Challenges for Computational Reliabilism: Epistemic Warrants, Endogeneity and Error-based Opacity in AI, ML and other Computational Methods

1. Introduction

The concept of computational reliabilism, originally coined in the context of epistemological problems related to computer simulations (Durán and Formanek, 2018), has been recently deployed to justify our reliance and trust in many other computational technologies, including machine learning methods in artificial intelligence such as deep neural networks (Durán and Jongsma, 2021). Roughly, these deployments can be understood as seeking to serve several— and often interrelated— purposes under the umbrella of a unified epistemological framework adept to account for a justified reliance on computational practices, methods and devices. In particular, an overarching hope of those championing such a framework is that computational reliabilism can:

- Respond to or circumvent the challenges related to epistemic opacity in computational methods, and in doing so,
- warrant or justify our beliefs regarding the reliability of computational processes and their results; and hence,
- To reassure us of the possibility of trust in computational methods, practices and artifacts even if these are insurmountably opaque.[1]

In this chapter I aim to elucidate what I deem to be three major challenges to computational reliabilism. I deem these challenges to have a bearing on its viability *both* as a general epistemological framework capable of dealing with the advent of computational methods, particularly in scientific inquiry, *and* as a pragmatic epistemic resolution to the justification problems related to the adoption of opaque computational methods, both of which are often cited as motivations for its adoption (Durán and Formanek, 2018; Durán and Jongsma, 2021). In particular, I focus on the following three challenges:

1. The challenge of warrant transmission and reliability-crediting properties
2. The challenge of the indispensability of endogenous features in artifactual reliability, and
3. The challenge of error-related opacity

2. Computational reliabilism as a kind of reliabilism

Computational reliabilism, a term coined by Durán and Formanek (2018), is the view that in the context of computational methods such as AI systems or computer simulations researchers are justified in believing or trusting the results yielded by such methods "*because* there is a reliable

[1] That reliability and trustworthiness are not the same thing is well known and well-documented in the literature. Nevertheless, that points 2 and 3 are distinct is not always clear or explicitly stated. However, that they are clearly distinct issues is evidenced by the fact that trust can be had in things whose reliability is questionable and by the fact that otherwise reliable methods and practices are sometimes not trusted. The relationship between epistemic efforts such as explainability, interpretability and trustworthiness, is a non-trivial issue that is the subject of important ongoing debates, some of which, despite their important insights, may be beyond the scope of this chapter.

process (i.e. the algorithm) that yields, *most of the time*, trustworthy results." (Durán and Jongsma, 2021 p.332 italics mine) [2] According to Durán and Formanek, computational reliabilism borrows closely from Alvin Goldman's epistemological reliabilism. In particular, it borrows from *process reliabilism* which suggests that a given inference/assertion can be deemed reliable if it is the product of a reliable process (2011). How such a process can be deemed reliable varies, but strictly speaking, the details do not matter that much: a reliable assertion need not include details about the reliable processes that produced it, the reliability in question is that of the assertion and not that of the processes by which it was produced, or so it is argued. Following from this, a process can be reliable even if the reasons *why* it is reliable are not accessible to an agent (Comesaña, 2010 p.571). In the context of perceptual claims, for example, the reliability of the process must be treated as somewhat of a brute fact. That is, we must simply accept that our perceptual system yields true beliefs more often than not, even if we do not know how or why it does so or how and why we know so.

Although at times Durán and Formanek "heartily endorse" views of reliability that suggest that scientists are justified in trusting the results of their computational methods simply because they "trust the assumptions upon which they are built" (2018 p.652), as suggested by Beisbart (2017) and others, a closer look at their views suggests that computational reliabilism— rightly— forgoes these kinds of assumptions (ibid). Unlike conventional reliabilism, computational reliabilism needs not presuppose the reliability of the methods from which the results are obtained. Rather, it is suggested, computational reliabilism takes in consideration 'reliability indicators' as "markers of methodological and epistemological competence of the computer, algorithms and social processes involved in the formation of beliefs." (Durán, this issue). Accordingly, Durán and Formanek suggest that computational reliabilism requires a *retrospective reliability chain* "that conditions the sources that attribute reliability to [computational methods] to be reliable in and by themselves." Furthermore, such sources, they accept, "must be *shown* to be reliable." (2018 p.656 italics mine).

As we shall see in the next section, this last point seeks to make the reliability chain referred to by Durán and Formanek somewhat distinct from a simple appeal to a chain of epistemic entitlements—non-evidentiary epistemic warrants used to justify ordinary knowledge claims and widely believed to be epistemologically acceptable (if not necessary) in everyday epistemic endeavors (Graham, 2012). By contrast, the way in which computational methods, or their constitutive algorithms, are deemed reliable, according to computational reliabilism, is through the consideration of factors—external to the algorithms themselves— that function as reliability indicators. These include "identifying methods (formal or otherwise), metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts that might increase the degree of warrant we have to believe the outputs of ML systems" (Duran, this issue p.5). Reliability indicators for machine learning methods, more specifically, include the following (Durán, this issue):

[2] Other versions of reliabilism related to computational systems expand the foci of the processes in question beyond just the algorithms and suggest that in contexts such as AI, a sociotechnical understanding of such processes is required. Although Durán and Formanek original framing of the computational reliabilism is directed towards algorithmic processes, there is a sense in which later work (See Durán and Jongsa, 2021 and Durán, this volume) the considerations are broadened to include both technical details and social practices.

- Technical robustness of algorithms
- Computer-Based scientific practice
- Social construction of reliability [3]

According to this view, we can rely on the results of ML because they are the product of robust algorithms, which in turn come from trustworthy computer science practices, and because reliability is a contextual and socially negotiated concept that includes the input of a diverse group of experts, practitioners and users. These methods, metrics, practices and the processes that include them produce reliable results more often than when such methods, metrics and practices are absent.

The "most of the time" or "more often than not" clauses above make sense of Durán's suggestion that, at its core, computational reliabilism is meant to be a frequentist approach to reliability (Durán and Formanek, 2018; Durán and Jongsma, 2021; this issue). Accordingly, the reliability of a process must be understood as a tendency to produce "high proportion of true beliefs relative to false ones." (ibid p.653) [4] For instance, Durán and Formanek explicitly state that computational reliabilism is predicated on whether or not "the probability that the next set of results of a *reliable* [computational method] is trustworthy is greater than the probability that the next set of results is trustworthy given that the first set was produced by an unreliable process by mere luck."(2018, p.654) Trusting the results of computational methods therefore, according to this view, "depends on a chain of reliable processes that, in the end, allow researchers to be justified in believing the results" (p. 655) Where this chain ends, however, is simply left unanswered (ibid, fn.6 p.655).

Still, as mentioned above, advocates of computational reliabilism deem the framework as capable of responding to the challenge of *essential epistemic opacity* in computational methods.[5] Because of this serious challenge, finding an epistemological framework capable of circumventing these obstacles seems highly desirable. Whether or not computational reliabilism is such a framework, as we will see, is not immediately obvious.

[3] While verification and validation techniques as well as expert knowledge are still discussed in the context of machine learning, Durán considers them as belonging to the categories of technical robustness and the social construction of reliability respectively.

[4] This is something echoed in Ferrario's (2023) approach to their own, highly formalized, version of reliabilism.

[5] This is the kind of opacity that is not relative to an agent's contingent epistemic limitations but is rather relative to their epistemic nature (Humphreys, 2009). Furthermore, in some instances, particularly in the context of computational methods, this kind of opacity arises in virtue of features of a system and not in virtue of any agential-relative features. As such, the causes of such opacity are agent-independent—I.e. the opacity in these cases arises from factors that are independent of any limitations on the part of the agent— and because no agential resources can correct them, they are also agent-neutral: i.e., any epistemic agent would face the same challenges. See Alvarado's discussion of Nagel's agent-neutral reasons (2023 p.136).

3. The challenges

In this section, I provide an overview of three challenges that I believe have a negative effect on the viability of computational reliablism as an epistemological framework that can appropriately capture the novel and/or unique epistemic issues associated with the use of computational methods, particularly in formal epistemic contexts such as scientific inquiry or in safety-critical contexts. In both of these contexts endogenous features— such as the nature and source of error— of a system are simply indispensable reliability indicators and computational reliabilism seems unable to account for them. As I will show, these challenges also elucidate the limitations of computational reliabilism even when it is framed solely as a pragmatic epistemic resolution to the justification problems related to the adoption of opaque computational methods.

3.1. The challenge of warrant transmission and reliability-crediting properties

Although argued for elsewhere by Symons and Alvarado (2019), it may prove useful to begin by briefly noting that one of the serious challenges for computational reliabilism as an epistemic framework is that it fails to account for the epistemic challenges related to warrant transmission, particularly as related to sanctioning novel computational technologies. As we saw in the previous section, the success of computational reliabilism strongly hinges on what Durán and Formanek call 'reliability indicators.' According to them these include "algorithmic-related methods and practices" that have "*reliability-conferring"* properties (Durán, in this volume p.7). According to these views, some of these practices can confer their reliability on to a process, a device that comes from such process, or the results of such a process. There is, however, a problem with the concept of reliability-conferring indicators and that is that there is no clear sense in which they can *confer* such reliability. This is particularly the case if the thing being conferred reliability to in virtue of established methods and processes is not itself a-yet-established method or technology. For example, the best practices by the best experts in the best of settings behind the invention of a never-before-seen technology are not enough to ensure that such technology is itself reliable. As I have argued for elsewhere (Alvarado, 2023a p.72), even in the case of technological devices that are now considered paradigmatic scientific instruments such as the classic optical telescope, neither the name and reputation of Galileo nor those of Kepler alone sufficed to establish it as a scientifically reliable instrument. What is more, neither Galileo's assertions that his *theoretical* approach resulted in a better instrument than the original, nor Kepler's initial efforts to drum up consensus about the validity of Galileo's instrument amongst the scholarly community proved to be sufficient warrant for its serious adoption (Alvarado, 2023). Rather, extensive testing as well as criteria for what constituted proper testing had to be designed and developed for telescopes to undergo before they were sanctioned as scientifically reliable instruments (Van Helden, 1994).

The difficulties regarding the transferability of epistemic warrants, or the lack thereof, from methods, practices and people to technical artifacts, become clearer in epistemic contexts in which epistemic endeavors that require hard rigor depend on one another, where the scale of resources required demand significant efforts in preparedness and constraint, and in contexts that could cause significant harms to others. Deeming the atomic bomb as a reliable weapon both in its power and its limitations, for example, solely in virtue of the fact that it was made with the best available methods and by the best physicists at the time, without probing the artifact itself to produce its own epistemic warrants, is clearly an inadequate epistemic strategy. As demonstrated by Symons and I (2019), epistemic warrants that justify reliance or trust on one method or on a group of experts are not simply transferable to the technical artifacts built with such a method or by such group of people. The notion that either existing reliable methods or human expertise can simply "*credit* reliability" forward to computational methods (or any technical artifact for that matter) as Durán suggests, is thus not immediately obvious.

Even in the case of highly cautious epistemic endeavors that require technological mediators, ensuring that such warrants transfer from one process to another requires that hard-to-attain epistemic conditions hold. Consider appeals to practices in computer-assisted mathematical proofs as an analogy to how we could trust computer simulations (Barberousse and Vorms, 2014). One condition that these appeals must assume holds in both settings is that the methods involved are transparent conveyers of epistemic content— that is, that in the process of manipulating epistemic content that is supported by one kind of warrant, say an apriori epistemic warrant like the ones supporting mathematical inferences, did not introduce elements into the process that required the support of another kind of epistemic warrant, namely a posteriori epistemic warrants such as those provided by empirical evidence (Burge, 1993; 1998). The analogy between computer-assisted mathematical proofs and computer simulations, as elucidated by Symons and I, was already a dubious stretch given the distinct practices and distinct epistemic norms involved in each endeavor. As noted by Winsberg (2010; 2019) and others, computational methods such as computer simulations, involve a vast and non-trivially motley set of experts, practices and methods. These practices and methods are far removed from the rigorously cautious process of mathematical proofs. Transparent conveying of epistemic content is simply not something that can be easily achieved in such settings. The fact that we cannot simply trust the mathematics involved in such processes (which require an a priori warrant) but that we have to also trust practices and equipment (which require their own a posteriori warrants) demonstrates the non-transparency of the conveyers at play. Hence, epistemic warrants fail to simply transfer from one step to the next and each part of the process must muster its own through distinct means.

It is worth noting that even in more ordinary epistemic contexts, the transfer of epistemic warrants is epistemically non-trivial. Take for example the problem of *epistemic alchemy* (McGlynn, 2014) in which a seemingly warranted proposition is derived from a poorly warranted one. In the case of epistemic entitlements, this would imply deriving a purportedly justified claim from claims supported by non-evidentiary warrants. While it may very well be the case that in ordinary epistemic practices this may not be too much of a problem (ibid), it is not clear that this practice or that epistemic entitlements themselves, such as the ones relied upon by Burge and others (see Barberousse and Vorms, 2014), are adequate in the context of scientific inquiry to start with. Rather, extensive sanctioning processes that are in fact reactionarily independent from— i.e., critically engaging with and seeking to overcome the limitations of— existing methods or current expertise consensus have to be designed and implemented for novel technologies to be appropriately sanctioned for scientific use (Alvarado, 2023a).

This points to the fact that making the case that warrant transmission occurs, if it occurs at all between practices, methods, experts and their artifactual products, takes a non-trivial philosophical effort. Unfortunately, other than noting that these reliability-conferring properties are not to be understood as ‘spooky’ (Durán, preprint), this effort is all but absent in accounts of computational reliabilism to date. Whatever version of computational reliabilism that strongly hinges on the notion of reliability-conferring, or reliability-crediting properties will have to seriously contend with the issues of warrant transmission. Until then, the future promissory note of existing computational reliabilism accounts seems to be of too high a denomination to just go unchecked, signaling a genuine and serious challenge to the viability of computational reliabilism as an epistemic framework.

To be fair however, this is a major epistemological problem for most accounts of reliability. Hence, in what follows I will take it as a given that there may be such things as reliability indicators. As we will see, even if this is the case, major problems emerge for computational reliabilism for in the case of the reliability of artifacts, these indicators happen to be found elsewhere from where reliabilist accounts conventionally look for them.

3.2. The challenge of the indispensability of endogenous features in artifactual reliability

As we saw above, in reliabilism, whether something is reliable or not does not depend on an agent’s internal evidentiary threshold for justification to deem it so, as evidentialist views suggest (Goldman, 2011). Rather, under views of this type, the reliability of a process can be determined solely by considering exogenous elements of a process. In the case involving conventional epistemic agents, such as ourselves, and our claims to knowledge, reliabilism is supposed to give us a theoretical framework that makes sense of the intuition that we seem perfectly capable of justifiably relying on either our own perceptive systems or on the testimony of others without necessarily having to invoke exhaustive mechanistic details about how either of them functions.[6]

As we saw in section 2 above, under views like these, the degree of reliability of a process depends on its proclivity, or lack thereof, to produce truthful outcomes and not on any internal properties of the process itself or on whether or not agents find them sufficiently justified to subscribe to them (Goldman, 2011; Goldman and Beddor, 2021). Process reliabilism extends this seemingly virtuous omission of conventional reliabilism to considerations about results or assertions that require or come from a procedure carried out by an agent or set of agents. Importantly, process reliabilism points towards exterior—often environmental—factors that can make or break an epistemically conducive setting. There is no point under these type of views, for example, in citing an agent’s good eyesight as a reliability indicator on a severely foggy day. It is the foggy day, or the lack thereof, that determines whether or not we have a reliable sighting; famously, seeing a farm from afar in a county littered with fake farm facades as tourist attractions is not sufficient to conclude that a farm has been sighted. Rather, a process by which a piece of knowledge could be relayed or come to be known under these circumstances has to be

[6] On a more common sensical sense, this safeguards the intuition that one does not have to know *that* one knows in order to know something and that one does not need to know *how* one knows something in order to know something, e.g., knowing that something was seen does not entail/require knowing how eyes, brains, nervous systems, etc., work.

reliable and the factors that deem it so are mainly external, exogenous to the agent in question, their inner workings or those of the process itself. Whiel pointing in the right direction, these examples fail to show that what reliabilism was really trying to keep at bay was the internalist notion that an agent's inner sense of justification was all that sufficed for them to claim knowledge. Relibailism, in other word, is an externalist epistemic framework. [7]

Durán and Formanek (2018) try to safeguard this intuition of reliabilism in *computational* reliabilism. They suggest, as noted above, that computational reliabilism uses "reliability indicators (RIs) as markers of methodological and epistemological competence of the computer, algorithm, and social processes involved in the formation of beliefs." Importantly, according to Durán and Formanek (*ibid*), these indicators, which "can be understood as algorithmic-related methods and practices" (Durán, in this volume p.7), are nevertheless to be considered "exogenous to the algorithm." This is in large part what makes computational reliabilism a reliabilist epistemological account in the first place.[8] *And*, it is also what allows it to be positioned, as we saw in previous sections, as a plausible solution to the challenges posed by the severe opacity many say to be characteristic of computational methods (Humphreys, 2009), particularly machine learning and other AI technologies (Burrell, 2016; Alvarado, 2020; 2021; 2022). In other words, one of the main appeals (if not *the* main appeal) of computational reliablism vis-à-vis computational methods is that it is not supposed to need to take into consideration the inner working of a process, its endogenous features, in order to provide justification of our reliance on its results. Hence, the opacity— in its many varieties and however severe— of the inner workings of a system, device or process is supposed to be a non-challenge. This virtuous omission of computational reliabilism is what substantiates the claims in the literature that nobody really is, and nobody should be, afraid of black boxes (Durán and Jongsma, 2021), even —or particularly— if lives are at stake (London, 2021).

Yet, a few issues for this view immediately emerge. First, the strong external/internal distinction that conventional reliabilism applies to considerations involving naturally-occurring epistemic agents, such as us, as well as to considerations regarding our surrounding environment, may not necessarily apply to considerations involving artifacts, or to processes generally understood as artifactual [9], or even to their products: be these propositions, results, computations, etc.[10] This is

[7] This will become important at the end of this chapter when we briefly discuss the fact that computational reliabilism, as championed by Duran and coauthors, seems to be more of an evidentialist non-externalist epistemic account and not an externalist reliabilism after all.

[8] It must be noted that Durán's insistence that the notion of reliability must be acknowledged as a social construct risks taking computational reliabilism more in the direction of evidentialism, a view which invokes an epistemic agent's internal evidential thresholds as relevant determinants in the assessment of claims.

[9] Science in general and methodology in particular can be thought of as artifactual processes of inquiry: like technical artifacts, scientific methodology is designed and implemented with an aim, intention and hence a function.

[10] There may be something to such an externalist/internalist demarcation in cases in which agents such as ourselves are involved—e.g., perhaps due to phenomenological, biological or cognitive features that clearly separate us from our non-organic surroundings. There is us, the cohesive organism with reflective cognitive capacities, and there are certain *external* conditions that ought to be met, processes to be carried out, in order for someone to be in a position where they can be said to know something (see the barn example above). That this demarcation—as fuzzy as it may be— can be meaningfully made at all in the case of technical artifacts is not immediately obvious.

because technical artifacts are constituted by features without which they could not carry out their designed or adopted central function and which emerge in virtue of material or design properties that can only be seen as endogenous—i.e., arising from within. At the same time, their designed/assigned tasks also determine what these essential arrangements and materiality will and can be. How they were made, what they were made of, and how they do what they do in virtue of how they were made, what they were made of and what they were made for, are indispensable considerations to fully capture the nature of artifacts (Kroes, 2002; Symons, 2010) and hence for understanding their functional capacities and limitations (Alvarado, 2023b). Understanding their capacities and limitations is, in turn, essential for reliability assessments and for properly-grounded trust allocation (Simon, 2010; Alvarado, 2023a).[11] As endogenous features, however, computational reliabilism is not supposed to need them to assess the reliability of a process, system, technology or their results. Yet ignoring them seems at best epistemically questionable, at worst epistemically irresponsible (Winsberg, et al, 2022).

Consider a medical doctor telling you that a given pharmaceutical cures acne 95% of the time but could kill the other 5%. Also, imagine that when asked about what information there is about the 5%, their reply is that there is not any more information beyond the rate and the extent of the risk: we know it fails this often, and we know that failure often implies death. Consider further that they tell you that they deem this substance to be very reliable because at least they know both the rate *and* the extent of failure. Consider even further that, rather than simply assuming the reliability of the process by which the medicine was brought about, the doctor cites some reliability indicators similar to those in computational reliabilism to justify their reliance on such a medicine. That is, rather than simply appealing to an epistemic entitlement and telling you that you should trust them, they insist that some endogenous factors should serve as reliability indicators to you as well. They tell us that not only does the medicine yield reliable results most of the time, but that the processes by which this medicine has been designed, developed and deployed are also reliable because they represent "methods (formal or otherwise), metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts that might increase the degree of warrant we have to believe the in the results of [such medicine]." The pill was developed in a world-renowned laboratory, it was produced with the best equipment and sanctioned by the most credible people. Imagine further that if you asked about how the most credible people sanctioned the pill, the answer is that you do not really need to know but if you insist and must know, they did it by testing it on roughly a hundred random people. Five of which died and ninety-five who got better. Once you take all these details into consideration, contemplate whether or not the question of whether this substance *is* reliable has been answered. In other words, can we still ask 'yes, but is this substance reliable?" without being conceptually confused? We can go back to the atomic bomb example and the cases of warrant transmission discussed in the previous section for reference: the answer is yes, the reliability of such substance has not yet been determined.[12]

[11] This is further in alignment with some developments in the epistemology of trust in science and technology which (rightly) suggest that in order for trust to be appropriately allocated, or 'grounded', the reasons that justify such trust must align with actual properties, namely capacities, of the trustee (Buechner, et al, 2013; Simon, 2020; Oreskes, 2021 p.55; Alvarado, 2022b; 2023b).

[12] It is in cases like these that a confusion may emerge between questions about reliability and trustworthiness. What the doctor is alluding in this case may be reasons for trust. They are citing trust indicators but not necessarily reliability indicators. Although tis is of course problematic, as we saw in the previous section, for the sake of

The argumentative element of the example above comes from the fact that the question—even if Durán's considerations are true—is not already answered and asking it continues to make sense. In other words, even if a technology comes from methods (formal or otherwise), metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts that are alleged to increase the degree of warrant we have to believe in the results of such technology, whether or not that technology itself, both in kind or in token, is *in fact* reliable, remains an open question (Symons and Alvarado, 2019; Alvarado, 2021; 2023a). This signals at the very least that these exogenous details are not sufficient to determine the reliability of such an artifact, and/or that there must be something else, perhaps something more, that does so.

Here is yet another example that illustrates this point. Consider the recent announcement by Google's GraphCast researchers that GraphCast has "significantly outperformed conventional weather forecasting methods in predicting global weather conditions up to 10 days in advance." (Edwards, 2023) Some may think that knowing that this system outperforms conventional weather prediction tools 90% of the time (Lam et al., 2022; 2023) could suffice to deem it as reliable or even as *more* reliable than the conventional tools it is being compared to. Nevertheless, notice that in such a context, since it is the kind of computational methods that is likely to be relied upon in order to make safety-critical decisions, knowing the source and nature of error— that is, why and to what extent it fails—becomes a non-trivial, indispensable consideration. The number of resources required for weather related interventions are such that decision-makers must be able to both prepare for and constraint their deployment with precision. Just knowing how often a system like this fails is not enough to assess its reliability. Further, note that in such cases, even just knowing how bad the system fails is not enough. It may prove irrelevant, for example, that the benchmark evaluation of one method over the other ranks as more accurate in average by 1 degree Celsius in normal weather, if once every month it is off by 10 degrees in the case of unpredictably weather events, which is when these systems are the most needed/useful. [13] What would help deem this system reliable is knowing how and why it does better, where it can fail and why it can fail, and what about its inner workings makes it vulnerable to unreliability: in other words, knowing about its endogenous properties.

Importantly, considering these aspects of a technical artifact is not only key in life-critical settings, but also in knowledge-critical processes—those processes in which acquiring, building on, or expanding genuine knowledge is the essential central task.[14]. In such contexts, questions

argument we can say that you may have reasons other than the actual/factual reliability of the substance to trust it: the reliability of others, social pressure, lack of alternatives, etc. As pointed by Dretske (2000), these are simply not epistemic reasons, so even if they may be adequate for non-epistemic contexts or technology (e.g., in medicine), they are simply inadequate for epistemic trust in epistemic contexts (Alvarado, 2023b).

[13] Thanks to Mats Krüger Svensson whose comment on the news aggregator website HackerNews inspired this example.

[14] In some foundational scientific projects, for example, in which the aim is to simply know more about a phenomenon in order so that the underlying mechanisms of other higher-level phenomena could be elucidated and hence understood, simply knowing that a process yields consistent results the majority of the time does not suffice. In this context the task is to know more so that we can know more. Hence, if we fail at identifying the elements that could undermine our capacity to know more, then we are failing at trying to know more and our ability to know more about the higher-level phenomena is on flimsy epistemic grounds.

such as ‘why/how/where did the error happen?’ become epistemically relevant even when the technology functions correctly “most of the time.” Particularly if we want to *appropriately*, that is justifiably, rely on it.[15] In a sense, this is the one thing that sections in scientific papers devoted to methodology seek to elucidate: that the methods used were crafted in such a way that those considerations that could directly undermine the validity of their results have been sufficiently addressed and hopefully assuaged. Importantly, when addressing the possibility of error in a technical artifact in these kinds of contexts, as we saw in the examples above, this does not simply mean gathering a rate of error, or even quantifying a magnitude of error when it does occur. Rather, it means having a general understanding of the nature and the source of such error such that we can speak to it, track it, and hopefully address it in future iterations of the process in question. These features, again, seem like centrally endogenous and not merely exogeneous features of both the technical artifact and the process by which it was made.

Hence, when it comes to understanding the capacities and limitations of a technical artifact, elements such as its materiality, material arrangement, structural integrity and hence possible points of failure are indispensable features to consider, e.g., a hammer made of a single piece of cast iron will be able to perform distinct tasks in distinct settings from a hammer made of plastic or even a hammer made of more than one part. In the case of computational methods, hardware and software architectures, the way code functions are timed, the kinds of errors these architectures and specific programs are prone to, and the kinds of languages and/or techniques deployed to perform a function must be considered. These material, structural, or functional elements can be categorized as constitutive of—or at least coextensively defined with—“algorithmic-related methods and practices,” yet whether these elements are exogenous or not to the algorithm itself, external or internal to a process or results whose reliability is under consideration, is not entirely clear.

And yet, whether these features are endogenous or exogenous, they must be accounted for. As briefly noted above, in the case of certain artifacts —computational or not— namely those used in both safety-critical settings, like medicine, and those used in foundational epistemic endeavors, such as science, endogenous features, such as the nature and source of failure points, become even more indispensable. This is not unique to computational methods, there is a reason why precision instruments and not less expensive alternatives are the norm in laboratories: because it is epistemologically relevant to consider endogenous properties of an artifact to deem it epistemically reliable (i.e., they serve epistemological values such as standardization, reproducibility, objectivity, etc.). Similarly, there is a reason why medical equipment is expected (and often forced) to undergo rigorous vetting and sanctioning processes before it is introduced into medical practice: because it is both medically and morally relevant to do so. Yes, it is an epistemic good to make sure that we can rely in the results of our instruments as we chose a path forward in a patient’s treatment, and yes, we have an epistemic responsibility not to take the reliability of systems that can affect people’s lives or well-being lightly (Harvard and Winsberg,

[15] The emphasis on appropriate justification for an epistemic technology is important here because, as a practical matter, we can and do rely on things for which we have no proper justification. We do this in certain instances of crises, but also in most our everyday activities. That we can do so, and that we do so, however, does not do much to the plight that in some circumstances doing otherwise is what should be done. In contrast with Durán and Formanek, I find little use in accommodating existing scientific practices in a normative inquiry. The reasons why we could or should rely on an artifact are not appropriately informed by the reasons why people actually do so.

2021; Winsberg et al., 2022; Harvard et al., 2022), but importantly, we are also morally required, in settings such as medicine or civil engineering, to safeguard people from error that may undermine the function of such systems. Doing so requires us to not neglect properties and features endogenous to both the artifact and the processes by which it was built. If this is the case then, as we will see, either computational reliabilism cannot account for such features and it is therefore an inadequate epistemological framework to deal with artifacts of any sort but particularly computational artifacts; or, computational reliabilism can and hence must account for such endogenous features, in which case it must directly address the challenge of epistemic opacity of a process' endogenous features and not merely circumvent it.

3.3. The problem of error-related essential epistemic opacity

In this section I aim to show that what is at stake in instances of epistemic opacity is not merely the inaccessibility to essential features of how a process *works*. This has been a central misunderstanding within the literature. Rather, what is at stake is the opacity of how such a process may *fail*. The inability of computational reliabilism to take endogenous features into consideration makes this challenge a serious problem for the viability of the framework as a solution to the epistemic opacity of computational methods. However, as we will see, the insurmountable essential epistemic opacity of certain kinds of error in computational methods such as machine learning— particularly neural networks, which are at the center of more recent developments in large language models and generative pre-trained transformers—makes it so that the challenge is even more severe than mere conceptual inadequacy on the part of the epistemological framework.

In the face of the challenges posed by the *essential epistemic opacity* of computational methods—elucidated first by Humphreys (2009) and later expanded by others (Burrell, 2016; Alvarado and Humphreys, 2017; Alvarado, 2020; 2021)— philosophers have come up with argumentative strategies that make sense of the fact that inquiry, even the inquiry that makes use of opaque methodology, can nevertheless be successful.[16] Some of these strategies aim to undermine the severity, both conceptually and practically, of the challenge of epistemic opacity. San Pedro (2021), for example, suggests that epistemic opacity can be contextually assessed as a

[16] I take it that the reader is sufficiently acquainted with the nuance surrounding the concept of epistemic opacity, particularly as it relates to its kinds, its sources, and their respective strengths and implications (For a thorough review of this concept see Alvarado, 2021). Hence, here I will only focus on the essential epistemic opacity particular to computational artifacts and not on the general sense of opacity first identified by Humphreys (2004) and later expanded on by others to include social and contingent sources of opacity: e.g., scientific infrastructure, technical literacy (Kaminski et al., 2017), state and corporate secrecy (Burrell, 2016), natural resources, etc. Elsewhere (2020; 2021;2023a), heavily borrowing from the work of Symons and Horner (2014a; 2014b) and expanding on Humphreys' (2009) views, I have argued that in the context of computational artifacts, essential epistemic opacity is uniquely— even if not exclusively or exhaustively— at play. This is particularly the case in AI/ML systems, whose architecture and analytic dimensionality further exacerbates the 'too many everything problem" (Alvarado, 2020).

matter of degrees and that certain mitigating procedures can ease its impact.[17] Similarly, after a thorough assessment of different kinds of sources for opacity in machine learning technologies, Jenna Burrell (2016) offers a series of best-practice remedies to assuage it. Others, mainly practitioners, simply deny that essential epistemic opacity is the case and offer evidence that suggests to them ways to make certain methods—for example, in machine learning—'interpretable.' Still others, accept that essential opacity is the case but deny that it makes a significant difference to the aims of inquiry (Durán and Jongsma, 2021; Duede, 2022). [18]

All of these argumentative strategies acknowledge that those seriously concerned with epistemic opacity have a certain commitment to the value of what is taken to be its opposite: transparency. And this is, to a certain extent, true. Transparency, however, can mean different things in different contexts, e.g., access to source code, surveyable or interpretable practices and components, explainable processes, etc. What is often assumed is that some level of access to 'how something works' or 'how something produces results' is what is at play. Hence, as a way of undermining the challenge of opacity, some of these strategies aim to undermine the value of such a commitment to transparency. One may argue, for example, that it is simply unsound to think that every single part of a process ought to be maximally accessible (intelligible, surveyable, etc.,) to an agent in order for them to successfully rely on it. If this is the idea of transparency, one may argue, then we go back to thinking about opacity in such general terms that it ultimately trivializes the concept in at least two ways: a) everything is opaque to everyone at one level or another, and if so, b) then opacity does not seem to pose any meaningful challenge after all since life and inquiry goes on successfully despite its ubiquity.

Additionally, it can be pointed out that while it is easy to understand *some* level of transparency of these methods as a desirable epistemic good, it is still possible to undermine the normative import of this good by diminishing the normative weight of this transparency: i.e., one could always argue that while it is *good* to have transparent methods, this transparency is by no means an epistemic necessity nor an obligation. It is well known, for instance, that transparency is not *sufficient* for explainability; someone could easily further argue that its *necessity* is equally questionable.[19] Even more devastatingly, it could be argued that the kind of maximal

[17] This, of course, ignores or neglects the fact that the concept of essential epistemic opacity implies that it is *impossible* for agents of a limited epistemic nature to overcome it. This is a topic I address elsewhere (Humphreys, 2009; Alvarado, 2021).

[18] Although some philosophers walk a fine line between the last two strategies, computational reliabilism is often championed as promising a way forward in line with the latter argumentative strategy: i.e., even in the case of essential epistemic opacity, so this strategy supposes, we can be justified in our reliance on certain computational methods and hence also in our trust on their results. Duede (2022), for examples justifies our reliance on deep learning techniques in science by attempting to contextualize its use in broader context of discovery. This is, as is often the case, yet another pragmatic defense for the utility of such an artifact. As I have argued here and elsewhere, if scientific inquiry is intrinsically an epistemic endeavor and not just merely a problem-solving one, this pragmatic approach is simply inadequate and can be reduced to an attempt to justify practices because 'practitioners practice them' and have happened to gain some utility from them, and not as I see it a critical epistemological inquiry.

[19] Full transparency of a complex and large source code, for example, in many instances of interest would not yield anything resembling interpretability or explainability. In other words, one can have access to all the details of a system without necessarily understanding them. Hence, transparency is not sufficient for understanding how something works.

transparency alluded to above could, in fact, get in the way of any actual epistemic good. A barrage of source code, for example, could obscure the functionality of a piece of software to someone investigating what it is supposed to do.

One way to counter this last point is too simply say that not every single aspect of a process of a system *can/must* be relevant. We must discern between those things that ought to be accessible to an agent and those that do not need to be. Consider the following, according to Humphreys, a computational system, process or device is *essentially* epistemically opaque iff

> "it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process." (2009)

Researchers of opacity have either taken this definition at face value and focused on the strength of the claim, i.e., the impossibility clause, or, as we saw above, on what the term 'all' implies. However, rather than focusing on the 'all' clause in the definition above, one can also focus on what Humphreys may have meant by the 'epistemically relevant' aspects of a process and try to see what those are. In a formal proof—be it in mathematics, geometry or logic— for example, one may find that *every* inferential step towards a result and hence every premise that informs such steps is an epistemically relevant element of the process. Durán and Formanek (2018) rightly point out, however, that in the context of computational methods, it is not immediately obvious which are the epistemically relevant elements of a process: is it every single line of code? Elements related to the overall aims and purpose of a software system? Or do the complete functions of underlying components such as compilers, circuitry, etc., count as epistemically relevant? Regardless of these difficulties, however, advocates of computational reliabilism claim to circumvent them by pointing out that transparency is simply not needed, and that having it may not do any good in the contexts where philosophers conventionally demand it. Although computational reliabilism does not simply assume the reliability of the methods in question, it can still make this latter claim by arguing, as we briefly saw, that there are some reliability indicators to help us ground our confidence and trust.

All of these views are partially correct: yes, the value of transparency is often implied in those that see opacity as a genuine epistemic challenge; yes, the value of this transparency is not immediately obvious, particularly since it seems that full transparency of all the elements of a process is neither sufficient nor necessary for explanatory endeavors; furthermore, even if such transparency is an epistemic good, in many circumstances of interest, the value of these goods may be outweighed when considered against exceptional ethical, prudential and/or practical urgencies. It could even be said, without much controversy, that it is true that prioritizing epistemic concerns in certain circumstances of ethical, prudential or practical relevance may prove frivolous, irresponsible or even reprehensible.

However, as we saw in the previous section, while we do not need a list of all the epistemically relevant elements of the technology being deployed, we definitely *need some very specific ones*, namely the elements related to error. As I have noted elsewhere "While the contents of an exhaustive list of epistemically relevant elements of a system may be debatable, that error and error assessment must be included is less controversial." (2022a) For example, while we may be able to forgo knowing exactly how an Aspirin *works* when we deploy it —as suggested by London (2019)— what a medical professional cannot forgo when deciding on its use, or the use of any pharmaceutical technology, is having a good idea of *how and why it could fail* and not just the rate at which it fails.

While knowing that a given medical practice could fail (where failing implies the loss of life or the loss of quality therein) can help an individual calculate the risks they are willing to take on their own life or those of others, or help an institution to calculate the mortuary and legal resources they will need when deploying a given technology, thinking that these considerations suffice in such a context is at best limited and at worst seriously misguided. A medical practitioner deploying these pharmaceuticals should know not just how often they fail, but more or less when they can fail. This is more often than not informed by knowing why they may fail, which is in turned informed in part by the chemical substances *in* the pharmaceutical, their relative ratio to one another and even the order in which they were mixed. Hence, they also need to know why and how it can fail. In other words, they need to have an idea of the sources and the nature of possible errors.

Simply stated, when it comes to technical artifacts in formal epistemic contexts such as rigorous scientific inquiry some things cannot *not* be in a list of relevant epistemic elements of *any* process: these include the nature and source of its possible errors.

Importantly, this is different from requiring a system and its inner workings to be fully transparent, or from requiring a full explanation of all the epistemically relevant steps a system takes to achieve a function. While the reliability indicators associated with computational methods such as machine learning suggested by Durán include "methods (formal or otherwise), metrics, expert competencies, cultures of research, and the like that make up for our best epistemic and normative efforts that might increase the degree of warrant we have to believe the outputs of ML systems," (2023, P.5) and while there is a sense in which the complete epistemically relevant elements of a given process remain elusive, there is an undeniable sense, as noted above, in which in certain settings knowing something about the ways in which something could fail is an indispensable element of *any* list of reliability indicators.

Hence, so much is clear: the heart of the problem of epistemic opacity is not necessarily not knowing the details of how something works. The real problem is access to the nature and source of error in order to run an adequate reliability assessment.

The problem for computational reliabilism is thus twofold. The first issue is that they must, as we saw in the previous section, concede that at least some of the internal details of a process or device qua process or device *must* be known in order to appropriately rely on it. Note how this is in sharp contrast to the view that computational reliabilism can rely only on elements exogenous to the process/device itself to ascribe reliability. If this is so, then the main appealing feature of computational reliabilism—namely that it can ensure an appropriate epistemic framework even without appeals to transparency or without caring for opaque elements of a process—seems to be no longer. The second problem is, that if they do concede to the points above, then they must face the challenge of *essential* epistemic opacity concerning the nature and source of error in most computational systems of interest, and particularly those involving machine learning.

In particular, these views have failed to acknowledge the severity of the challenges posed by the irreversibility of probabilistic, stochastic models whose results are multiply realizable and overdetermined (Symons and Boschetti, 2013). They also neglect to consider the catastrophic/strange nature of error in ML technologies such as deep neural networks (Alvarado, 2022). Both of these are instances in which tracking the source of error and the nature of such error is simply unachievable. In the first case, the fact that a computational system may take several distinct paths to reach its results even in the case of identical input, makes it so that determining which path was actually taken is not possible (Symons and Boschetti, 2013). If this is the case and there is an error, tracking where the error happened is simply not achievable. This is a particularly severe challenge in the case of computational technologies such as deep neural networks and transformer models which require a significant number of layers of analysis.

In the second case, the kinds of error that these computational methods are prone to, are simply too extreme and unpredictable. This is exemplified by issues such as hallucinations in generative AI (Ji et al., 2023; Lee, 2023), but also in the case of adversarial attacks to neural networks in which a random positioning of a single pixel could alter the analysis of an image. Importantly, in recent research similar results to these adversarial attacks have been found to emerge even in non-modified images that involve complex textures such as those of natural foliage or intricate fence designs. Elsewhere (Alvarado, 2022) I note the following:

> "There are at least thousands of naturally occurring examples of images that contain information that works just as an artificially induced adversarial attack would and that can mislead well-known image classifiers. Furthermore, it has been shown that even well-established defense strategies that ensure some widespread image classifiers remain resilient versus some artificial adversarial attacks are nevertheless unable to defend against naturally occurring examples."

As further noted there, these naturally occurring adversarial examples can include simple elements such as "inclement weather conditions and obscured objects, and it can also include objects that are anomalous." (Hendrycks et al, 2019) This means that the neural network does not need to be intentionally targeted to produce unexpected error. Rather, features intrinsic to natural images can make neural networks yield unpredictable errors. Although this signals that we can at least recognize certain patterns that elicit this kind of error, the fact that non-manipulated images of random scenes and objects can cause this kind of error means that there can be an infinite number of patterns out there that can do the same in a myriad of fields: telescope imagery, x-rays, etc. Non-surprisingly, this kind of issue can even be found in cases of text-based analysis (Zhang, et al, 2023). If this is so, then any kind of machine learning data analysis may be prone to such a problem. Interestingly, these errors are "different "bugs" from traditional software" (Sun et al., 2018) Moreover, this is not the only way in which errors of this type in these types of systems are opaque. For example, due to their highly distributed computing, neural networks also fail in what is called a 'graceful' manner with minute deviations from weigh to weigh and layer to layer (Alvarado, 2022). This makes it so that such errors carry on distributed in the very many weights of deep neural networks and the automated optimization changes that these systems undergo in their many layers of analysis. And lastly, the arrangement of these networks structures is such that this error is not easily detected either internally by the network or externally by those supervising it. Thus, these errors are delivered with the same degree of confidence as truthful results.

This last point illustrates, once more, the essential epistemic opacity of computational methods and the severity of the challenge it represents for epistemic frameworks seeking to justify our reliance in such methods, particularly when they are used in context where reliability is crucial. The section before this one showed the indispensability of endogenous features of processes and devices to the reliability assessments of artifacts. This discussion also showed the further indispensability to reliability assessments of the nature and source of error, particularly in knowledge-critical and safety critical contexts. In this section, we saw that it is precisely these two elements, the source and nature of error that matters most when we talk about epistemic opacity. The last point in the paragraph above illustrates that such a challenge in the context of computational methods such as deep neural networks may very well be, as Humphreys (2009) once suggested of other computational methods, *essentially* insurmountable and thus, a genuine challenge to computational reliabilism. Importantly, the essentiality of such an opacity need not arise from internal features of an epistemic agent such as its nature, as Humphreys suggested. Rather, it can be traced to endogenous features of the system in question and as such be agent-neutral and agent-independent. That is, they arise in virtue of features that have little or nothing to do with an agent's epistemic limitations (hence, agent-independent) and it is the kind of opacity that would apply to any given agent with sufficiently similar epistemic limitations (agent-neutral) (Alvarado, 2021).

4. **Conclusion**

In this chapter I have presented three distinct yet interrelated challenges to computational reliabilism and its viability as an epistemic framework that can provide justificatory grounds for our reliance in novel and opaque computational technologies, particularly those related to novel AI methodology such as machine learning. In particular, I argued that the challenges related to warrant transmission, the indispensability of endogenous features such as the source and nature of error in reliability assessments, and the ultimate opacity of the latter, as exemplified by the irreversibility of such systems and the intractable nature of errors related to adversarial disruptions, represent a serious problem for the viability of computational reliabilism. With regards to the second of these challenges, I also showed that it is the same virtues for which the epistemic framework is lauded that make it incapable of accounting for such endogenous features as the source and nature of error in an artifact. With regards to the third challenge, I also showed that, contrary to conventional interpretations in the literature, what is at play in the challenge of epistemic opacity is not just any kind of general transparency regarding how something works, but rather the access—or lack thereof— to the nature and source of error. In doing so, I hope to have also shown that these challenges are not only serious but maybe even insurmountable for such an epistemic framework as computational reliablism as currently articulated in the literature. Nevertheless, I do not aim for this analysis to constitute a final devastating critique of computational reliabilism. My hope is rather that, along with other recent efforts (Smart, et al., 2021; Ferrario, 2023), these observations serve as a starting point to rescue it from its present inadequacy.

Bibliography

Alston, W. P. (1995). How to think about reliability. *Philosophical Topics*, *23*(1), 1-29.
Beutel, G., Geerits, E., & Kielstein, J. T. (2023). Artificial hallucination: GPT on LSD?. *Critical Care*, *27*(1), 148.

Alvarado, R. (2023). *Simulating Science: computer simulations as Scientific instruments* (Vol. 479). Springer Nature.

Alvarado, R. (2022). What kind of trust does AI deserve, if any?. *AI and Ethics*, 1-15.

Alvarado, R. (2022). Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics*, *36*(2), 121-133.

Alvarado, R. (2021). Explaining epistemic opacity. (Preprint) http://philsci-archive.pitt.edu/id/eprint/19384

Alvarado, R (2020) Opacity Artificial Intelligence, Machine Learning, Big Data and Democratic Processes. Chapter in *Big Data and Democracy,* Macnish, K., & Galliott, J. (Eds.). (2020). Edinburgh University Press.

Alvarado, R., & Humphreys, P. (2017). Big data, thick mediation, and representational opacity. *New Literary History*, *48*(4), 729-749.

Beisbart, C. (2017). Advancing knowledge through computer simulations? A socratic exercise. In *The Science and Art of Simulation I: Exploring-Understanding-Knowing* (pp. 153-174). Springer International Publishing.

Buechner, J., Simon, J., & Tavani, H. T. (2013). Re-thinking trust and trustworthiness in digital environments. *11th Computer Ethics: Philosophical Enquiry (CEPE 2013)*, 1-15.

Burge, T. (1998). Computer proof, apriori knowledge, and other minds: The sixth philosophical perspectives lecture. *Philosophical perspectives*, *12*, 1-37.

Burge, T. (1993). Content preservation. *The philosophical review*, *102*(4), 457-488.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Comesaña, J. (2010). Evidentialist Reliabilism. *Noûs*, *44*(4), 571–600. http://www.jstor.org/stable/40959693

Dretske, F. (2000). Entitlement: Epistemic rights without epistemic duties? Philosophy and Phenomenological Research, 60(3), 591–606.

Duede, E. (2022). Deep learning opacity in scientific discovery. *Philosophy of Science*, 1-13.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*(4), 645-666.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

Edwards, Benj (2023) AI outperforms conventional weather forecasting for the first time: Google study. https://arstechnica.com/science/2023/11/ai-outperforms-conventional-weather-forecasting-for-the-first-time-google-study/

Ferrario, A. (2023). Justifying our Credences in the Trustworthiness of AI Systems: A Reliabilistic Approach. *Available at SSRN 4524678*.

Goldman, A. I. (2011). Toward a Synthesis of Reliabilism and Evidentialism?: Or: Evidentialism's Troubles, Reliabilism's Rescue Package. *Evidentialism and its discontents,* 254-280.

Goldman, Alvin and Bob Beddor, "Reliabilist Epistemology", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2021/entries/reliabilism/

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

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). Natural adversarial examples. *arXiv preprint arXiv:1907.07174*.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615-626.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Kroes, P., & Meijers, A. (2002). The Dual Nature of Technical Artifacts-presentation of a new research programme.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A., ... & Battaglia, P. (2022). GraphCast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

Lee, M. (2023). A Mathematical Investigation of Hallucination and Creativity in GPT Models. *Mathematics (2227-7390)*, *11*(10

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15-21.

McDermott, M. B., Wang, S., Marinsek, N., Ranganath, R., Foschini, L., & Ghassemi, M. (2021). Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, *13*(586), eabb1655.

Simon, J. (Ed.). (2020). *The Routledge handbook of trust and philosophy*. Routledge.

Smart, A., James, L., Hutchinson, B., Wu, S., & Vallor, S. (2020, February). Why reliabilism is not enough: Epistemic and moral justification in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 372-377).

Sun, Y., Huang, X., Kroening, D., Sharp, J., Hill, M., & Ashmore, R. (2018). Testing deep neural networks. *arXiv preprint arXiv:1803.04792*.

Symons, J., & Alvarado, R. (2019). Epistemic entitlements and the practice of computer simulation. *Minds and Machines*, *29*(1), 37-60

Symons, J., & Boschetti, F. (2013). How computational models predict the behavior of complex systems. *Foundations of Science*, *18*, 809-821.

Symons, J. (2010). The individuality of artifacts and organisms. *History and philosophy of the life sciences*, 233-246

Van Helden, A. (1994). Telescopes and authority from Galileo to Cassini. *Osiris*, *9*, 8-29.

Winsberg, E., Brennan, J., & Surprenant, C. W. (2020). How Government Leaders Violated Their Epistemic Duties During the SARS-CoV-2 Crisis. *Kennedy Institute of Ethics Journal*, *30*(3), 215-242.

Winsberg, E., & Harvard, S. (2022). Purposes and duties in scientific modelling. *J Epidemiol Community Health*, *76*(5), 512-517.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1-41.