Over What Range Should Reliabilists Measure Reliability?

Buijsman, Stefan

DOI
10.1007/s10670-022-00645-4

Publication date
2023

Document Version
Final published version

Published in
Erkenntnis

Citation (APA)
Buijsman, S. (2023). Over What Range Should Reliabilists Measure Reliability? *Erkenntnis*, *89*(7), 2641-2661. https://doi.org/10.1007/s10670-022-00645-4

Erkenntnis
https://doi.org/10.1007/s10670-022-00645-4

ORIGINAL RESEARCH

Over What Range Should Reliabilists Measure Reliability?

Stefan Buijsman[1]

Received: 23 February 2022 / Accepted: 19 November 2022

Abstract
Process reliabilist accounts claim that a belief is justified when it is the result of a reliable belief-forming process. Yet over what range of possible token processes is this reliability calculated? I argue against the idea that *all* possible token processes (in the actual world, or some other subset of possible worlds) are to be considered using the case of a user acquiring beliefs based on the output of an AI system, which is typically reliable for a substantial local range but unreliable when all possible inputs are considered. I show that existing solutions to the generality problem imply that these cases cannot be solved by a more fine-grained typing of the belief-forming process. Instead, I suggest that reliability is evaluated over a range restricted by the content of the actual belief and by the similarity of the input to the actual input.

1 Introduction

Process reliabilists (which I will shorten for convenience to ‘reliabilist’ in this paper) hold that a belief is justified when it is the result of a reliable belief-forming process (Goldman, 1979). In order to evaluate whether a belief is justified, we thus need to know two things: which process type was used to form the belief (a familiar issue known as the Generality Problem) and how reliable that process is. My focus here is on the latter question, which in turn can be split into two parts. First: what set of token processes needs to be considered to determine reliability? Is it the entire range of token processes in the actual world? (which seems to be the standard assumption, and when I write about all possible inputs I intend to index these to the actual world unless otherwise specified). Is it instead, as Henderson and Horgan (2006) argue, a range of token processes in both the actual world and different possible worlds? Or is it a much smaller

✉ Stefan Buijsman
s.n.r.buijsman@tudelft.nl

[1] TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 04 January 2023

range of token processes? That is the question I will focus on here, looking specifically at process types for which we have very detailed information about their reliability, namely testimony from machine learning algorithms (i.e. AI algorithms). I will leave aside the second question: what calculation is used over the set of token processes to score the degree of reliability? Instead of engaging this question, which can be found e.g. in the discussion around justified credences (Dunn, 2015; Pettigrew, 2021), I will assume for simplicity that the ratio

$$\frac{\text{token processes producing true beliefs}}{\text{total token processes}}$$

suffices (e.g. if the process type leads to true beliefs 8 out of 10 times, then 8 out of 10 token processes produce true beliefs and reliability is 80%).

That leaves the question: which token processes should count as part of this ratio? It is rarely discussed, but it seems that the standard assumption is that all token processes in the actual world count, though the indexation is sometimes argued to be to the possible world where the process was used, or to a more intricate construction of possible worlds. In all these cases, however, every token process in the subset of possible worlds is to be considered. Henderson and Horgan (2006) explicitly argue against a more local evaluation of reliability and Lyons (2019) likewise briefly discusses and rejects the idea that reliability might be evaluated over a subset of token processes. Yet I argue in Sect. 2 that, based on belief-forming processes where one simply believes what an AI system outputs (i.e. testimony from AI systems), there is a case to be made for the need to evaluate reliability over a smaller set of token processes, for all process types. After presenting my positive argument for this claim, I consider a possible objection, namely that the process type in question should be narrower, and that the evaluation is still over all tokens, but only those falling under the narrower type. To forestall that objection I look at different ways of typing the belief-forming process (i.e. different answers to the Generality Problem) in Sect. 3 and argue that none of the currently acceptable answers fit with the fine-grained typing approach to the AI testimony cases I present. After that, I consider a few other objections and views from the extant literature on the range of evaluation in Sect. 4, showing that this literature is largely tangential though there is one view (by Graham, 2012) that gets close to covering the AI case by pushing for local evaluation under ‘normal circumstances’. That, I argue, is hard to specify more precisely in the AI context. I also argue against restricting the evaluation to typical inputs; those the process could easily have in nearby possible worlds. If these solutions are unavailable, how then should the smaller range I argue for be determined? I close by considering possible answers to this question in Sect. 5.

2 Arguing for Small Ranges of Reliability

The basic format for the cases I discuss here will be that of a user acquiring beliefs based on a machine learning system.[1] As Wheeler (2020) argues, these cases can be understood as testimony that a user receives from the AI system. To really isolate this process of testimony, I consider only the case where the user is presented with very little information aside from the system output. They were probably notified that the system was tested and found to be accurate, but often lack more detailed knowledge or the time/opportunity to verify AI outputs for themselves. This is in fact quite common in cases where users receive testimony from AI systems [see e.g. criticisms of epistemic dependence and the phenomenon of 'quasi-automation' where there is a human overseeing the outputs of an AI system but merely rubber-stamping these (van den Hoven, 1998; Wagner, 2019) but also the phenomenon of automation bias (Parasuraman and Manzey, 2010)], and thus makes for a realistic case study. I focus on these cases for a few reasons: (1) information about reliability is readily available and process individuation is comparatively straightforward. (2) As opposed to e.g. testimony from humans, there is good reason to think that some of these processes are overall unreliable but locally reliable (see below). Yes, people will lie some of the time, but probably not more often than speak the truth. (3) AI systems are relevant because they conceptually disentangle non-accidental reliability from standard notions of normality and typicality, and as such are a good case to base the discussion in Sect. 4 on. So, while my aim is to argue that the reliability of all belief-forming processes (regardless of whether they involve AI systems) should be evaluated on a smaller range, I will do so by focussing on testimony from AI systems. This is simply one kind of belief-forming process that people may use, and is in my opinion not fundamentally different from other kinds of testimony or belief-forming processes. Consequently, there does not seem to be a reason to restrict the range of evaluation for testimony from AI systems and not for other belief forming processes. I hope the reader will therefore bear with me in the discussions of AI cases (as opposed to more common cases discussed in the reliabilism literature) as they are one of the easiest and most realistic I can think of where we see this issue of local reliability without global reliability.

The case I will present first is, however, a somewhat unrealistic one based on computer vision algorithms, because it is the simplest case with which to illustrate the intricacies around the accuracy of machine learning models. Consider, therefore, a case where a user is sitting behind a computer and is told by an AI system what objects are present in an image (where the user can't see the image herself, so there is no independent way of verifying the computer output).

[1] A machine learning system is one where the algorithm adjusts its parameters based on a set of training data for which the right output is known. For example, linear regression is a machine learning tool, where e.g. parameters a and b in formula $y = ax + b$ are adjusted to minimize the error with respect to the training data. The AI systems under discussion are all deep learning systems (neural networks, to be precise), which have millions to trillions of parameters all adjusted based on a training set.

Is someone in this situation justified in believing that e.g. ‘this image contains a dog’? The reliabilist will, naturally, answer that this depends on whether the belief-forming process is reliable. The designer of the machine learning system will then likely answer that when the algorithm was tested, on a so-called validation set (a portion of data set aside at the beginning to verify, after the training phase where the parameters of the model are adjusted, how well it works on new inputs), it had an accuracy of x%. For a state-of-the-art system this tends to be high (> 90%), though complete accuracy is avoided as this typically means worse performance on new inputs (‘overfitting’). In any case, based on that figure it seems that the model is reliable and so the belief should be justified. Note, however, that this accuracy score is based on a fairly small range of token processes, where a token process is (simplified into) an input–output pair for the algorithm and the question is whether the output is correct (and so leads to a true belief in the user).

Is this accuracy score the reliability measure that the reliabilist is after? It would seem natural to desire a measure based on more than just a small set of possible inputs, and so to measure reliability over all possible inputs of the algorithm. That would seem to exhaust the possible token processes (if we abstract away from the exact time the user sees the output, etc., which is in fact not possible) and so give a much better idea if the process leading to ‘this image contains a dog’ is sufficiently reliable to lead to a justified belief in the user receiving the testimony. The issue is, however, that when considering *all* possible inputs virtually every machine learning algorithm will be unreliable. There are several reasons for this.

First, the input for a computer vision algorithm is a range of values corresponding to the pixels of the image that is analysed. A large number of these possible images (e.g. one containing only static) will be nonsensical. While it is possible to add an output category ‘nonsense’/‘I don’t know’ to an algorithm, this typically doesn’t happen. Even if one were to add one a large number of incorrect outputs is likely to slip past for the nonsensical inputs. The simple reason is that it is difficult to find a good way to score how confident the model is in a prediction (Guo et al., 2017; Luppers et al., 2020), and so computer vision models will likely have a hard time classifying nonsensical inputs as such. We see this, for example, with adversarial attacks where a normal image is changed slightly [possibly just a single pixel; Su et al. (2019)] and the algorithm gives a dramatically different output. For example, an image of a yellow bus is recognized as an ostrich after small perturbations, or an image containing three dogs leads to the outputs ‘train (0.95), person (0.53)’ (Akhtar and Mian, 2018).

Second, even for inputs that correspond to realistic photographs, computer models can be seriously wrong. Alcorn et al. (2019) tested a computer vision model on images of vehicles presented both straight on and rotated in different poses. For the typical positions the model performed perfectly, recognizing for example a school bus and motor scooter with 100% confidence. Yet rotate them somewhat, e.g. in a position where the motor scooter makes a wheelie, and the model suddenly saw (still with a high confidence) a parachute. Similarly, an overturned school bus was classified as a snowplow and an overturned fire truck as a bobsled. In fact, 97% of possible poses for these vehicles were incorrectly classified (Alcorn et al., 2019). Likewise, natural adversarial examples abound, where e.g. a dragonfly resting on a yellow shovel is classified as a banana, or a mushroom on some wood is classified as a nail (Hendrycks et al., 2021).

Third, there is a general issue about the moment when the model is applied (similar to the temporal generality problem noted by Weatherson (2012) and the historical v.s. current time-slice measurement of reliability already found in Goldman (1979), who opts for a historical theory). It may, namely, be the case that an AI system is based on outdated information: the patterns it uses to determine the output did hold for the training data, but in the meantime the situation in the world has changed and so those patterns are no longer a reliable basis for the output. Known as concept drift (Žliobaité et al., 2016) this introduces a further question regarding the evaluation of reliability: when evaluated historically the model may well be reliable, even if it is unreliable in the current time-slice (because it becomes unreliable after a while). For example, algorithms predicting the demand of products worked reliably before the pandemic, accurately predicting what inventory was required to fill the incoming orders. When the covid-19 pandemic started, however, purchasing patterns changed drastically (we suddenly bought far more hand gel and face masks). The result was AI systems that yielded highly inaccurate predictions. The same thing happened for fraud detection algorithms, which normally would flag people suddenly buying garden equipment as suspicious behaviour, yet this was no longer a good indicator for fraud in the new situation (Heaven, 2020). It is a serious challenge that a model can be reliable for a period of time, and should lead to justified beliefs, yet is highly unreliable at a later time. One can say, then, that machine learning systems will not be reliable when all possible inputs are considered, even though they can be highly reliable on a subset of inputs. Furthermore, even restricting it to all possible inputs that are also realistic images fails to guarantee reliability of these systems.

So, if it is at all possible to acquire justified beliefs based on these machine learning systems, then the reliabilist will have to say that the relevant reliability is not that of the algorithm over all inputs, but instead the reliability over a smaller set of inputs. The question is: can we get justified beliefs from AI testimony? I think one has to answer this with a strong yes. The same computer vision models that are susceptible to these numerous mistakes power self-driving cars, which though far from perfect (particularly in atypical situations) do rather well on highways, and are likewise applied in numerous other scenarios. Furthermore, performance of machine learning models is often better (in circumscribed domains) than that of humans. For example, a simple model to predict house prices can do better than humans on the types of houses it was trained on, yet do worse on other types of houses [and users unaware of how AI works chose to rely on the model *more* in the new, out-of-distribution, cases, see Chiang and Yin (2021) and Poursabzi-Sangdeh et al. (2021)]. My point is: although testimony from AI systems will not be reliable when evaluated over all possible/plausible inputs, it can still be highly reliable over the smaller range of inputs for which they are actually used. Since we do consider it possible to acquire justified beliefs from AI testimony in these situations, it follows that the range of evaluation for reliability should be smaller, and not all possible inputs (in the actual world, or in a different subset of possible worlds) are used to determine reliability.

That is the basic argument: a process where a user simply believes what the AI outputs (i.e. a case of AI testimony) can plausibly lead to justified beliefs in some cases. Yet the AI system is not going to be reliable when measured over all possible inputs in the actual world/a subset of possible worlds, so unless the process type is quite narrow (see the next section) it follows that the reliabilist should evaluate reliability over a subset of token processes falling under the process type. Not just for AI testimony, but generally speaking (cf. also some non-AI examples of other locally reliable processes in Sect. 5). The question of how this narrower evaluation should go (as the relevant subset likely differs per token process) is the subject of Sect. 5. First though, there is the question whether I do not type the belief-forming process too broadly in my argument.

3 Fine-Grained Process Typing as an Answer?

One way to deal with the differences in reliability mentioned above is to look at the typing of the belief-forming process. While 'believing the output of AI system X' may not be a reliable process, perhaps the more fine-grained 'believing the output of AI system X in circumstances C at times T' is reliable. For each of the cases where I argue that a more local evaluation (i.e. not across all possible token processes) is required it is equally possible to say that a more fine-grained process type should be named, for which reliability is evaluated globally. Is that a viable response? In theory it of course is, but as I argue in this section it doesn't mesh well with existing accounts of process typing, developed in answer to the generality problem. Most of these appeal to cognitive factors to determine the right typing, so I start with this array of answers. One exception to that rule is Beebe (2004), who looks at all statistically relevant factors. I treat his account last, as it is potentially problematic for my view, but argue that precisely that fine-grained typing gets it into trouble. It is therefore unlikely that my argument for more local ranges of evaluation can be answered by arguing for more fine-grained process typing.

3.1 Cognitive Approaches

Treating them in historical order, Comesaña (2006) has presented a solution to the generality problem (i.e. a principled way of typing belief-forming processes) using the basing relation, of a belief being based on certain evidence. So, the view defended specifically is:

> *Well-Founded Reliabilism*: A belief that p by S is epistemically justified if and only if:
>
> (i) S has evidence E;
> (ii) the belief that p by S is based on E; and
> (iii) the type *producing a belief that p based on evidence* E is a reliable type. (Comesaña 2006, p. 38)

Applying this to the case of AI testimony, where a user of an AI system forms beliefs by simply believing whatever the model outputs, we get a fairly general type. The evidence will be ‘the AI says that *p*’, and so the process type is *producing a belief that p based on the AI outputting that p*. No distinction can be made here on the inputs that are fed into the AI system, as the evidence on the basis for the resulting belief is always the same: the AI provides output *p*. Of course, there might be further evidence, e.g. that the specific case is an adversarial case (and so less reliable) or that the system has been tested and found reliable for cases such as the one in question, but this would be additional information. The simple scenario where the user has no information other than that an AI system is provided and gives a certain output is certainly possible, happens in practice, and is typed by this account in a manner that forces us into the more local evaluation of reliability I argue for.

Essentially the same typing comes out of the cognitive convergence account presented by Olsson (2016) and Jönsson (2013). They argue that there is a wide-spread convergence in the way people type belief-forming processes, something which Jönsson (2013) also tests empirically. The idea is that we tend to describe belief-forming processes using the same verbs, and that the description we choose in this way gives the relevant type for the reliabilist. Specifically, (Jönsson 2013, p. 264) settles on the following specification:

> (J^{AC}) A believer B is justified in believing p relative an attributor A to the degree j iff $j = f(r_{AC}(t_{AC}(bfp)))$, where *bfp* is the token-process that led B to believe that *p*, t_{AC} is either the type in terms of which A is actually thinking about *bfp* at the time of attribution (J_{ACa}), or the type in terms of which A is disposed to think about *bfp* at the time of attribution (J_{ACb}), and r_{AC} is the degree of reliability that A estimates $t_{AC}(bfp)$ to have.

If we again take the bare situation where a user believes what an AI system outputs, without further information on the reliability of the output or whether the input is similar or not to the training data, then I consider it highly likely that the ascription will type the process using the AI as a whole (e.g. ‘B believes that there is a cat on the picture because the AI says so’, and not say e.g. ‘B believes that there is a cat on the picture based on an AI model operating under circumstances C with output type O.’ The typing found by Jönsson (2013) is uniformly of the former kind, with people answering that someone is seeing, hearing, speculating, etc. with little mention of more specific conditions under which this takes place. So, on this account too, it seems that the more coarse-grained typing that I used in my argument for local evaluation follows.

Finally, then, there is the solution to the generality problem offered by Lyons (2019), in terms of (as it happens) algorithms and parameters. He doesn’t mean computer algorithms here, but instead means the algorithms that describe our cognitive processes. As such, it straightforwardly applies to our belief-forming process when receiving AI testimony. Lyons’ central claim is that they should be typed as follows:

> Process tokens Γ and Δ are tokens of the same (relevant) cognitive process type iff the complete algorithmic characterization of Γ is the same as the

> complete algorithmic characterization of Δ, and Γ and Δ have the same parameter values (Lyons, 2019, p. 474).

Basically, if two token processes follow the same steps from input to output and have the same parameter values (to which I return in a moment), they belong to the same process type. The first question then is: do all token processes of AI testimony, where the user simply believes the output of an AI system, have the same complete algorithmic characterization? Well, to start with the AI system will operate in accordance with the exact same algorithm in all these cases, so there is no issue there (i.e. no basis for a more fine-grained typing that undermines my argument). I also think that it is plausible that the *cognitive* (i.e. human) processes follow the same steps, as in all cases the user simply believes the computer output, without additional information. Perhaps some outputs will be very odd (e.g. a natural language system writing about birds with a cleft palate, or when the user sees the input picture of a yellow shovel and gets the AI output 'banana') but we know that often users do not spot mistakes made by AI systems (Chiang and Yin, 2021; Parasuraman and Manzey, 2010) and can get rid of a lot of these cases by obscuring the input, or considering more difficult tasks (for humans) than object recognition. On the algorithmic characterization front, then, this account too leads to a broad typing that necessitates local evaluation.

The parameters are where it gets interesting, and where the crucial difference is with the account by Beebe (2004) that I consider in the next subsection. Parameters on this account are namely only those factors that cause a systematic difference in (human) processing, as opposed to simply any factor that is statistically relevant. For (Lyons, 2019, p. 489) "[t]ruth, reliability, and the like still aren't parameters because they don't cause or constitute differences in processing." So while there certainly are parameters that have a systematic influence on the accuracy of AI systems, these parameters are (almost always, to the point where it takes serious effort even from the developers to spot biases etc. in a system) obscured from the user. So, although they are certainly statistically relevant (hence me considering the next account in its own subsection) they are unlikely to have any systematic effect on the cognitive processing of a user who is only confronted with the output of the AI system. In short, on this end too, the typing will be the coarse-grained one I assumed, where all possible token processes of AI testimony fall under the relevant type. As a result, these solutions to the generality problem all fit with my claim that the evaluation of reliability has to be more local than simply all possible token processes falling under the type. The exception is Beebe's account, up next.

3.2 Statistically Relevant Factors

The final account of process typing I consider is similarly based on algorithms (Beebe, 2004, p. 180) gives three different factors that together determine the relevant process type for a given process token. Specifically, he does so via three conditions:

> The reliability of a cognitive process type *T* determines the justification of any belief token produced by a cognitive process token t that falls under *T* only if all of the members of *T*:

(a) solve the same type of information-processing problem *i* solved by *t*;
(b) use the same information-processing procedure or algorithm *t* used in solving *i*; and
(c) share the same cognitive architecture as *t*.

So, does an AI system as a whole fit these requirements, or does this condition force us into a more fine-grained typing? I think that this first set of conditions still maintains the coarse typing I assumed in my argument. For consider: the AI system will be explicitly designed to solve an information-processing problem (e.g. computer vision, natural language processing or fraud detection) and will use the same algorithm to solve this problem for all possible token processes. Furthermore, as I already argued above, it seems likely that a user who is only presented with the output of the AI will exhibit no differences in cognitive processing of this output. If this is where the account stops, then it too would be a case where the typing I suggested follows.

However, like Lyons (2019) and Beebe (2004) considers that parameters which influence the processing procedure are also relevant to the typing of the process. His worry is, namely, that there are still a large number of types that all meet the three conditions laid out, including ones that are irrelevant. For any process type *A* it is easy to add an irrelevant partition, e.g. *A* performed on Wednesday v.s. *A* performed on any day but Wednesday. Both sub-processes match the three conditions above and so qualify as viable candidates for the process typing. To avoid those cases Beebe invokes statistical relevance, where a factor *F* is statistically relevant for process *A* under circumstances *C* iff $P(A|C\&F) \neq P(A|C)$. The process type that is selected, then, is "the broadest homogeneous type (with respect to the production of true beliefs) within which t falls" (Beebe, 2004, p. 188), with a homogenous type being such that no more subdivisions can be made using statistically relevant factors.

This is an issue for my argument, as there are plenty of statistically relevant factors that partition an AI system, leading to a more fine-grained typing. For example, whether something is an adversarial case or not will be statistically relevant: AI systems are (by definition) unreliable for adversarial cases and so more reliable for the remaining cases. Furthermore, whether an input is an outlier is a statistically relevant factor: the AI system is less reliable for outliers than for inputs similar to the training data, and so on. There are lots of statistically relevant factors that partition the general process, and so Beebe's account leads to very fine-grained process types. So fine-grained that it becomes problematic. As Dutant and Olsson (2013) have argued at length, the process types collapse into ones covering only true instances and ones covering only false instances. The example presented by Lyons (2019) of forming beliefs about $m \times n$ using addition (which only works for 2×2) is a nice example here. If the process type is to 'do addition to solve $m \times n$

with $m, n \in \mathbb{N}$' then it is highly unreliable (as it is only true for $m = n = 2$). Yet the type 'use addition to solve $m \times n$ with $m, n \leq 1{,}000{,}000$ is already more reliable, so $P(m \times | m, n \in \mathbb{N}) \neq P(m \times n | m, n \leq 1{,}000{,}000)$. And it keeps getting more and more reliable the finer we type the process, until we reach 'do addition to solve $P(m \times n | m = n = 2)$', which is again a different probability from all other types. Vice versa, the same statistical relevance pushes you into typing the remainder of the process as $P(m \times n | m, n \in \mathbb{N} - 2)$, i.e. the process type over all natural numbers except 2. In short, by considering all statistically relevant factors the typing collapses into truth or falsity and thus trivializes the epistemology.

It seems then that if we want to type processes relying on cognitive factors alone, as the accounts in the previous subsection do, then this leads to a coarse-grained typing of processes based on AI testimony. Since the user of these processes reacts no different to the different cases considered by the AI (nor, for that matter, does the AI system itself as it will follow the exact same steps it does for every input) the typing is the same. If, on the other hand, we do push for a more fine-grained typing by considering all statistically relevant factors—and not just the cognitive factors—then this leads to a trivialization of the process typing. This can only be avoided if one manages to find a way to select *some* but not all statistically relevant (non-cognitive) factors as being important for process typing. I am doubtful that this is possible, and I think we can therefore say that more fine-grained typing is an unlikely answer to the cases I've presented here. Local evaluation of reliability is the more plausible solution. Yet, how does this claim fit in with the existing discussion on the evaluation of reliability? That is what I consider in the following section.

4 Existing Discussion on the Evaluation of Reliability

4.1 Arguments Against Local Evaluation

Now that I've argued that the process shouldn't be given a more fine-grained typing, I can consider the extant literature on the range of evaluation. As I mentioned in the introduction, there are at least two discussions that mention the possibility to evaluate reliability more locally and one account (Graham, 2012) that has a kind of local evaluation built in. To start with, Henderson and Horgan (2006) present a case where Athena and Fortuna are in fake-barn country, with the added feature that all yellow buildings happen to be real barns, but that there are plenty of barn facades of other colours around. Athena has a belief-forming process based on general experience with barns, and so believes of many facades in fake-barn country that they are barns. Her process is locally unreliable, but her extensive experience with recognizing barns gives the intuition that her beliefs are justified. Henderson and Horgan (2006) proceed from the idea that what we ultimately want for justification is the kind of reliability afforded by the refinement of belief-forming processes based on their successes and failures. They then argue that the kinds of processes that result from this adjustment to the environment will be transglobally (so across a range of possible worlds) reliable. Since Athena has had the appropriate kind of training in recognizing barns she meets this standard of transglobal reliability.

On the other hand, Fortuna has no experience with barns and first saw one yesterday, which happened to be yellow. Based on that experience she forms the belief that the building she sees is a barn every time she sees a yellow building. In fake-barn country this process is highly reliable (as all and only yellow buildings are real barns), but it is not (trans)globally reliable. So, Henderson and Horgan (2006) consider that the test for whether the process results from the right kind of adjustment to the environment is to see if it is reliable over a very wide range of situations/possible inputs. So how does this argument affect my argument from Sect. 2 that we should evaluate reliability more locally?

I follow the response by Graham (2014) to this and a few other cases. He takes issue with the difficulty of evaluating reliability globally or even transglobally, as well as the fact that perception often isn't reliable across such a wide range of cases (e.g. some birds detect mates simply by the colour of their beaks, which is another nice example of a process that is locally reliable but not globally reliable). He argues instead that it is sufficient for the overall goal of identifying the right kind of reliability to require that a process be non-accidentally reliable to solve these cases. The real issue with Fortuna, he claims, is that her belief-forming process is reliable by sheer luck, and that is enough to say that her beliefs are unjustified. He furthermore holds that Athena is justified because her belief-forming process is non-accidentally reliable under normal conditions (which I discuss below). Henderson and Horgan (2006) can be read as trying to specify such an anti-luck requirement through transglobal reliability, but as the bird example and my AI examples show, this is too stringent a requirement to capture the anti-luck intuition. If we keep it instead to an anti-luck requirement on reliability that does not mention the range of evaluation, then issue is solved quite naturally, and it also leaves intact my argument based on AI testimony: the situations where this is highly reliable are typically similar to the situations on which AI systems were trained to be reliable. So it is no accident that AI systems perform well in these circumstances. This does mean that one is owed an account of when a process is non-accidentally reliable (if it isn't transglobal reliability), a question that I'll leave for future work. As for the converse (Athena) question, I discuss that more fully in Sect. 5.

I turn therefore to the second objection to more local evaluations of reliability, by (Lyons, 2019, p. 488). He discusses the above-mentioned example of forming beliefs about $m \times n$ via the process $m + n$. Although this is reliable for $m = n = 2$ he consider the process just as unreliable/the output just as unjustified as for other inputs. And indeed, I think it would be wrong to reduce the range of evaluation so far. That, at least, is one response to this objection: the range of evaluation needs to be some minimum size. And, moreover, the selection of the range of evaluation shouldn't be simply based on reliability: just because a process gets some cases correct shouldn't imply that it is therefore reliable for those cases. The range of evaluation needs to be determined independently from the question whether the process is reliable over that range. I also find it tempting to say that this is a case of luck (that the process gets it right for $m = n = 2$), though this is hard to clearly define the idea in the case of necessary mathematical truths. I therefore mostly see this example as a challenge for any account that aims to specify how the range of evaluation is to be determined. It should rule out the 2 × 2 case, but still allow for justified beliefs based on testimony from (locally accurate) AI systems.

4.2 Which Possible World(s) to Evaluate in?

As mentioned, Henderson and Horgan (2006) argue that reliability should be evaluated transglobally: in the actual world and over a range of possible worlds. This is part of a wider discussion on the appropriate range of evaluation for reliability. For example, Goldman (1979) and Lyons (2013) consider that reliability should be evaluated in the possible world in which the process is employed ("reliability, i.e., reliability in the agent's world and/or environment" (Lyons 2013, p. 1)), though Goldman also mentions the option that reliability is evaluated in the actual world. Goldman (1986), on the other hand, holds that reliability should be evaluated in 'normal worlds'. Finally, Sosa (1993) and Comesaña (2002) argue for a two-dimensional composite of the possible world where the process is employed and the actual world. A large part of this discussion concerns the New Evil Demon problem, and the question whether reliabilists can/should say that a person in that possible world is justified in his or her beliefs. That issue is tangential to the one under discussion here. Whether the local reliability of the process is evaluated in the possible world where it is used, or in the actual world, or in some two-dimensional composite, is a separate matter. In all these cases we need to figure out which set of cases to look at within the possible world(s) in which reliability is evaluated. As such, my argument here is compatible with most of these views. In fact, one of them, defended by Graham (2012), already implies the more local evaluation that I argue for (in the possible world where the process is employed, to place it in context of the views mentioned in this subsection). I turn to that view next.

4.3 Normal Circumstances: A Possible Local Range

One view in the discussion on where reliability should be evaluated gives a specific result on what local circumstances matter. Graham (2012) presents an etiological account, on which the circumstances in which the belief-forming process acquired its function (of reliably producing true beliefs) is what matters: "It's reliability in normal circumstances that matters, for reliably in normal conditions individuates and explains what counts as normal functioning" (Graham, 2012, p. 456). This sounds like a promising way to spell out how AI testimony is to be evaluated, but the difficult part here is determining what the normal circumstances are in these cases.

Let me start, however, with the question whether AI testimony is covered by the account. I suspect that it is: "my account of entitlement applies to any belief-forming process that has reliability as an etiological function, whether the function derives from a history of natural selection or from a history of learning through reinforcement or other means." (Graham, 2012, p. 457) As the AI systems in question (supervised machine learning) are explicitly trained to minimise the error of their outputs, I think we can say that they have the function of reliably performing the task they are designed for. And so the full belief-forming process of AI testimony seems to fit the definition here, as believing the outputs of a system trained to be reliable should in turn be a process with the goal of forming reliable beliefs.

The question then is what the normal circumstances will be where the AI is to be evaluated. In the cases Graham discusses the normal circumstances are those in which the process acquired its function. So, if we apply that to the AI case then it seems that the training data is the basis of what counts as normal circumstance for AI testimony. It won't be just the training data, of course, just as the normal circumstances for perception aren't restricted to exactly those in which it first evolved, but apply just as well in the somewhat different environment we are currently in. So, anything sufficiently similar to the training data will count as normal circumstance. Is that a reasonable range to pick? I think there are some difficulties, to do with the possibility that the training set isn't representative.

Take the case of Alcorn et al. (2019) again, where it turned out that computer vision algorithms fail to correctly identify cars, scooters, etc. in a wide range of circumstances. Similar studies have found that these algorithms don't recognize stop signs if there are a few pieces of black and white tape on it (in fact the AI saw it as a sign indicating a higher speed limit (Eykholt et al., 2018)). The training data contained none of these circumstances, and by now a whole range of 'natural adversarial cases'—plausible scenario's where AIs make serious mistakes—is known (Hendrycks et al., 2021). Part of the reason will be that the training and original validation set were insufficient to capture the phenomenon the AI was meant to detect. Generally speaking, then, there is no reason why the training set should capture the full range of what we consider 'normal' cases, because the creators of these datasets could easily have overlooked cases that are normal. The environment captured by the training data could be said to be normal yet simply different from the environment in which it is deployed, but I think that it's a better analysis to hold that the data fails to capture the normal environment. There may not be a clear environment that matches the exact contours of the training data (e.g. with only white men) and we do consider the fault to be with the data and the cases it identifies, rather than with the choice of use cases. So what I think is lacking is this initial situation where the process is functioning as intended, which we see with young birds in their natural environment. Graham wants to capture that idea, and focussing on the training set won't do. Nor will it do for calculations of reliability, precisely because these datasets need not give an accurate picture of the situations in which they are used in practice.

Instead, then, consider the circumstances intended by the designers of the system. AI systems, after all, get their function not through a process of selection, but by design. Will these do? It does get the self-driving car case right, as the intention of the designers is for the algorithm to correctly read stop signs etc. So, going by the intention of the designers will lead to the conclusion that this system is unreliable if there are too many mistakes (in practice) and won't provide justification for any outputs. Yet designer intentions are also a rather fickle basis for the reliability measure. Perhaps they have lofty inspirations, intending the system to be much more generally applicable than it actually is. The converse can happen just as well, as seems to be the case with the natural language system GPT-3. The designers probably intended simply to build a larger scale version of the previous language model they developed, and were subsequently surprised by how many language tasks the model could tackle. It turned out to be more versatile than expected. Surely that shouldn't imply that we are unjustified to use the system in the cases where it was unexpectedly accurate, simply because those circumstances weren't intended by the designers.

I think, then, that the two most natural interpretations of 'normal circumstances', a historical one (based on the training data) and one based on the designers of the system, are unlikely to give the right results. Might other options work? We could work with a pre-theoretical notion of normal circumstances and/or normal use, avoiding the reductive analyses that I've given above. And here we see one reason why I picked AI systems as a case, rather than more familiar processes. For the non-accidental reliability of AI algorithms needn't be linked to normality as closely as most biological processes are. As an example, in the case of facial recognition algorithms there are many faces that would be viewed as normal on a pre-theoretic notion—most of them, in fact. If there are no serious birth defects, or scarring from accidents etc., we'd quickly regard a face as normal. Yet, notoriously, facial recognition algorithms were (and to some extent are) biased in terms of their accuracy. Gender classification based on images of faces worked particularly well for white men, slightly worse for white women and far worse for other people (Buolamwini and Gebru, 2018), with accuracy gaps as high as 34.4 percentage points. That means that an algorithm intended for use on the entire population, and in fact used on the entire population, was highly reliable for one group (for white males in particular accuracy was around 99%) and close to chance for other groups (65.3% accuracy was the worst case). So, if we evaluate the algorithm over normal circumstances, both seen in terms of circumstances of use [e.g. a range of evaluation defined by what inputs the process receives in nearby possible worlds) and in terms of normal faces, it will score much worse than when evaluated over a subpopulation. Such cases are easy to come by, as object recognition has a similar disparity in accuracy for objects in western v.s. objects in other countries (soap being recognized when in a bottle, but not when it is a bar of soap, for example; De Vries et al. (2019)].

Are such systems still providing us with justifiable outputs in the subgroups for which they are highly reliable? It seems plausible to think that we are justified in believing these outputs in those cases, as there is a situation of non-accidental high accuracy (the situation is one where the AI was trained primarily on faces of white men). In fact, because the AI was trained on this subgroup Graham would likely identify the recognition of faces of white men as the function by his etiological definition. Now, defining normality by the training set isn't a water-tight move as discussed above, but it is how algorithms can become non-accidentally reliable and that makes for a decent argument that we are justified to believe the outputs of the algorithm in those cases (though not for all faces).

In short, the challenge is that what is 'normal' pertains to the world, and those normal circumstances might not be the ones in which the algorithm is accurate. A prominent reason for this is that the training set of the algorithm isn't necessarily representative of the world. That can be a problem for reductive definitions, because the training set misses important aspects of a phenomenon that we think should count as normal circumstances. But it can also be a problem for fixing the range of evaluation in terms of a pre-theoretic notion of normal circumstances or normal use, because that specific range (e.g. normal faces) need not be the one for which the algorithm is non-accidentally accurate. It is for that reason that I'm skeptical that 'normality' fully captures what we're after here, and also why I've picked AI as an example throughout—as it is easy to find cases of non-accidental reliability without lining up with our notion of normal circumstances. Perhaps there is still some way to make the idea work with a notion of normality, but I think it helps to broaden the scope of possible solutions a bit. Therefore I turn next to a discussion of how this range of evaluation may be determined without using normal circumstance as the guiding notion.

5 Determining the Range of Evaluation

When we have a belief-forming process that produces a certain output, how reliable is the process that produced it? That is the central question I am posing here, and where I've argued that the answer requires a method of determining the range of evaluation over which this reliability is determined. While I have focussed on examples from AI systems, where accuracy is easy to measure and process individuation is fairly straightforward, the question applies generally. I'll still use AI examples for some technical details here and there, but hope to formulate an answer that works in other contexts too, and circle back to the examples from other authors at the end of this section. So, then, how should we identify the range of evaluation, if not through normality?

Two perspectives offer themselves. First, the question is to find a local range of evaluation, which means that we want to evaluate the process over all cases that are 'close enough' to the input–output pair in question. As Graham (2012) already notes, a range of circumstances is needed in which the process is reliable. If we look at the literature on AI systems, that can translate into a distance measure that determines which inputs are close to the actual input.[2] Note that this is quite different from looking at 'normal circumstances' as Graham intents, or looking at e.g. those inputs which a process might easily get in nearby possible worlds. Those notions look at the circumstances in which the process is calibrated or employed, whereas the suggestion here is instead to ignore those aspects and look simply at the range of cases obtained from making changes to the inputs.

[2] Karimi et al. (2020) has a survey on such measures, though they are specifically looking for the nearest input with a different output. As a concrete example, (Wachter et al., 2017) uses a weighted Manhattan distance over input x made up of dimensions k (e.g. the different pixels, or variables such as age, income, etc. found on a loan application):

$$dist(x, x') = \sum_{k \in D} \frac{|x_k - x'_k|}{\text{median}_{j \in P}\left(|X_{j,k} - \text{median}_{l \in P}(X_{l,k})|\right)}$$

The lower part of the fraction corrects for the median average deviation found in variable k, to handle the fact that e.g. income will vary much stronger than age.

My suggestion is thus a fairly direct notion of similarity: the range of evaluation consists of those cases that don't differ too much from the actual case for which we need to determine if the process is reliable or not. So, if the actual input is the face of a white man, then we primarily look at other faces of other white men – faces that look broadly similar, and more similar than faces of (say) asian women. The mentioned gender classification algorithm will be highly accurate for that group of faces. However, if we take another input, and look at the face of an African-american woman, then the range of similar faces will be one over which the algorithm isn't all that accurate. Note that we probably need to restrict this range of similar faces to ones that are plausible—there might be changes to a face that result in very atypical images, such as adding a third eye. And this brings us closer to an idea that we might be looking at a range of close possible worlds, though it is not quite a range of inputs that the process could easily receive (as the similarity metric doesn't look at our use of the belief-forming process). Still, the idea might sound familiar.

For compare my suggestion to one of Williamson's formulation of the notion of safety: "[I]n a case α one is safe from error in believing that [a condition] C obtains if and only if there is no case close to α in which one falsely believes that C obtains" (Williamson, 2000, pp. 126–127). Of course, that is a rather more stringent requirement than my suggestion for the range of evaluation. Understandably so, as I am looking at justification rather than knowledge. I also wouldn't be surprised if we prefer to have a wider range of evaluation than the range we'd pick for the safety requirement of knowledge. Still, the safety requirement follows the same idea of looking at sufficiently close alternative inputs and the comparison may help elucidate how my suggestion differs from that of Graham.

However, I don't want to deviate completely from the account that Graham (2012) presents, or from the intention of Henderson and Horgan (2006) in looking at appropriately adjusted belief-forming processes. While the range of evaluation is fixed in a different manner, I am in favour of thinking that it is non-accidental reliability that matters. This idea of reliability being non-accidental is, I think, plausibly spelled out by Graham's connection to natural selection or reinforcement learning, or (in the AI case) the training data and in line with the general aim of Henderson and Horgan (2006). It is the calibration of the process, in whichever way that happens, that makes it non-accidentally reliable for a certain set of inputs (where, again, a more developed account of non-accidental reliability is left to future work). If we didn't have that aspect in the mix, then the barn cases from Henderson and Horgan (2006) would quickly present a problem again, as they are relying precisely on the idea that all similar inputs (the other barns in fake barn country) are ones for which the process is reliable. One more remark needs to be made, however, on the restrictions to the range of reliability before I go through those examples in more detail.

For there's a second perspective which I think is important: whether the process type reliably produces beliefs with similar contents, as opposed to under similar circumstances. To take computer vision as an example again, there is the question whether the AI system reliably recognizes e.g. a hammer when it says that there is a hammer on the picture. There is nothing in the above distance measure that will guarantee that reliability is measured over a large number of hammers (or a large number of cases where the AI thinks there is a hammer), so no way to guarantee that the evaluated reliability is that of detecting hammers. Rather, it would be the reliability of detecting objects under circumstances similar to those of the token process in question. For other processes we might wonder the same: is the process generally reliable, but not for the type of content that it currently produces? Vision in young birds might work reliably to determine whether their parents are nearby, but not reliably to spot predators, or cardboard, etc. And so it makes sense to narrow down the range of evaluation by the type of content. Now, that doesn't mean that we should restrict the evaluation to cases where the young bird is correctly identifying a parent, but instead to narrow it down to the cases where the bird forms the belief that there is a parent present, in addition to the cases where there really is a parent present (but the bird might not notice). The same process might have very different reliability for those different types of content, after all[3]

And indeed, in the AI case Wong et al. (2021) find that the accuracy of computer vision models varies widely depending on the output label, i.e. on the type of content one looks at. For example, these models are good at recognizing teapots, cats and hammers, but particularly bad at recognising screens and printers (measured over the image dataset which is also used for training the models). That seems to be a highly relevant difference in reliability. The fake barn cases, for example, are presented as problematic for a local evaluation by Henderson and Horgan (2006) because in those circumstances a globally unreliable process *for recognizing barns* happens to be locally reliable. The question is not whether perception is generally reliable in fake barn country, but focusses on the content of the belief in question (i.e. 'that is a barn').[4] We want to know whether we are justified to believe the content, and so what should matter according to a reliabilist is whether the belief-forming process reliably produces beliefs with those kinds of content. The range of evaluation for the process relevant to any particular belief can then already be restricted

[3] For AI systems we can define this more formally, e.g. using the method from Wong et al. (2021) that starts with a score Q for input x paired with output y given the correct classification z:

$$Q(x,y) = \begin{cases} ((C(y|x)^{\alpha} & \text{if } y = z \\ (1 - C(y|x))^{\beta} & \text{if } y \neq z \end{cases}$$

Here α and β are different factors rewarding/punishing correct/incorrect outputs. Now, this doesn't set up a local range of evaluation, as it is just for a single input–output pair (so almost a token process). They showcase the difference with looking at the output type for the range of evaluation in their computation of a combined TrustScore (as they call it) which integrates the different individual scores with respect to the output label, e.g. all cases where the model says 'hammer', or all cases where it says 'screen'. Formally specified, for some set of input–output pairs x,y (which occur with probability P(x, y)) the aggregated score T is:

$$T = \iint P(x,y)Q(x,y)dxdy$$

By restricting the integral to e.g. those cases where the output $y =$ 'hammer' with a minimum confidence you get a measure specifically for all cases where the model claims there is a hammer in the picture.

[4] Of course Henderson and Horgan (2006) do this by giving a fine-grained typing of the processes ('believing a building to be a barn based on its yellow colour'), but I take it that my reconstruction of the case with a coarse-grained process is equally problematic and also calls out for the solution that the reliability needs to be non-accidental.

to situations in which the process produces beliefs with similar contents, but also to cases where the correct outcome would be a belief with similar content. It might be the case, after all, that there are lots of cases where the process misses e.g. a parent that is present, or a barn of a different colour (in the Henderson and Horgan (2006) case), and those false negatives should be taken into account when looking at the reliability. Similar contents, then, both in the actual resulting beliefs and in what would be the correct beliefs.

I thus suggest to have a two-pronged limitation of the range of evaluation: on output (similar contents), and on input (sufficiently similar to the actual input). Reliability over this range should be non-accidental, which is nicely covered by Graham's approach. With those restrictions in place, I think the range of examples discussed so far can be accommodated quite nicely. I've already sketched how this might work for the gender classification case, and think we can say something similar for the differences in reliability in object detection algorithms (a bottle of soap isn't all that similar to a bar of soap, and the AI is non-accidentally reliable for the first but unreliable for the second). The adversarial cases naturally fall in the range of evaluation, and if there are too many plausible cases where the algorithm is wrong, it seems fair to call it unreliable. That being said, I didn't want to restrict my focus to AI, so how about the other examples from the epistemology literature?

To start with, there are the two fake barn cases of Henderson and Horgan (2006). On the one hand there is the locally reliable process (in fake barn country) which is not globally reliable. The accidental nature of this reliability is to blame, as I've discussed above. This is certainly captured by keeping Graham's appeal to a simpler notion of non-accidentality than transglobal reliability. I'll assume for arguments sake that all sufficiently close inputs (or most of them) fall under fake barn country, but this conclusion isn't all that obvious: my distance function wasn't directly related to the circumstances of use, and e.g. changing the colour of the barn seems to count as a fairly small change to the input conditions. Since I'm not looking at a local range in terms of the geographical location, or in terms of the other barns that Fortuna might easily see, the outcome is likely to be that she is unjustified because her belief-forming process is not widely applicable. Still, I don't doubt that other examples can be found where one is lucky within the local range of evaluation, so hence I'll consider that situation as well.

The converse case of Athena, who has a globally reliable process for determining whether something is a barn which is locally unreliable, is not as easy to handle. I think that here again it helps to stress that the local range I'm suggesting isn't one particularly bound to the location in which Athena finds herself. Still, if fake barn country is large enough, and contains sufficient variety to fit in the full range of evaluation, then Athena's process will be deemed locally unreliable even if it is globally reliable. And so she won't be justified in (notably large) fake barn country, even though she is justified in her barn-related beliefs sufficiently far outside of fake barn country. Provided that we set the range of evaluation wide enough, I think this conclusion isn't that problematic. In that case there will be some fake barn countries (ones small enough to not take up the entire range) that maintain justification, and some where it is so systematically different that our belief-forming processes are not just wrong due to bad luck but are wrong because they systematically misrepresent the environment. Of course, Athena still might not be to blame for continuing to form beliefs using her globally reliable process; as in other examples (young birds in lab settings, receiving AI testimony without knowing that the system is unreliable in that particular case) that might be the only rational thing to do. Still, a reliabilist will have to say that she lacks (externalist) justification. And so ultimately I think this ends up with a balancing act between the two cases, where we want ranges of evaluation that are reasonably large, but not too large.

Just by opting for reasonably large ranges of evaluation we avoid the problems with cases such as the evaluation of $m \times n$ using the operation $m + n$, presented by Lyons (2019). Since the range of evaluation isn't selected based on differences in reliability, but looks at similar inputs, more than just the number two would be part of it. Consequently, the suggested belief-forming process will count as unreliable. The trick is again, though, to get the balancing right. If the range is set too broadly, we might run into difficulties with the perceptual processes in young birds discussed by Graham (2014). Take this too broadly, and birds in normal conditions (that is, not in a lab but in a nest) might be evaluated partly by the very different conditions of looking at cardboard cutouts. If, however, we keep the range of circumstances to consider small enough—but still sizeable—then the reliability of their belief-forming process would be tested by whether they'd still spot their parents if the shape of their beaks was somewhat different, if the lighting conditions would be changed, etc. We shouldn't, note, look at situations where the spot on their beak isn't red, but some other colour, since it's precisely the red that the belief-forming process uses. Here, again, the plausibility of the changes should help, as it's plausible for the beak to have a slightly different shape, but not for the dot on it to be green. Just as it makes sense to change facial features in some ways, but not in others (like adding a third eye).

6 Conclusion

Over what range should reliabilists measure reliability? I have argued that it should not be that of all possible token processes of the process type. AI systems are a salient example where the process type is (often) unreliable when measured over all possible token processes, but reliable for a substantial more local range of token processes. As it seems undesirable to say that no justified beliefs can be formed using AI systems, the range of evaluation should be restricted. I have defended that argument by looking first at the typing of the process, showing that almost all existing accounts of process typing entail a coarse-grained typing of the belief-forming processes appealed to in my argumentation. Furthermore, the one exception to the rule collapses into a trivially narrow typing, precisely because it appeals to all statistically relevant factors rather than only cognitive ones.

A smaller range of evaluation therefore is the better way to handle these cases. I suggest that this range is determined first and foremost by the content of the belief resulting from the token process, where resulting beliefs that have or should have similar contents are what we care about. After that initial selection, the range is further restricted by looking only at sufficiently similar inputs for the other token processes. One way to phrase this requirement is to say that the range is restricted by a distance function over the input variables of the token process which focusses on plausible (/probable) changes to the inputs. Of the options considered here, that seems to be the best way of dealing with the scenario where a user forms beliefs based on the output of an AI system, as well as other (e.g. perceptual) situations where belief-forming processes are locally reliable but globally unreliable. Finally, it handles the few objections in the literature on the range of evaluations well, though my discussion is largely tangential to that one as I am not concerned with the question of what possible world(s) should be picked for the evaluation of reliability. Naturally, there are still open issues. We are owed an account of non-accidental reliability. Furthermore, by focussing on the range of evaluation I abstracted from other issues such as the time frame, the internal processes of users of AI systems, and the way reliability is calculated (e.g. do worse errors count more heavily?). These do of course matter for any complete account of the cases I've discussed, which have hopefully been shown to be fruitful ones for further discussion.

References

Alcorn, M., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W., & Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4845-4854).

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6,* 14410–14430.

Beebe, J. (2004). The generality problem, statistical relevance and the tri-level hypothesis. *Noûs, 38,* 177–195.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency, PMLR* (Vol. 2018, pp. 77–91).

Chiang, C., & Yin, M. (2021). You'd better stop! Understanding human reliance on machine learning models under covariate shift. In *13th ACM web science conference 2021* (pp. 120–129).

Comesaña, J. (2002). The diagonal and the demon. *Philosophical Studies, 110,* 249–266.

Comesaña, J. (2006). A well-founded solution to the generality problem. *Philosophical Studies, 129,* 27–47.

de Vries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 52–59).

Dunn, J. (2015). Reliability for degrees of belief. *Philosophical Studies, 172*(7), 1929–1952.

Dutant, J., & Olsson, E. (2013). Is there a statistical solution to the generality problem? *Erkenntnis, 78,* 1347–1365.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., & 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).

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). Reidel.

Goldman, A. (1986). *Epistemology and cognition*. Cambridge: Harvard University Press.

Graham, P. (2012). Epistemic entitlement. *Noûs, 46*(3), 449–482.

Graham, P. (2014). Against transglobal reliabilism. *Philosophical Studies, 169*(3), 525–535.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Heaven, W. (2020). Our weird behavior during the pandemic is messing with AI models. MIT Technology Review May 11, 2020. https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/

Henderson, D., & Horgan, T. (2006). Transglobal reliabilism. *Croatian Journal of Philosophy, 6,* 171–95.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15262–15271).

Jönsson, M. (2013). A reliabilism built on cognitive convergence: An empirically grounded solution to the generality problem. *Episteme, 10,* 241–68.

Karimi, A., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint arXiv:2010.04050

Kuppers, F., Kronenberger, J., Shantia, A., & Haselhoff, A. (2020). Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 326–327).

Lyons, J. (2013). Should reliabilists be worried about demon worlds? *Philosophy and Phenomenological Research, 86*(1), 1–40.

Lyons, J. (2019). Algorithm and parameters: Solving the generality problem for reliabilism. *The Philosophical Review, 128*(4), 463–509.

Olsson, E. (2016). A naturalistic approach to the generality problem. In *Goldman and his critics* (pp. 178–199). Wiley-Blackwell.

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410.

Pettigrew, R. (2021). What is justified credence? *Episteme, 18*(1), 16–30.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–52).

Sosa, E. (1993). Proper functionalism and virtue epistemology. *Noûs, 27,* 51–65.

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation, 23*(5), 828–841.

van den Hoven, J. (1998). Moral responsibility, public office and information technology. In I. Snellen & W. van de Donk (Eds.), *Public administration in an information age: A handbook* (pp. 97–112). Amsterdam: IOS Press.

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.

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104–122.

Weatherson, B. (2012). The temporal generality problem. *Logos & Episteme, 3*(1), 117–122.

Wheeler, B. (2020). Reliabilism and the testimony of robots. *Techné: Research in Philosophy and Technology, 24*(3), 332–356.

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

Wong, A., Wang, X., & Hryniowski, A. (2021). How much can we really trust you? Towards simple, interpretable trust quantification metrics for deep neural networks. arXiv preprint arXiv:2009.05835

Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In *Big data analysis: New algorithms for a new society* (pp. 91–114).