumption that trusting a prison system is the same thing as trusting a prisoner. Whatever the best analysis of trustworthiness turns out to be, the properties that make AI technology trustworthy are different from the properties that make the social systems surrounding that technology trustworthy. Here, the primary concern is the technology itself. The question at issue in this paper is: What, if anything, could make the technology itself trustworthy?

Let us begin the analysis of trustworthy AI technology by asking what, in general, trustworthiness is. When someone trusts you to take good care of their car, they not only estimate that the probability that you will take good care of the car is sufficiently high to overcome their scruples, but they also demonstrate a willingness to presume that you *want* to take care of the car and that your motivation is not exclusively self-interested. On the basis of examples like this, some philosophers have suggested that trustworthy agents must be motivated by prosocial attitudes.[3,4] If the presumption of motivation by prosocial attitudes is conceptually necessary for trust, then the question of whether AI is trustworthy is not compelling. Hardly anyone thinks that present-day AI has the capacity for prosocial motivation.

According to an alternative philosophical analysis, the distinctive feature of trust manifests itself in the effect it has on the parties involved in a relation of trust. On this view, relations of trust are characterized by cooperation-supporting feedback loops. When you trust someone, that person becomes more willing to trust you in return.[5] If trust requires this kind of reciprocity, then the question of which sorts of AI can be trusted is, once again, uncompelling. Present-day AI systems are not capable of trusting a human in any interesting sense, and are certainly not capable of learning to adjust their degree of trust in light of experience. Consequently, there can be no feedback-sustaining mechanism in the case of human relationships with AI.

These brief observations give us some reason to suspect that debates about trust in AI are, as multiple authors have recently warned, uncritically anthropomorphic.[6,7] In a recent article, Mark Ryan divides up existing theories of trust differently than this paper does, but also argues that neither group of theories applies well to AI.[8] According to Ryan, all theories of trust are too anthropomorphic to apply to AI. However, Thi Nguyen has developed a new theory of trust that Ryan overlooks, and Nguyen's theory is specifically designed to apply to technological artifacts. Might Nguyen's theory overcome the charge of anthropomorphism and offer us an appropriate way to think about trustworthy AI?

Nguyen says that trust is an unquestioning attitude. When we trust a person, we adopt the policy that confirming what they say is unnecessary. We do not need to waste cognitive effort on convincing ourselves that the person is speaking the truth. When we trust an artifact, we do something similar: We rely on the artifact, but waste no conscious effort on verifying that it is doing what it should. Nguyen's discussion focuses on the sorts of artifacts involved in action, such as the sort of rope used in rock climbing. In that context, to trust the rope is to ignore that it is there at all, and instead, to focus entirely on the execution of appropriate bodily movement. Nguyen's view of trust is not weighed down by anthropomorphic implications. Nevertheless, even on his artifact-centric view, trustworthiness is not the kind of thing we should want from AI systems because it should not be our goal to build a relationship with AI systems in which we can simply ignore what they do. This is especially clear in examples of ML systems that categorize people and their preferences. It is now widely appreciated that the use of such systems can regiment and reinforce pre-existing social biases.[9,10] It is consequently unlikely that we will ever have the ethical luxury of being able to ignore the workings of AI systems entirely.

One might criticize this dismissal of Nguyen's view by saying that it fails to properly imagine how it should be applied in the case of AI, where the agent who judges whether a technology is trustworthy is not an individual, as it is in the case of the rock climbing rope, but instead a group of humans, or perhaps an entire society. To respond to this criticism, one can appeal to a simplifying assumption, which says that if a technology deserves to be regarded as trustworthy by an entire society, then it deserves to be regarded as trustworthy by at least one person in that society. In particular, it ought to be trustworthy for a person who is directly involved in making choices about the deployment or regulation of AI. No one who is directly involved in making choices about the deployment or regulation of AI has right to ignore it, or its likely effects. As a result, Nguyen's analysis is not a good candidate for fleshing out the concept of trustworthy AI, even when we presume that the agent making the judgment of trustworthiness is an entire society.

There is another sense in which Nguyen's analysis of trust might be relevant to AI. Even if everyone on the planet is engaged with AI in some sense, not everyone will be involved in its regulation. We want a world in which most of those people, most of the time, can trust the system of AI regulation, in Nguyen's sense: They can ignore it and get on with their lives. This proposal is entirely compatible with the view developed in this paper. Notice, however, that it does not involve an assessment of the trustworthiness of

the technology itself. It is tantamount to an admission that AI technology is not a candidate for trustworthiness. Rather than placing trust in the technology, we must hope to cultivate trust in the social infrastructure that manages the technology. As mentioned at the outset of this section, the concern in this paper is the technology itself, rather than the social infrastructure in which it is embedded. This proposal, therefore, is entirely compatible with skepticism about the concept of trustworthy AI.

If the foregoing considerations are right, then trustworthiness does not seem particularly suitable to guide our dealings with AI. The suggestion that we should cultivate trust in AI either rests on the false assumption that AI systems have human capacities for social cognition, and is, therefore, an incoherent goal, or it turns out to be dangerously uncritical.

A Non-Anthropocentric Version of Trust?

An Objection

One might object to the argument above by saying that, although Nguyen's analysis of trust is oriented toward artifacts, it nevertheless remains inappropriately bound by intuitions that arise from the contingent peculiarities of interpersonal relationships. According to this line of criticism, within debates about AI ethics, the term *trust* denotes a purely epistemic phenomenon, something rather like reliability. At this point, it will be helpful to narrow the scope of the discussion a bit to ML models, including deep neural networks. Since most of the AI technology that has had recent impact on medicine has indeed been based on ML (broadly construed), this narrowing of scope is not severe. Moreover, it helps us to be more precise about the relevant epistemic properties.

Here, then, is a version of the objection worked out in more detail.

1. The intended meaning of the claim that an ML model is trustworthy is just that it confers epistemic justification on the beliefs it is taken to support.
2. The degree of epistemic justification an ML model confers on the beliefs it is taken to support is amenable to accurate quantification.
3. The quantities used to represent the degree of epistemic justification of an ML model are free from anthropocentric bias.
4. Therefore, the claim that an ML model is trustworthy, when properly interpreted, is not necessarily subject to anthropocentric bias.
5. We, therefore, no longer have any reason to doubt that trustworthy AI (as realized by ML models) is a coherent ethical goal.

This compactly formulated objection naturally skates over a number of complex questions about the nature of epistemic justification. Before a response can be developed, therefore, the claims that comprise the objection deserve to be unpacked and clarified.

Justification is at least usually conceptualized as a property of beliefs. The first clarificatory question that deserves to be answered is about the identity of the agent whose beliefs are under investigation. The answer to this question is straightforward: The agent or agents in question are the human users of the ML model. Although the argument presented here does not presume that it is *impossible* for a machine or a program to have beliefs, that is not the proposal currently under consideration. For the present purposes, the predictions of the ML model are to be regarded as a potential source of justification for the beliefs of the human users.

The second clarificatory question that deserves to be answered is about the quantities mentioned in premises 2 and 3. The two quantities most often used to represent the degree of epistemic justification in ML are reliability and robustness. It is difficult to define either concept precisely because their meanings shift considerably from one scientific context to another. Roughly speaking though, reliability is a measure of how often a model gets its classifications correct within a given distribution of testing data, and robustness is a measure of how well the model's performance stands up to perturbations in that distribution. Robustness and reliability are primary examples of a family of quantities that make up what

we will call the *quantitative error profile* of an ML model, which is a quantitative summary of the model's errors. The term *quantitative* refers to the frequency of errors made, rather than the more general fact that numbers are involved in the description. That is, the quantitative error profile reflects a collection of strategies for counting errors, rather than describing what they are like. A quantitative error profile in the relevant sense need not be restricted to machines. One can also speak clearly about a scientific expert who has amassed a long track record of predictions in some well-specified domain, as having a quantitative error profile.

With the notion of a quantitative error profile on the table, a response to the objection above can now be developed. The first part of the response says that reliability and robustness are indeed valuable properties, and that, ceteris paribus, a model is more desirable if it has a better quantitative error profile. The second and more critical part of the response, which will now be worked out in some detail, says that the objection misconceives the nature of human epistemic justification in the context of ML models. To see this, note that premise 2 says that the justification that we humans have for believing a claim on the basis of the judgment of an ML model is amenable to accurate quantification. The phrase "accurate quantification" suggests that the degree of justification is exhausted by the quantitative error profile of the model in question. Premise 2, therefore, expresses a version of the epistemological doctrine of reliabilism.[11] With respect to epistemic justification, reliabilism says that the justification for a belief is a function of the reliability of the historical process that generated that it,[12,13] where *reliability* means the quality of the quantitative error profile.[14] In this context, the belief in question is the claim that derives from the classification judgment made by the ML model, and the process that generated it is an activity of the classification algorithm executed on a trained model.

To make the connection to the doctrine of reliabilism more concrete, imagine a scenario in which a human expert and an ML model make mutually exclusive judgments. For example, an oncologist says that a cancer patient is eligible for chemotherapy, but a medical AI given the same patient data says that she is not. Imagine, furthermore, that the quantitative error profile of the medical AI is better than that of the oncologist. Perhaps experience indicates that the oncologist gets 950/1,000 judgments right, whereas the medical AI gets 990/1,000 judgments right. Then, the medical AI is, in the sense relevant to epistemology, more reliable than the doctor. Does it follow, necessarily, that the epistemic justification for the claim that the patient is ineligible for chemotherapy is stronger than the claim that she is not? According to the reliabilist interpretation of premise 2, the answer is yes. The quantitative error profile of the medical AI is better than that of the oncologist, and that is the only fact that is relevant to determining which claim is more justified.

A Response

In order to respond to the objection developed in the previous section, an argument that casts doubt on premise 2 shall now be developed. The argument draws on recent work by Paul Humphreys, who is interested in the predictive failures of deep neural networks.[15,16] His discussion focuses on image recognition in convolutional neural networks (CNNs), but the reasoning can be generalized to any multiclass classification problem in ML. CNNs take images as input and deliver probability distributions over labels as output (at least in the typical setup in which a Softmax classifier is used as the output layer.). If you give a CNN a picture of an umbrella, it will output the label "umbrella" along with a probability that describes its confidence in that prediction. It will also ascribe a probability to all the other labels on which it is trained. Contemporary CNNs can now perform such classification tasks as well as humans, and they can do it on data sets that include many thousands of images. However, as is now widely appreciated, CNNs are also susceptible to "adversarial attack."[17] An adversarial attack is a carefully engineered intervention on the input data that delivers a new set of images that look nearly identical to the originals, but which forces the ML model to make substantial classification errors it otherwise would not make.

Famously, some adversarial attacks result in very strange classification errors. Perhaps the most famous example of such an error, originally published in Ian Goodfellow's 2014 paper, is a photograph of

what is obviously a panda bear, which, after having been adversarially manipulated, is classified as a gibbon with 99% probability. In his argument, Humphreys offers a hypothetical example in which a person, rather than a machine, makes similar error. In Humphreys' argument, a human called *Roger* appears at first to be extremely reliable image classifier. His performance on classification tasks prompts us to believe that he is knowledgeable about the kinds of things depicted in the photographs. Occasionally, however, Roger makes an extreme error. In Humphreys' example, Roger examines photographs of cubes, and is nearly always correct. When shown a picture of a cube, he labels the picture as a cube, and when given a picture of something else, he says it is something else. Occasionally, however, when an error does crop up, Roger labels the cube as something completely different, like, for example, a hippopotamus. Humphreys suggests that this kind of howler would undermine our previous confidence in the claim that Roger's beliefs about cubes are well justified. Similarly, when a convolutional deep neural network makes an error of that sort, we ought to lose confidence in the suggestion that beliefs that derive from the judgments of an ML model are well justified.[18]

The thought experiment about Roger is merely the setup behind Humphreys' epistemological argument. Although the argument itself will not be reconstructed here, it is important to articulate the conclusion that Humphreys hopes to establish. Humphreys' conclusion is the view that epistemic reliabilism, as it is typically conceived, is not an appropriate epistemological standard to use when reasoning about neural networks. For our purposes, the point that Humphreys' thought experiment illustrates is that, in a domain in which such radical errors are possible, the degree of epistemic justification for a claim is not exhausted by the quantitative error profile of the model that supports it. In addition to the quantitative error profile, the nature and magnitude of the error must be taken into account. As a result, the degree of epistemic justification an AI model confers on the claims it is taken to support is not amenable to accurate quantification. Hence, premise 2 in the objection above is, therefore, false.

The argument just presented highlights the fact that there is an information asymmetry between the error profiles of ML models, on the one hand, and the error profiles of humans on the other. Cognitively normal humans do not make mistakes of the sort Roger is described as having made (if they did, we would suspect that they suffer from a psychiatric disorder). We humans make many kinds of errors of course, and we make them frequently. Nevertheless, the kinds of errors that normal humans make under favorable epistemic conditions are typically not as alien[19] as the cube/hippopotamus confusion.

This difference is the lynchpin in our response to the objection above. When reasoning explicitly about human knowledge, we can usually suppress or disregard information about the nature and magnitude of typical errors because that information is woven into our shared cognitive backdrop. After all, we ourselves are prone to errors of the same sort. By contrast, when extracting knowledge from ML models, it becomes necessary to represent and reason explicitly about the nature and magnitude of model errors. We cannot simply extract the degree of epistemic justification for a given claim from the quantitative error profile of the model that made the associated judgment. The claim that the degree of epistemic justification can be quantified by means of the quantitative error profile reflects a willingness to suppress information about the nature and magnitude of error in the ML model in the same way that we routinely do in the human case. However, in the ML case, this suppression of information about the nature and magnitude of errors is unwise because the ML model does not share our distinctively human cognitive backdrop. This willingness to suppress error information reflects a tacit assumption that ML classification judgments work according to the same principles as human ones. However, this is itself an unjustified form of anthropocentrism.

The argument just presented shows that even the purportedly non-anthropocentric, purely epistemic interpretation of what it means to trust AI is subject to a subtle form of anthropocentrism. Since both the interpersonal and the purely epistemic interpretations of trust rest on unjustified anthropocentric assumptions, the dilemma from the "Troubles with Trustworthiness" section remains intact. Depending on which notion of trust one has in mind, the goal of "trustworthy AI" is either dangerous because it pushes us to adopt an unquestioning attitude, or incoherent, because it rests on one of two anthropocentrically biased assumptions.

Strange Error

Thus far, the central claim has been that trustworthy AI is not the goal around which efforts at developing AI ethics ought to be centered. At this point, it might be tempting to say that it would be better to cultivate an attitude of *distrust* toward AI models. However, that recommendation is equally anthropocentric, and equally mistaken. What is needed instead is a careful and rigorous approach to error analysis. In this section, it is shown that the kinds of errors described in the last section make analysis unusually tricky, in a way that will have consequences for ethical reasoning about the use of AI. To develop this line of thinking, the concept of strange error will be introduced. The concept of strange error is similar to, but not identical to, the concept of adversarial examples.

Our definition of strange error is as follows: Strange errors are errors that (1) result from perturbations to the input data that are either unnoticeable to humans, or otherwise strike them as irrelevant to the classification task, and (2) would strike humans as radically incorrect, if they knew the ground truth.

Let us call the first property subtlety and the second property radicality. Adversarial examples (as they are typically understood) are also subtle, but need not be radical. Moreover, whereas adversarial examples are, by definition, intentionally engineered to produce classification errors, strange errors need not be produced by intentionally engineered perturbations to the input data. They can also occur accidentally. A good example of this comes from the ImageNet data set. ImageNet is an enormous database of images, most of which are scraped from the Internet. CNNs are often trained with ImageNet data and can generally perform very well on it. However, there is a subset of the ImageNet database called *ImageNet-A*, which is constituted by a subset of ImageNet images that were painstakingly selected by hand precisely because they generate radical errors (Figure 1).[20]

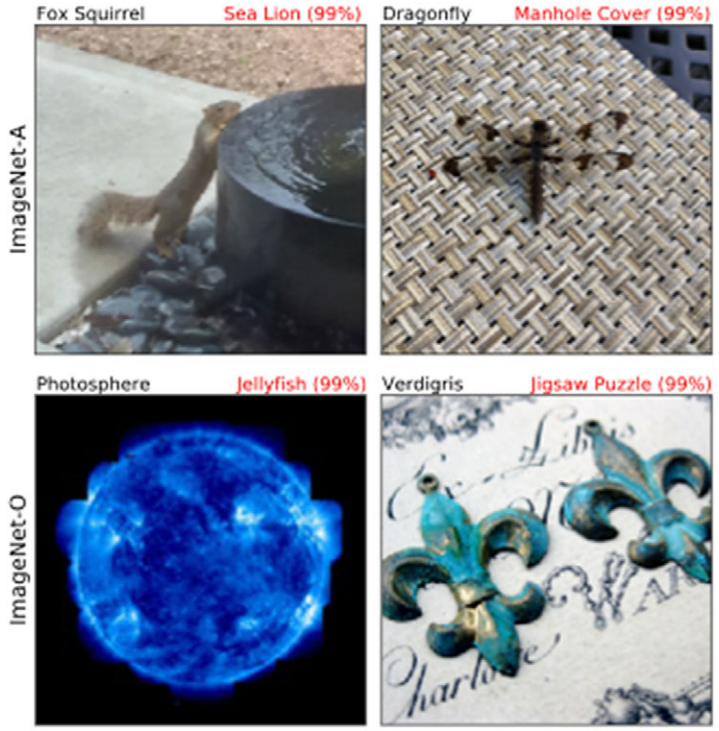

Figure 1. Examples of naturally occurring images that prompt radical errors when given as input to ResNet-50, a standard convolutional neural network. The horizontal black labels refer to the correct label. The red labels refer to the highest probability classification by ResNet-50. ImageNet-O is another subset of ImageNet, selected according to slightly different parameters than ImageNet-A. Taken, with permission, from Hendrycks et al. 2021 (see note 20).

How much do we need to be worried about strange errors in medicine? Some prominent examples of medical AI involve binary judgments about the presence or absence of a disease. In such cases, it is not at all clear that a mistake can be radical. However, medical AI is increasingly used on multiclass classification problems, in which many labels are possible for each datapoint. In multiclass classification problems, the space of possible answers is large, and consequently contains room enough for radical errors to crop up. Imagine an AI tasked with diagnosing neurological problems. It is trained with supervised learning on imaging data to identify hundreds of possible neurological problems. In such a case, there are as many possible labels as there are diseases. Imagine, furthermore, that someone is suffering from a brain tumor, but that a medical AI, as a result of some inscrutable pattern in the input data, gives a diagnosis of meningitis. That would be an example of a strange error in medicine. Moreover, medical AI is not restricted to image processing. Natural language processing on electronic health records is also being used to predict the probability of an open-ended list of medical problems.[21] In these cases, radical errors are to be expected.

The problem of strange errors in medicine is reinforced by considerations regarding the medical infrastructure in which ML models are used. Finlayson et al. argue that, because medical data are balkanized, and because hospital infrastructure is hard to update, medical AI is acutely and increasingly susceptible to adversarial attacks, some of which will produce strange errors in our sense.[22] If Finlayson et al. are right, medical AI will be confronted with many strange errors in coming years. This fact fortifies the argument in the previous section because, in a regime in which strange errors are common, quantitative error profiles will be a particularly impoverished guide to epistemic justification.

Is there anything we can do defend medical AI against the risk of strange errors? Cameron Buckner argues that, in order to understand adversarial examples, we need a theory of artifacts in deep learning.[23] Buckner's argument can readily extended to strange errors in our sense. If we had a theory about the patterns in input data that tend to generate strange errors, we could work to anticipate particular kinds of strange errors, and thereby mitigate associated risks. Thus far, however, and despite a surge in recent literature on strategies for defense against adversarial attack, there does not yet appear to be any general algorithmic solution to the problem.

Culpability and Complexity

Thus far, the concept of strange error has been defined, but nothing has yet been said anything about the concept of risk. We use the term "risk" in its informal sense, rather than in its decision-theoretic one. According to the informal sense of the term, risk can be contrasted with uncertainty along a dimension of objectivity. Uncertainty has a subjective connotation, whereas risk has an objective one. Consider the following example from Sven Hanson. You see a snake in the grass, which you suspect may be poisonous. In fact, it is not. You are then uncertain about the snake's ability to poison you, but you are not actually at risk of being poisoned.[24] The objective connotation of the term "risk" is appropriate for present purposes because one can be in position to know that a model is disposed to make strange errors, despite the fact that, precisely because they are strange, one cannot know much about what they will look like when they do crop up.

Risk is, therefore, an epistemic situation in which one knows that the model on which our beliefs are based is disposed to make errors, and the consequences of these errors might be severe. Notice that this definition does not exclude the possibility of non-strange errors. Typically, the users of an ML system will be faced with both strange and non-strange errors. With respect to the non-strange errors, the usual decision-making practices are applicable. One could, for example, resort to normal cost–benefit analysis.

It is now time to turn to the question of how strange risk influences moral culpability. Culpability is the condition in which a person or other agent is subject to judgments of praise and blame. Are the engineers who build an ML system culpable for its errors? According to most philosophical analyses of culpability, it involves an awareness condition. That is, you must be aware of both the probable consequences of your action and the fact that those consequences have moral significance. One must then ask: Are the engineers who build ML systems aware of the errors their models are likely to make, and

of their moral significance? If the model is disposed to make strange errors, then the answer is "no." As was argued in the "A Non-Anthropocentric Version of Trust?" section, neither the nature nor the magnitude of strange errors can be systematically anticipated.

Why exactly can you not anticipate the nature and magnitude of strange errors? Recall that one of the two defining properties of strange errors is radicality and that radical errors are possible only when the space of possible classifications is large. Strange errors occur because humans conceptualize that space differently from machines. For humans, two classifications in that space are close together if they are similar to one another (elephant and hippopotamus). The similarity gradient is, of course, a function of our own psychology. ML systems might replicate that similarity gradient under some conditions,[25] but this is not something one can expect in general. Strange errors occur when an ML model orders some subspace of classifications very differently, such that what counts as a highly probable second-best classification for us is radically different from what counts as a highly probable second-best classification for the ML model (cube and hippopotamus).

It is this difference in the way humans and machines order conceptual space that gives rise to the radicality property. On its own, the radicality property makes anticipating the nature and magnitude of strange errors difficult. Once we take the subtlety property into account, the prospects for anticipating the nature and magnitude of strange errors decline further. To say that strange error is subtle is to say that human perceptual judgments cannot easily identify the input conditions under which such radical mistakes are likely to occur. Together, these facts make it practically (if not logically) impossible to anticipate the nature and magnitude of strange errors.

It is nearly a platitude in technical articles on AI ethics that ML models are subject to unknown failure modes which appear only after deployment. In other words, unpredictable mistakes occur once the model is applied to real-world data. The concept of strange error helps refine this idea and helps clarify its ethical consequences. When reasoning about the ethics of these unknown failure modes, one must ask *why* the failure mode was unknown, and whether it could have been known before deployment. These questions are important because, unless they are answered, it will be impossible to say who, if anyone, is culpable for the error. Considering the discussion of strange error, one can say that the nature and magnitude of the errors could not have been anticipated. Moreover, the practical impossibility of anticipating these factors does not stem from the cognitive limitations of any particular person[26] or group. Unless and until we have a theory of ML artifacts, of the sort Cameron Buckner calls for, there exists neither person nor group that can be expected to have anticipated the nature and magnitude of strange errors.

One positive proposal worth considering, which is pre-figured in Paul Humphreys' talk, is to impose something like a strangeness-limiting mechanism in the deployment of medical AI. Humphreys suggests an idea along these lines as an amendment to traditional reliabilism, in order to make it more appropriate for an age in which scientific work relies heavily on ML techniques. Epistemology and ethics have been closely linked not only since the introduction of AI. It is, therefore, not difficult to convert Humphrey's theoretical suggestion into a practical one. His suggestion is that we develop an explicit measure of strangeness, which we extrapolate from subjective judgments of similarity, and then use that explicit measure to delimit the space of models that count as epistemically reliable. More precisely, where g is the function that describes the input–output mapping of an ML model, and X is the input for that model, we insist that there exists no X such that $g(X)$ is farther from an accurate representation of the input data than δ. Here, δ refers to a conventional threshold within the similarity space, which itself is derived from collective judgments of the relevant scientific community. The ethical variant of this epistemological criterion is to say that we must only deploy ML models that meet this strangeness-limiting constraint. This is because risks can be minimized in this way and, in particular, the danger of serious harm from medical AI systems can be avoided.

This strangeness-limiting constraint has intuitive appeal, but faces at least two difficulties. First, although it is perfectly coherent to say that a model confers justification only if the strangeness of its judgments is bounded, one can never be in position to know with certainty that the strangeness of its judgments is bounded. In other words, one can never be in position to know that there exists no input X such that $g(X)$ is farther from accuracy than δ. This is because any such claim would be based on inductive

inference from a limited sample, and inductive inference is inherently risky. In a practical setting, the best one can do is to say that no judgments have yet been observed that exhibit a degree of strangeness greater than δ, despite having tested the model on many data sets. It is unclear whether this suggestion amounts to a different strategy than the one we already follow. We try to test our ML models rigorously, but, nevertheless, they continue to surprise us with the occasional strange error.

The other difficulty is that, if, contrary to the deflationary reaction expressed in the previous paragraph, we do end up ruling out a substantial number of ML models because they fail to satisfy our strangeness-limiting constraint, we would almost certainly end up discarding models that are accurate and strange for those that are non-strange, but also less accurate. This kind of trade-off is difficult to justify in cases where accuracy has greater influence on the net utility of medical judgment than does the absence of strangeness. It seems, therefore, that, unless and until we have a rigorous theory of ML artifacts of the sort Cameron Buckner envisions, we will have to live with strange errors and the resulting risks.

The argument of this section has been that the people who would otherwise be culpable for the negative consequences of AI error are not, in fact, culpable because they cannot have known the nature and magnitude of the errors involved. A natural response to this claim is that ignorance does not always undermine culpability. In some circumstances, a person may be culpable for a negative outcome even if they were ignorant of the possibility of that outcome. *Ignorant culpability*, as it is sometimes called, can occur when the person neglected to undertake some action that would have relieved them of their ignorance. Following Smith,[27] we can refer to such neglected knowledge gathering activities as *benighting omissions*. Under strange risk, there are no benighting omissions. Absent a systematic theory of adversarial examples, there is nothing one could have done to gather knowledge of the nature and magnitude of strange errors.

From this, we conclude that the primary pathway by which ignorance leads to culpability is blocked under conditions of strange risk. Must we therefore give up on the notion of culpability altogether when operating on the basis of ML judgments? This view has been advocated by the so-called "fatalist" camp in discussions of responsibility gaps generated by ML technology,[28] and is exemplified by the original paper on responsibility gaps.[29] However, there are at least two reasons to look for a more nuanced position.

First, not all errors are strange. Even if an ML model is epistemically opaque, and therefore generates behavior that is resistant to mechanistic explanation, you can still study its behavior, and look out for worrisome patterns. This is indeed what you have to do when you construct the quantitative error profile of the model. If your error analysis is not thorough, or if you do not allow that analysis to constrain your choices about how to deploy your model appropriately, you bear some culpability for whatever non-strange errors that do crop up.

Second, even if you cannot predict the nature and magnitude of model errors, you can still say something about them. Even in the case of strange errors, you can study the model's quantitative error profile. Moreover, once you have read this paper and know that strange errors occur, you can incorporate that knowledge into your risk analysis. You are then culpable for taking appropriate account of your own second-order uncertainty about the nature and magnitude of the errors your model will make.

In the medical context, this means that software engineers as well as physicians and operators of medical facilities using AI will have to grapple with the existence of strange errors. In the future, informed consent may even need to involve information about strange error. Of course, it will be an (additional) challenge to communicate this concept to lay people. Patients need to know what they are getting into when AI is used in medicine. Only then can they decide for themselves whether they want to use this technology or not. If the problem of strange errors is concealed, then developers, operators, and users are culpable if harm occurs that patients could not have foreseen.

For these reasons, it is unwise to make any absolutist claims about culpability in the domain of AI. Nevertheless, there is no denying that ML in general, and strange error in particular, make it rather difficult to say who is culpable for what. On top of the fact that strange errors make it practically impossible to anticipate the nature and magnitude of error, it is also the case that AI models are not constructed by a single person. In medical cases, they are not even constructed by a single institution. As Finlayson et al. note, most neural network models used in medical AI are developed by some AI company

or research group for general classification tasks, and then adopted and fine-tuned by an independent group for medical applications. Medical AI is, therefore, particularly susceptible to the so-called many-hands problem, according to which liability cannot be pinned to any particular person. Given the combination of strange error and the many-hands problem, it would be wise to explore options for restitution that do not depend on identifying a culpable party, but this should be done in parallel to attempts to assess culpability, and to hold bad actors liable. Strange error makes culpability assessment more complex, and less certain, but cannot undermine it altogether.

Although we prefer to speak of "culpability," rather than the more general notion of "responsibility" in this context, the conclusion that has just been drawn does entail a view about responsibility gaps. According to the view defended here, it is unlikely that responsibility gaps will be generated in which no one has any culpability for harm. However, the view here is also compatible with John Danaher's proposal that ML models may generate what he calls a *retribution gap*.[30] A retribution gap is a situation in which the victim of some harm wants retribution, but cannot identify any particular party deserving of that retribution. Danaher may well be right that, as ML systems become enmeshed in our practical life, harms will appear for which retribution is not a coherent goal. As we learn to live with strange errors and risks resulting from them, we may be compelled to give up some of our retributive habits, even if we retain the capacity to ascertain various kinds of culpability.

Conclusion

This paper has shown that trustworthy AI is not the best goal around which to organize our efforts to develop an ethics for AI. Even when interpreted as a purely epistemic concept, the goal of constructing trustworthy AI is subject to a subtle form of anthropocentrism, embodied in the assumption that the errors of ML systems are generated by the same kind of process underlying human error. It is practically impossible to anticipate the nature and magnitude of strange errors, and this fact puts considerable pressure on our standard practices of assessing moral culpability. This pressure is all the more salient in medical contexts, where relations of trust between patient and clinician are destined to remain ethically significant, even as we are forced to search for other standards by which to regulate our relationships with AI.

Acknowledgments. We would like to thank Thomas Grote and Juan Durán for insightful comments on an earlier draft. We would also like to thank participants at the INM-7 Research Colloquium in Jülich, and at the XAI 4 Workshop at the Delft University of Technology, for posing excellent questions and partaking in what we took to be an unusually fruitful pair of discussions. Finally, we would like to dedicate this paper to the memory of Paul Humphreys, who died rather suddenly while this paper was being written. Paul discussed the material in this paper with us informally, but it would have been much improved if he had had the chance to comment on the written version.

Notes

1. AI HLEG. Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence; 2019.
2. See note 1, AI HLEG 2019, at 4.
3. Baier A. Trust and antitrust. *Ethics* 1986;96(2):231–60. doi:10.1086/292745.
4. Jones K. Second-hand moral knowledge. *The Journal of Philosophy* 1999;**96**(2):55–78. doi:10.2307/2564672.
5. McGeer V, Pettit P. The empowering theory of trust. In: Faulkner P, Simpson TW, eds. *The Philosophy of Trust*. Oxford: Oxford University Press; 2017:14–34.
6. Shevlin H, Halina M. Apply rich psychological terms in AI with care. *Nature Machine Intelligence* 2019;**1**:165. doi:10.1038/s42256-019-0039-y.
7. Ryan M. In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics* 2020;**26**(5):2749–67.

8. We were not aware of Ryan's paper until after submitting this article for publication. Thanks to Thomas Grote for making us aware of it.
9. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Computational Surveys* 2021;**54**(6):1–35. doi:10.1145/3457607.
10. Christian B. *The Alignment Problem: Machine Learning and Human Values*. New York: WW Norton & Company; 2020.
11. The objection above is framed in terms of epistemic justification, which, according to some philosophers, is incompatible with reliabilist epistemology. This is because reliabilism is sometimes conceptualized as a complete departure from the older Western tradition in which having epistemic justification for a proposition is a matter of having introspective access to reasons for that proposition. Nevertheless, some prominent reliabilists, such as Alvin Goldman, are happy to speak of the empirical basis for knowledge in terms of justification, as we do here. In any case, it would not be difficult to reconfigure our argument so as to eliminate appeal to the concept of justification.
12. Goldman A. What is justified belief? In: Pappas GS, ed. *Justification and Knowledge: New Studies in Epistemology*. Dordrecht: Reidel; 1979:1–25.
13. Durán JM, Formanek N. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines* 2018;**28**(4):645–66.
14. Sometimes, epistemological reliabilism is concerned only with the actual track record of the process in question. However, many strands of reliabilism are designed to build on the slightly older tradition of causal justification for knowledge, which emphasized so-called "safety and sensitivity conditions," which are expressed as counterfactuals. These counterfactual conditions are closely tied to the concept of robustness in ML. Thanks to Eva Schmidt for pointing out this connection.
15. Humphreys P. *Predictive Failures in Neural Nets. Lecture Series in Evidence, Model and Explanations. Philosophy of Science India*; 2020; available at https://www.youtube.com/watch?v=2VFPXbrCqzM (last accessed 1 June 2022).
16. Paul Humphreys had written a mature, but unpublished draft of a paper that corresponds to the talk referenced above. He discussed that paper with us in personal communication but was not quite ready to share a draft. Sadly, he developed brain cancer during the summer of 2022 and died quite suddenly. Since the manuscript is now destined to remain unpublished, the best we can do is cite the online talk.
17. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples; 3rd International Conference on Learning Representations; 2015.
18. Here, a comparison with the view of Durán and Formanek 2018 (see note 13) is instructive. They focus on whether epistemic reliabilism can handle the problem of epistemic opacity, and they argue skillfully that it can. We take everything we say here to be compatible with their arguments, which suggests that the problem of strange risk is fundamentally different from the problem of epistemic opacity.
19. Heinrichs B, Knell S. Aliens in the space of reasons? On the interaction between humans and artificial intelligent agents. *Philosophy and Technology* 2021;**34**(4):1569–80.
20. Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D. Natural adversarial examples. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Institute of Electrical and Electronics Engineers (IEEE); 2021:1526215271.
21. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports* 2016;**6**(1):1–10.
22. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019:**363**(6433):1287–9.
23. Buckner C. Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence* 2020;**2**(12):731–6.
24. Hansson SO. Risk. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University; 2007, available at https://plato.stanford.edu/archives/fall2018/entries/risk/ (last accessed 1 June 2022).

25. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 2014;**111**(23):8619–24.
26. Van de Poel I, Royakkers LM, Zwart SD, De Lima T. *Moral Responsibility and the Problem of Many Hands*. New York: Routledge; 2015.
27. Smith H. Culpable ignorance. *The Philosophical Review* 1983;**92**(4):543–71.
28. Santoni de Sio F, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology* 2021;**34**:1057–84.
29. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 2004;**6**(3):175–83.
30. Danaher J. Robots, law and the retribution gap. *Ethics and Information Technology* 2016;**18** (4):299309.

Cite this article: Rathkopf C and Heinrichs B (2024). Learning to Live with Strange Error: Beyond Trustworthiness in Artificial Intelligence Ethics. *Cambridge Quarterly of Healthcare Ethics* 33: 333–345, doi:10.1017/S0963180122000688