Draft. Comments are welcome.

CAN WE TEST AI LIKE WE TEST DRUGS? A GENERATIVE ANALOGY BETWEEN MACHINE LEARNING AND CLINICAL TRANSLATION

Emanuele Ratti[1], Department of Philosophy, University of Bristol

Lena Zuchowski, Department of Philosophy, University of Bristol

Abstract. In the past few years, machine learning (ML) has been widely (and to an extent, successfully) implemented in medicine. However, uncertainties surrounding ML have made it difficult to establish the basis of its epistemic warrants. In the literature, a parallel has been drawn between medicine and ML, suggesting that we should model epistemic standards for ML on the standards of clinical translation. By developing tools from Hesse's (1966) work, we characterise the nature of this parallel as a generative analogy between the process of clinical translation and the process of building ML systems. We identify more precisely the epistemic warrants of clinical translation that are typically only mentioned when appealing to the analogy, and we show in which sense such warrants apply analogically to the context of ML. In particular, we interpret the epistemic warrants of clinical translation in reliabilist terms, and we show how this can inform a new form of ML reliabilism, which is compatible to existing reliabilist accounts in philosophy of AI.

Keywords: clinical translation; reliabilism; machine learning

1 INTRODUCTION

In the past few years, philosophers and machine learning (ML) practitioners alike have discussed to what extent we can rely on (or trust[2]) complex computational systems like ML tools, in particular, given the (ineliminable or essential) opacity and the uncertainties characterizing them (Alvarado and Humphreys 2017; Boge 2022). This question has been posed especially in the context of medical ML. In an influential article, London (2019) has proposed to treat these predicaments in the same way similar problems are handled in medicine itself. Medical knowledge itself is severely fragmented, and it can be characterized as "atheoretic, associationist, and opaque"

[1] mnl.ratti@gmail.com

[2] For the present article, it does not matter whether one is talking about trust or reliability, as discussions on trust assume reliability of ML tools as a precondition.

(London 2019, p 17). A paradigmatic case is our knowledge of pharmaceuticals, which is characterized by multiple unknowns, such as the uncertainties regarding mechanisms of action[3] or lack of theoretical justification. But in medicine these unknowns are epistemically sidestepped by methodologies that have been proved effective in establishing reliable results. Given that with medical ML (and ML in general) we are facing a similar situation of atheoreticity, associationism and opacity, he draws a parallel between the warrants given by methods used in clinical translations with our prospects of relying on ML models/systems (London 2019). This parallel is not limited to London's work, and its popularity can be appreciated in the explosion of calls for more randomized controlled trials (RTCs) for medical AI, which is an implicit recognition that we should apply to ML roughly the same epistemic methodologies we apply to clinical translation (Genin and Grote 2021). But although this parallel can potentially be articulated into a concrete guide to overcome significant challenges raised by opacity or lack of methodological standards in ML, it has not gone beyond the level of a fascinating suggestion.

In this article, we analyse this parallel by interpreting it as a *generative analogy*. In Section 2, we clarify what we mean by 'generative analogy' by constructing an account based on Hesse's work (1966). In Section 3, we interpret the epistemic warrants provided by the first of the two analogues (i.e., the process of clinical translation) through a reliabilist framework. In Section 4, we systematically investigate to what extent such a framework for clinical translation is applicable also in the context of ML, and how the analogy can lead to the formulation of epistemic warrants for the ML context that are compatible with reliabilist accounts already discussed in the literature (Duran 2024).

2. GENERATIVE ANALOGIES

[3] While, at first sight, mechanisms of action are not essentially opaque (Humphreys 2011), one can make the argument that a given mechanism of action can be incredibly complex, and that it is unlikely that a human agent, *qua* cognitive agent, will ever fully understand it through just a mechanism sketch or schema. It is also not unreasonable to compare the complexity of the mechanism of action of a molecule (involving millions of entities and activities) with the complexity of a ML model or the process of algorithmic optimization. So, it is possible to say that, at least in practice, mechanisms of action are essentially epistemically opaque.

In this section, we will provide a brief introduction to the Hessian account of analogies in science (section 2.1) before further developing a hitherto somewhat neglected kind of Hessian analogy, namely generative analogies (section 2.2).

2.1 Hessian Analogies in Science

An analogy is typically understood as a comparison between two entities (the analogues), which are viewed as suitably similar such that this comparison can be used to draw one or more conclusions about the properties of the analogues. Hesse (1966) distinguishes between two kinds of analogies in scientific reasoning: a *positive analogy* exists between two analogues in virtue of some properties that they have in common, in contrast, a *negative analogy* exists if an analogue has one or more properties that the other does not, or if the same property is instantiated differently in the two analogues. Both positive and negative analogies can be used in scientific reasoning to form hypotheses about additional shared or non-shared properties of the analogues. In particular, if the properties in a positive analogy imply the existence of an additional property in one analogue, then one can deduce that (barring any screen factors) the second analogue should have this property as well. A classic example of such analogous reasoning is Reid's (1785) argument (reconstructed in Bartha, 2022) for life on Mars, which states that based on a number of similarities between our planet and Mars, it is "not unreasonable to think, that those planets may, like our earth, be the habitation of various orders of living creatures" (Reid, 1785, p 24).

Hesse (1966) also provides a more nuanced analysis of analogical reasoning in terms of what she calls *vertical* and *horizontal relations*. Horizontal identity, difference or similarity relations are established between sets properties of two different analogues, while vertical identity, difference or similarity relations are established between properties of the same analogue. For the purpose of our paper, it is sufficient to review the former. Reid's (1785) reasoning about life on Mars is clearly an example of establishing horizontal relations between a set of properties (p1, p2, p3, …) on Earth and the same set of properties on Mars. In particular, the reasoning can then be paraphrased as follows: If Earth has properties (p1, p2, p3, …) and these properties cause life to flourish this planet (p4), and Mars likewise possesses properties (p1, p2, p3,...), then one is justified in drawing the inference p4 is possessed by Mars too, despite other differences between the planets. Hesse (1966) assigns such deductive reasoning from horizontal relations between positive analogies a *predictive purpose*: in general terms, if Analogue 1 (A1) has properties (p1, p2, p3, …)

that strictly imply property p4, then, if Analogue 2 (A2) possesses the same properties (p1, p2, p3, …), one is warranted to predict that A2 also has property p4.

Prediction is not the only purpose that Hesse (1966) ascribes to reasoning from horizontal relations between analogues. Such reasoning can also have a *persuasive purpose*, which Hesse (1966) illustrates by referring to the analogy between the father-child and the state-citizen relationship, which is often used to highlight "the consequences, of a moral or normative character, which follows from the relations of the four terms already known" (Hesse 1966, p 63), and thereby persuade people that this model is worth adopting.

But recognizing that reasoning from horizontal relations between positive analogues can have different purposes, opens the door to identifying further such purposes beyond the two recognized by Hesse (1966). In the following, we will argue that there is an important third purpose, namely, a *generative purpose* to analogous reasoning in science.

2.2 Generative Analogies

Consider the following well-studied episode from the history of molecular biology. By 1953, there were competing hypotheses on the mechanism behind protein synthesis, none of which involved nucleotides in any systematic way (Stegmann 2016). Watson and Crick's model of DNA (1953) inspired the well-known (at the time) physicist George Gamow (Kay 2000), who proposed to conceptualise the problem of the relation between DNA and amino acids purely as a coding problem, namely as the problem of finding the correct translation between a code with an alphabet of four letters (DNA bases), and a code with an alphabet of twenty letters (amino acids). The analogy here is between the kind of problem the scientists set out to solve, rather than between sets of properties, as in the case of the predictive analogy discussed above. The outcome of this kind of analogical reasoning is not a prediction about the properties of the second analogue A2 but the *generation of a research strategy*, which has been successful in the case of analogue A1 and, due to the positive analogies between A1 and A2, one can now reasonably expect to be successful in the case of A2 as well. For the problem of finding the correct relation between DNA and amino acids, this generative analogy to coding problems proved successful: in their analysis of this episode, Kay (2000, p. 129) writes "to molecular biology the tropes of (...) information theory, linguistics, and computer-based cryptanalysis" and thereby mobilised additional research resources from outside the field. In particular, scientists began testing different coding schemes and testing them on state-of-the-art computational resources. The analogy was therefore truly generative in the sense that it generated both a viable, novel research strategy as well as additional research based on this strategy.

A second example of generative analogy is again from biology, but this time at a more general level. The 'machine' analogy has served a great deal of purposes throughout the history of biology. Especially with regards to organisms, comparisons with machines have played three pivotal roles (Nicholson 2013). The first is theoretical, i.e. comparisons between organisms and machines have provided a foundation for the conceptualization, representation, and even explanation of organisms themselves. The second is rhetorical, i.e. as pedagogical tools for non-experts. These two uses raised a number of problems, which do not concern us here. But a third function is of interest, given that it can be described as a case of generative analogy. Nicholson (2013) calls it *heuristic*, and he refers to the fact that a comparison between machines and organisms have led to the formulation of "methodological tools that facilitate the empirical investigation of the target phenomenon" (p 674). Notoriously, organisms are self-organizing, self-producing, and self-maintaining entities, and their parts are interdependent and do not exist in isolation. This is not the case with machines, at all. But for the purposes of understanding distinct parts of organisms, these can be studied independently from the whole, exactly as it is done with machines. The analogy with machines has opened up important avenues of investigation, by abstracting "away the intimidating complexity of the broader physiological context of the organism as a whole, and focus[ing] their attention on well-defined interacting parts" (p 675). Treating organisms *as if* they were machines - analogously to machine - has stimulated the creation of important methodological strategies in fields such as molecular biology, which philosophers of science have commented at great length in the context of mechanistic philosophy (Bechtel and Richardson 2010; Craver and Darden 2013).

We will argue in this paper that the parallel between ML and medicine is best viewed as a *generative analogy*. This means that the analogy does not predict properties; rather, it generates a research strategy. In section 3, we will argue for the interpretation of this analogy as a generative one. In section 4, we will outline the research strategy we think this analogy generates.

3. A NOVEL GENERATIVE ANALOGY BETWEEN CLINICAL TRANSLATION AND MACHINE LEARNING

As stated in the Introduction, parallels between ML and medicine have been drawn, with the goal of providing a more solid ground for the epistemology of ML by exploiting specific methodological moves used in medicine. Because what is done in medicine is envisioned to be used in ML, we interpret these parallels as attempts to draw some form of analogical reasoning. In the following we will briefly outline the possible attempts to interpret this analogy (section 3.1) before arguing for an alternative, novel generative analogy between clinical translation and ML (section 3.2). Before proceeding, we should specify what we mean here by 'what is done in medicine'. In order to be aligned as much as possible to the context where the analogy was formulated the first time (London 2019), here medicine is restricted to 'clinical translation'.

3.1 Possible Interpretations of the Parallels as Analogies

If we apply the received view on analogies as expressed by the Hessian's predictive account to the comparison between the process of clinical translation and ML systems, this is what we have. The first analogue (A1), we claim, is a particular method used in clinical translation, and the second analogue (A2) is the construction of ML system. The general strategy is then to establish a predictive analogy between those two analogues by identifying a set of properties (p1, p2, p3, …) that is shared by both and then use the existence of those properties predictively (section 2.1). 'Use predictively' means to predict the existence of a fourth epistemic property that has already been established to predicate on the properties (p1, p2, p3, …) for clinical translation (A1) and should therefore also do so for ML (A2).

For example, London (2019) describes the process of clinical translation as characterised by the following three epistemic properties: associationism (p1), atheoreticity (p2) and opacity (p3). In particular, knowledge of pharmaceuticals and their effects proceed by providing evidence that strengthens the association between the use of a drug and a beneficial effect (p1); the establishment of this association does not crucially rely on theoretical knowledge (p2); and there is a pervasive lack of knowledge of mechanism of actions of many molecules which implies the impossibility of (mechanistically) explaining why they have the desired pharmaceutical effect (p3). The existence of these three epistemic properties is relatively unequivocally accepted (London 2019). These properties are seen as impediments to the mission of clinical translation.

However, despite the fact that these seem to undermine the usual process of scientific discovery and justification, it is accepted that there is mechanism by which drug efficacy can be reliably established and that get around these problems, namely Randomised Control Trials (RCTs). The fact that this is possible can be seen as the fourth property (p4) in this scenario.

To establish the parallel (which we interpret as an analogy) between clinical translation (A1) and ML (A2), London (2019) then argues that ML systems possess the same three properties (p1, p2, p3). In particular, with respect to associationism (p1), he argues that ML tools do not track causal relations (with few exceptions), but patterns and regularities, i.e., they display associationism. With respect to atheoriticity (p2), it is a defining feature of ML that those tools learn from statistical associations between data sets rather than rely on explicitly coded theories. With respect to opacity, this is also a generally recognized and much discussed (e.g., Alvarado and Humphreys, 2017) feature of ML, whereby opacity applies both to the model that is generated as a result of training a ML algorithm, as well as to the process of optimization itself (Boge, 2022). But it is important to show why p1-3, in the case of ML, raise challenges, which is something that London does not specify in great detail.

Consider a standard example of a medical ML application, a deep neural network (DNN) that classifies magnetic resonance imaging (MRI) scans as cancerous or non-cancerous. The weights given to each node in the network become intractable due to the volume of data that would be required to keep detailed records of those weights throughout the optimization process, and it remains unknown to the user how each pixel or group of pixels feeds into the decision-making process, i.e., the ML-process used in this classification process is opaque (p3). Non-exhaustive attempts at analysing the ways different areas of an image influence classification in DNNs have shown that the network is influenced by features that are not trivially obvious to human users, e.g., the classification of a scan might depend heavily on pixels not directly related to the parts of the image that would be ascribed to as ‘cancer-relevant’ according to existing knowledge and theories about cancer, i.e., the process is atheoretical (p2). The fact that the statistical association established through the network do not seem to track existing theories about cancer implies that the results of the DNN need to be interpreted as establishing mere associations between certain image properties and the existence of cancerous cell, i.e., the process is one of associationism (p1).

The case of DNN image classification is a standard example, and can be used to highlight more explicitly the epistemically problematic aspects of p1-3 (Freielseben and Grote, 2023; Grote et al., 2024). In fact, p1-3 as applied to ML, reflect classic examples of errors in ML itself (Freiesleben and Grote 2024). For example, DNN image classification is liable to the use of *shortcuts*, i.e., the establishment of associations based on non-medical information that has erroneously been encoded in the image and ultimately leads to misclassification of images not subject to the correlation exploited in the shortcut. This is because ML tools unveil associations between inputs and outputs, but these associations do not necessarily stand for more robust relations – there is always an uncertainty related to the associations that is difficult to quantify (p2). Due to the fact the DNN system is opaque (p3) and does not rely on theories that would limit the scope of relevant associations (p1), such shortcuts are usually only detected once the algorithm starts misclassifying images. Another typical error in ML are *natural distribution shifts*, which is when a mismatch between deployment and training distribution cause a drop in performance. Grote and Freiselben (2023) discuss an example based on COVID-19: imagine that a ML algorithm has learnt to classify a person as having COVID-19 on the basis of certain symptoms (e.g., cough, fever, loss of sense of smell, etc). The problem is that the virus mutated rapidly, and such symptoms might not be as prevalent in COVId-19 cases after a few years. But we can anticipate the natural distribution shifts only if we know that the ML system indeed uses such symptoms to classify (which we might not, given p3), or if we have a way to connect the components of the model learnt by the algorithm to what we (theoretically) know about the disease (which we might not able to do, given p1). Similarly, in comparison to traditional image classification methods, ML image classification algorithms are more easily affected by *adversarial attacks*, i.e., deliberate tampering with the algorithm. In particular, due to the opacity (p3) of the algorithm, such attacks cannot be detected by tracking the internal workings of the process. Those examples seem to indicate that p1-3, because they strictly connected to typical errors in ML (shortcuts; natural distribution shifts; adversarial attacks) create, indeed, serious epistemic challenges.

The analogy with clinical translation makes available the prediction that there should be a process that could be used to establish reliability nevertheless, namely RCTs (p4). To establish this, the analogy between clinical translation and ML needs to be treated as a predictive one (section 2.1), i.e., one needs to assume (i) that the fact that RCTs are a suitable method of establishing the efficacy of drug treatments is predicated on the three relevant epistemic properties p1-3 and (ii) that both analogues possess those three properties. We will identify several problems with the drawing of this analogy below (section 3.2). However, it is undoubtedly the case that the design of state-of-the-art efficacy tests for ML medical image classification has been influenced by the argument from predictive analogy outlined here and that the mechanisms used to establish efficacy are modeled on RTCs. An example of such a close boot-strapping to pharmaceutical RCTs is the state-of-the art study on the clinical safety of AI-supported screen reading by Lang et al. (2023). In this study, a group of 80033 women participating in mammography screening were randomly assigned either two human screen readers or a human reader and an AI classification system. Initial results after a two-year period showed that cancer detection rates in the two groups were similar, but that the medical professionals' workloads in the AI-supported group were significantly reduced (Lang et al., 2023, p. 936). The trial will continue for two years so that the long-term effects of AI-supported screen reading on clinical outcomes will also be studied. The predictive analogy between pharmaceutical discovery and ML has therefore clearly been an influential one.

3.2 Criticism of the Predictive Analogy Between Clinical Translation and Machine Learning

There are two aspects in which the analogy outlined above (section 3.1) does not straightforwardly fit the structure of a Hessian predictive analogy. Firstly, the analogy is not between two objects (e.g., Earth and Mars) but between two complex, epistemic processes, i.e., clinical translation and ML[4]. Similarly, the predicted property p4 is not a first order property of either analogue, but a process for testing the reliability of the two analogue processes (RCT). However, this leads to the second significant difference to a 'straightforward' predictive Hessian analogy: even for the first analogue A1, it is not clear that the properties (p1, p2, p3) deductively entail the property p4. Instead, it seems to be more correct to say that empirical practice has proven that despite the properties (p1, p2, p3), RCTs are an effective methodology for establishing the benefits of newly developed pharmaceuticals. Given that ML systems and their construction as a process have the same properties (p1, p2, p3), one can assume (but not deduce) that there is a high likelihood that a

[4] Admittedly, London (2016) himself remains somewhat ambiguous about this. On the one hand, he (2019) introduces the epistemic properties (p1, p2, p3) of medical knowledge by discussing drugs and pharmaceuticals as molecules; on the other hand, in other passages, London (2019) suggests that the analogy should be between the methods used to establish the reliability of the process of clinical translation and the methods used to validate or construct ML systems. Our unpacking of the analogy in section 3.1 indicates that the latter is the intended meaning.

similar process to RCTs might also be effective in establishing the effectiveness of ML. However, this implies that the analogy is actually not a predictive one[5]. Instead, the reformulation highlights the similarity between this analogy and the analogy between DNA sequencing and coding problems, or the analogy between machines and organisms (section 2.2)

We maintain that the best function to ascribe to the analogy between clinical translation and the construction of ML is a *generative* one. This seems to remain in line with London's (2019) interpretation of the relationship between those two analogues: he suggests that we can learn a lot in the context of ML from the process of clinical translation, given that the process of clinical translation faces similar epistemic challenges as the construction and evaluation of ML systems. However, this does not imply that the exact same process to establish efficacy should be used. The analogy is generative because we can draw inspiration from the process of clinical translation to overcome the challenges posed by properties p1-3. This interpretation of the analogy implies that a viable research strategy can be derived from considering similarities to drug development, but that the process of establishing reliability cannot be translated one-to-one onto the case of ML.

3.3 The Nature of Clinical Translation

Now that we have established that the analogy is a generative one, we have to show which aspects of A1 are indeed 'generative'. By drawing from London's epistemology of clinical translation, we identify those aspects that can be 'generative' in the context of ML.

3.3.1 The process of clinical translation

Let us start by specifying the process of clinical translation, and in which sense it is still reliable despite the challenges raised by p1, p2, and p3.

The process of clinical translation is typically associated with clinical trials, most notably RCTs. London, in his work with Kimmelman (Kimmelman and London 2015; Kimmelman 2012), has comprehensively constructed a reliabilist epistemology for clinical translation. Their starting point is that the output of clinical translation is not mere 'hardware' such as pharmaceuticals,

[5] We focus solely on arguing against the predictive analogy interpretation as it is the mainstream interpretation of the AI and clinical translation relationship. Given the technical sophistication of the two areas, a persuasive analogy has not been entertained and does not seem applicable here. We therefore do not see any need to address this analogy specifically, in particular, given that we will provide a positive argument for a generative analogy interpretation below.

devices, etc, nor is it just establishing a causal connection between an intervention and an endpoint. What is at stake in clinical translation is *information*. This is based on the assumption that "drugs alone are not therapeutic agents" (2015, p 29). In fact, drugs are chemical or biological substances which can have a therapeutic effect only in the presence of specific conditions and constraints. The process of clinical translation then starts with a level of uncertainty where there is a hypothesis that a molecule might have a beneficial effect in treating a certain condition. But the hypothesis is, typically, based on either tenuous evidence or vague background knowledge: it may be that a molecule has a structure such that it is conceivable that its mechanism of action might be co-opted for therapeutic purposes; or it has been observed in other contexts that the molecule had provided benefits to treat a given condition; etc. In all these cases, the starting point of clinical translation is characterized by p1, p2, and p3: there is uncertainty as to what the molecule does when administered in different physiological systems (opacity and atheoreticity); the association between the molecule and beneficial effects is tenuous or merely hypothetical (associationism). The process of clinical translation can then be construed as a mechanism to diminish or sidestep the uncertainties[6] related to p1, p2, and p3.

But what kind of mechanism is it? According to Kimmelman and London, it is a mechanism for collecting pieces of information on the *molecule-in-context* that will unlock the therapeutic potential of the molecule itself. They call the components of such a mechanism *intervention ensemble* (IE). IEs have a complex structure. They are divided into three main dimensions (Kimmelman 2012). Within the *treatment dimension*, components are typically information on dosage, schedule administration target, co-interventions, timing, risk mitigation, and others. The *population dimension* includes components such as information on diagnostic criteria, indications, contraindications, age. Finally, the *outcome dimension* includes components such as information on endpoints, duration, and the like.

In clinical translation, the values and boundaries of the components of these dimensions of IEs are changed and tested until two goals are achieved. First, the process needs to identify optimal values of variables within an IE (e.g. dose, timing of drug administration, etc) at which "a drug achieves the most favorable risk-benefit balance" (p 29). Second, the process has to identify more

[6] For an overview of quantitative and qualitative uncertainties in clinical research, see (Djulbegovic 2007)

precisely the boundaries of the dimensions beyond which IE is not clinically useful anymore - thereby guiding concerns of external validity.

Another feature that London and Kimmelman emphasize is that this process of finding the right materials, identifying how to coordinate them, and determining optimal values for the dimensions and boundaries of the IEs is an exploratory, experimental, and gradual process of converging to the optimal IE among an infinite number of possible ensembles. This is an important feature, which differs from the typical assumption of clinical translation as a linear process.

To sum up, clinical translation can be construed as a process that, in an exploratory and piecemeal fashion, provides information on the conditions under which a certain molecule will unlock its therapeutic potential. These different pieces of information constitute an IE. The more one goes far away from these ideal conditions, the less reliable the molecule's therapeutic potential will be. In the context of prescribing a drug, it is then a judgement-call by physicians to evaluate whether the context of a patient meets the characteristics of the relevant IE so that a certain drug will be beneficial to the specific patient.

3.3.2 Clinical translation as a reliable process

IEs can confer two types of reliability to the processes and products of clinical translation. Before describing them, let us qualify more precisely the 'reliability' part. We understand reliability in a slightly different way in which this term has been taken in epistemology. A classic form of reliabilism is about knowledge, and it states the conditions under which an individual knows something. It is said that an individual knows that *p* iff *p* is true, they believe *p*, and *p* is formed through a reliable process, where a reliable process is one producing a high proportion of true beliefs over incorrect ones (Alvarado and Humphreys 2017; Goldman 1979). Here we take 'reliable' in a broader sense, because it is not about beliefs and sentences, but it is about the outputs of the process of clinical translation and the use we make of them, and the process of clinical translation is 'reliable' when it tends to produce a high proportion of correct outputs. There are, in particular, two types of outputs. The first output are pharmaceuticals that do something specific. In this first sense, a well-constructed IE confers reliability for producing a pharmaceutical doing something specific (e.g., treating a condition under specific circumstances). Second, clinical translation produces information that is used in the clinical context. In this second sense, IE provides information to reliably use the products of the process of clinical translation in new clinical contexts. The work here is to identify the conditions that make such a process reliable in these two senses.

We call the first sense of reliability *ensemble-reliability* (that is, the internal reliability of the ensemble). This means that the set of practices to build a IE, which consists in establishing the nature of the three dimensions and the boundaries of each component, leads *per se* to a reliable output, in the sense of a drug delivering a beneficial effect. The 'reliability-conferring' properties of a well-constructed IE are information on how the different dimensions of IE either diminish or sidestep the uncertainties associated to p1, p2, and p3. In other words, an IE is reliable by decreasing the likelihood of errors induced by those challenges and uncertainties. By doing this, for a given IE, there will be a tendency to produce more favourable outcomes than not, where 'tendency' is specified individually for each IE by adopting relevant quantitative metrics. Therefore, a proper IE will give us reasons to believe that the IE is reliable.

In particular, by strengthening the associations between an intervention and a clinical end-point (e.g., by finding the right co-interventions and optimal values, and addressing issues of confounding factors), IE will diminish the uncertainties introduced by associationism (p1). Understanding the specific clinical context and finding optimal values/boundaries within which the molecule will be effective will sidestep the need for any information about the mechanism of action to warrant the use of a given drug (p3), and will avoid the pitfalls of the incompleteness and uncertainties of speculations about mechanisms of action[7] (Howick 2011). Finally, while pharmaceutical theory might be useful at various junctures of clinical translation (Aronson et al 2018; Kimmelman and London 2015), the justification of the results of a RCT are not dependent on theory or domain knowledge, i.e., reliability can be established despite atheoreticity (p2). We can express these ideas by saying that the process of clinical translation is not really a translational process in the colloquial sense; rather, it is a process of progressively enveloping a molecule within a given (metaphorical) environment that is the IE, where IE constraints the action of a molecule in such a way that the molecule's therapeutic potential is realized more often than not, and where most confounding factors are eliminated such that the statistical association between the molecule

[7] In our understanding, this view is typical of movements like Evidence-Based Medicine (EBM), which downplays the importance of mechanistic evidence. While EBM is certainly a major player in shaping the epistemology of medicine, there are also other scholars who stress the importance of mechanisms, thereby arguing in favor of a more inclusive evidential pluralism (see for instance, Parkinnen et al 2018)

and the effect is robust. All in all, building an IE by following the proper vetted standards will provide good reasons to rely on the outputs of the process of clinical translation.

We will call the second sense of reliability conferred by IEs *use-reliability*. Having built a robust IE is not enough to guarantee that the product it envelopes will be reliable in a new context. For instance, just knowing that a drug x is efficient in treating the condition y is not enough to safely prescribe x to treat all instances of y. However, the information gathered by constructing x's IE can be used to assess whether the conditions under which x is shown to treat y are similar to the conditions of the context in which one wants to prescribe the drug. Recent calls to report protocols of trials such as SPIRIT or CONSORT (Chan et al 2013; Schulz et al 2010), which are typically taken to put one in a position of evaluating ensemble-reliability, are meant to also improve 'use-reliability'. In other words, information contained in a well-constructed IE will give use reasons to believe that the pharmaceutical will work properly in a context that is similar to the original context of the IE.

4 USING THE GENERATIVE ANALOGY BETWEEN CLINICAL TRANSLATION AND MACHINE LEARNING

Having introduced IEs (section 3), we can use the analogy between ML and clinical translation to generate similar, but not identical, epistemic warrants (understood broadly) to address the uncertainties associated with p1-p3 for ML. Our proposal is that we should take inspiration from IEs to generate an equivalent ensemble-template for the ML context, which will envelope MLs in such a way that gets around or diminishes p1, p2, and p3. We call the ML equivalent to IEs *learning ensemble* (LE, see Figure 1). Ideally, a LE will confer both ensemble-reliability as well as use-reliability (section 3.3).

We generally define a learning ensemble LE as a set of components c1, c2, …, cn that (within the boundaries b of specific conditions sc1, sc2,...,scn) will lead a ML system S to reliably function in the intended way (*ensemble-reliability*). If those conditions apply in a new context, and they are within the defined boundaries, then we can safely infer that S will be reliable in the new context as well (*use-reliability*).

4.1. Components, dimensions, and reliability of learning ensembles

Using the generative analogy (section 2) leads us to believe that there may be similar dimensions to LEs as there are to IEs. Furthermore, as in the case of IEs, one can distinguish between components within the dimensions. These components and dimensions constitute the information to meaningfully evaluate the system's reliability and to determine whether a certain ML system will reliably work in a new context.

In IEs, the dimensions comprise information on the treatment, the population, and the outcomes. In order to identify the dimensions of LEs (and their components), we take inspiration from recent initiatives in medical AI that are aimed at improving how practitioners report the way ML systems in medicine are constructed: SPIRIT-AI (Cruz-Rivera et al 2020), CONSORT-AI (Liu et al 2020), and TRIPOD+AI (Collins et al 2024). These reporting initiatives are 'AI extensions' (as the titles of the SPIRIT-AI and CONSORT-AI articles suggest) of reporting guidelines for RCTs and computational models in medicine. These reporting guidelines provide the necessary information (i) to evaluate whether ML systems have been designed according to vetted standards, which are viewed as conducive to a reliable performance, and (ii) to implement ML systems in new contexts. Therefore, we can ascribe to these initiatives roughly aims of improving both ensemble- and use-reliability. Drawing on these initiatives, we identify three dimensions of LEs. These are:

1. the '*boundaries of reliability*' *dimension* (D1), i.e., the specific circumstances in which a ML system has been built, the methodological steps followed, and the justification for these steps
2. the *performance dimension* (D2), i.e., the various metrics used to evaluate the performance of the ML system, and the justification for using such metrics
3. the *functional dimension* (D3), i.e. the intended uses of the ML system, the evidence that the intended uses are achievable, and the relation of these uses to domain knowledge and domain norms

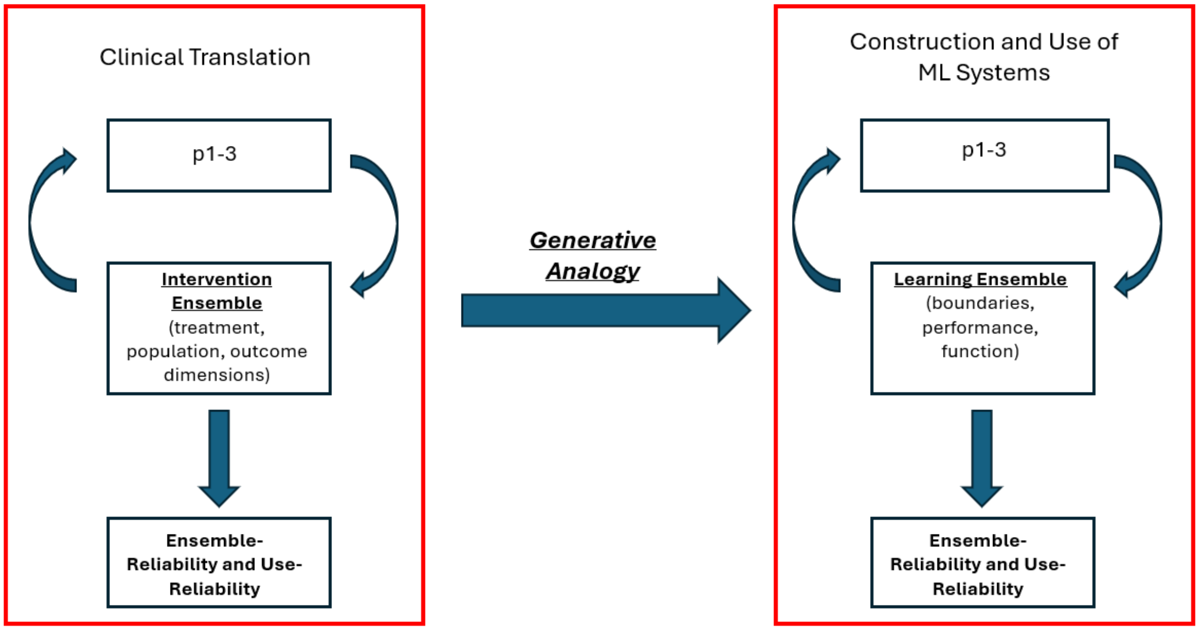
Clinical Translation
p1-3
Intervention Ensemble (treatment, population, outcome dimensions)
Ensemble-Reliability and Use-Reliability
Generative Analogy
Construction and Use of ML Systems
p1-3
Learning Ensemble (boundaries, performance, function)
Ensemble-Reliability and Use-Reliability

Figure 1. Generative analogy between clinical translation and machine learning

In the next sections, we will outline how different components of D1-D3 can potentially confer ensemble- and use-reliability to the ML system. We are not providing a comprehensive and exhausting list of components and their reliability-conferring properties. Rather, we will exemplify some properties of LEs for MLs.

4.1.1 D1: The Boundaries of Reliability

D1 contains information on what we call 'the boundaries of reliability'. A ML system is typically built to perform well in specific circumstances. We understand 'circumstances' (as it will be apparent below) in a broad sense. The circumstances are the components of D1. To achieve ensemble-reliability, values for circumstances that allow an optimal performance need to be specified. To achieve use-reliability, it has to be possible to compare the circumstances of implementation to the original circumstances in which the ML tool has been built. In the following, we will discuss some examples of components within D1.

First, the ML system will work reliably if used by a specific user group. This may be for different reasons e.g., the ML system is designed in such a way that its output or its functioning can be properly understood only by individuals with certain competencies. This is especially relevant for use-reliability. For instance, SPIRIT-AI (extension 11a iv) stresses that poor clarity on how to use the human-AI interface may lead to confusion about “whether an error occurred due to a human deviation from the instructed procedure, or if it was an error made by the AI system” (p 1359). In both cases, poor clarity on the intended user makes the use of the ML system unreliable. This is especially important to diminish the uncertainties related to opacity, not necessarily understood as an essential and ineliminable opacity, but more an opacity based on a lack of technical expertise related to the use of ML systems.

A second class of components should specify the infrastructure supporting the ML system. SPIRIT-AI and CONSORT-AI (briefly) mention aspects concerning the specifications of suitable hardware or software. These aspects are essential to the ensemble-reliability as it needs to be ensured that the ML system has been developed and tested on reputable and well-functioning computational infrastructures, platforms, and reputable software. If an ML system has been constructed on the basis of dubious computational infrastructures, then this may raise concerns about its reliability. Reputable and well-functioning infrastructure will give us reasons to believe that the ML system is properly executing its computational and coding components. Similarly, this component is important to ensure use-reliability: if a ML system has been designed to perform well on specific hardware or software, then in a novel context of implementation comparable hardware and compatible software must be used. In this case, uncertainties diminished are not necessarily related to p1-3 or the typical errors in ML, but they are about the errors that malfunctioning software and hardware might lead to.

One of the most crucial components of this dimension is input data. Both TRIPOD+AI (Collins et al 2024) as well as SPIRIT-AI (Cruz Rivera et al 2020) emphasise the importance of reporting crucial information about the data used to train the ML system, e.g., sources of such data, rationale for using this data, and information on its representativeness. To evaluate ensemble-reliability, the reliability conferring properties are related to the way data is processed and its provenance. Firstly, it is important to justify why a certain data set has been used for achieving a specific goal, and whether the representativeness of data is sufficient for the task at hand, as described in TRIPOD+AI (items 5). Moreover, it is important to specify in the protocol the minimum quality-requirements for the input data, e.g., as stressed by SPIRIT-AI (extension 10 ii). Given this information, it is possible to evaluate whether the input data is of sufficient quality to avoid the well-known problem of ‘garbage-in/garbage-out’. High-quality data can diminish issues of shortcuts, because such data can be curated to eliminate artifacts which might lead to shortcuts in the first place. High-quality data can also be used to diminish the nefarious consequences of adversarial attacks during adversarial training (Dong et al 2022). All this information gives us reasons to believe that p1-3 can be sidestepped, given the connections between these and typical ML errors – because we know that the probability of these errors is diminished, it gives us a confidence boost in believing that the system generates correct outputs. Information about input data is also crucial to gauge use-reliability. At a basic level, information about data will suggest issues of generalization via possible mismatches between the training data distribution and the data distribution of the context of implementation (Freiesleben and Grote 2023), which can result in unreliable performance of the ML system in the novel context. Other crucial information include the extent to which given input data is a proxy-measure or the provenance of input data, such as it being unprocessed or vendor-specific post-processing data (SPIRIT-AI, extension 10 ii). This type of information is useful to establish whether proxies are acceptable to the new context of implementation, or whether the processing procedures of a specific vendor are not compatible with the data used in the deployment context. Furthermore, knowing that a certain ML system is continuously retrained with data of a specific provenance can be important to anticipate possible natural distribution shifts or performativity in the novel context of implementation (FDA 2019; TRIPOD+AI item 12f; Freiesleben and Grote 2023; SPIRIT-AI, extension 11a iii). This information is then relevant to overcome instances of p1 and p3 (given their connections to natural distribution shifts).

4.1.2 D2: Performance Criteria

D1's 'components' provides information on the conditions under which a ML system systematically and consistently has been shown to work well. However, D1 does not contain information about actual measures of the performance of the ML system, i.e. a measure of the system's record of providing the right output. In process reliabilist epistemologies, it is typically said that a process is reliable "not so because it was successful once, but rather because there is a tendency to produce a high proportion of true beliefs relative to false ones" (Duran and Formanek 2018, p 653). *Mutatis mutandis*, what is at stake here is how to interpret the intuition that a ML system should produce a higher proportion of correct outputs over incorrect ones. What it means for a ML system to perform (within the conditions set up in the first dimension) reliably is specified by D2, which we will call the *performance* dimension.

There exists a large body of literature on measuring performances of ML systems. For instance, D2 overlaps significantly with the reliability indicators of technical performance of algorithms in computational reliabilism (Duran 2024). Duran explicitly mentions validation methods, showing how a judicious choice of specific validation methods and their results provide "a good indication of the algorithm's accuracy and margin of error" (p 10). Such quantitative measures of performance are important to evaluate ensemble-reliability. Another important aspect is that ML systems should also report the level of uncertainty of a certain output (Grote 2021). This can diminish or sidestep the challenges related to p2, because it suggests how confident is a ML system of the strength of the association between an input and an output. Additionally, integrating the output with the use of Explainable AI (XAI) tools can, in part, diminish uncertainties. For instance, XAI tools, while might not provide positive evidence for something to work (Duran 2021), can nonetheless flag cases of shortcuts (and hence sidestep p1 and p3), by showing that a ML system has indeed leveraged, e.g., pixels of an image that are unrelated to the particular clinical condition. Further information relevant to determine ensemble-reliability are justifications of how and why a given performance metric has been chosen. This is because any metric is subject to various tradeoffs between false-positives and false-negatives, e.g., the precision-recall tradeoff and the sensitivity-specificity tradeoff. Ratti and Graves (2022) notice that, depending on the goals of the ML system, using one tradeoff rather than another might make a difference as to how reliable the ML system will be, e.g. if the purpose of the system "involves diagnosing a disease, then sensitivity-specificity may be a better tradeoff, while if the system is retrieving disease information from patient records, then precision-recall might be better" (p 811).

For use-reliability, information on which specific trade-offs exist is important: if a user thinks that in the new context of implementation false positives should be avoided more than false negatives, then knowing how the tradeoff between false positive and false negative has been treated in the ML will provide crucial information to decide whether to use the ML system in the novel context.

4.1.3 D3: Function of ML systems

Philosophers of technology have drawn a distinction between *effect function* and *purpose function* (Vermaas 2009; van Eck 2015). On the one hand, effect functions designate the desired effect of the behaviour of a technical artifact. In the case of an electric screwdriver, the effect function will be to tighten or loose screws. On the other hand, purpose functions designate those state of affairs in the real-world that effect functions are taken to contribute to. In the case of the electric screwdriver, the purpose function might be to hang a painting. There can be misalignments between effect functions and intended purpose functions, such that effect functions are not conducive of purpose functions. Drawing on this distinction, we think that additional information has to be provided to make sure that the ML system not only performs well (D2) within a specific set of conditions (D1), but also to make sure that its purpose function is well-specified, that it is aligned with the context of use, and that there is evidence that the purpose function is indeed facilitated by the system. We call this third dimension (D3) the *functional* dimension. This dimension is not necessarily related to the uncertainties raised by p1-3, but it is still essential to better define and characterizes LEs.

The main components of D3 are the relevant effect and purpose functions. ML systems' effect functions are typically predictions or classifications. In contrast, the purpose functions to which ML systems should contribute may vary. In the case of medical AI, the FDA has recently (2019) stressed the importance of specifying how the ML system is intended to contribute to healthcare practices, e.g. whether it should be used to diagnose or to drive (rather than to inform) clinical management. For instance, if the purpose function is to provide a diagnosis in such a way that the ML system trumps all further reasoning, then we might reasonably require performance metrics to be much more stringent, given the potential consequences. In other words, specifying the purpose function can provide further information to contextualise performance and, ultimately, to evaluate ensemble-reliability. Specifying the purpose function is important also for use-reliability. For example, if a ML system has been developed to inform clinical management then there is no reason to assume that it can be used reliably to drive clinical management. The importance of the functional dimension is underscored in the reporting initiatives mentioned above. TRIPOD+AI (items 8a and 8b) alludes to the importance of providing clear outcomes and instructions on how to interpret them. SPIRIT-AI (extension 11a) is even more explicit in requesting clarity on how the ML system will contribute to specific decision-making procedures or clinical practices, and on how to interpret its outputs.

In addition, the purpose function will specify how ML designers have envisioned the context in which the system is intended to be used. In the sciences, this is a serious issue. Termine, Ratti, and Facchini (2024) have shown how ML practitioners have the tendency of constructing ML systems without paying much attention to the scientific context in which these are intended to be used. If the purpose function of a ML system developed in a given scientific context is completely at odds with the norms and domain-knowledge of that context, then its reliability is jeopardised. This is because, while the ML system might provide more correct outputs than not (D2), the outputs themselves might be misleading or based on poorly supported assumptions. Therefore, a purpose function that is well aligned to a given scientific context, will give use reasons to believe that outputs will not be inconsistent with domain knowledge and norms. Duran (2024) calls this reliability indicator 'knowledge-based integration', and he shows the perils of these misalignments with an example taken from facial recognition for crime detection, where "the categories their CNN purports to use (...) are posited in isolation from established evidence, models of criminal psychology, social studies of crime, and the relevant theories of criminality" (p 15). By identifying the extent to which the purpose function of the ML system is aligned with domain-knowledge and domain norms, one can evaluate both ensemble-reliability and use-reliability.

Please note that all of this has nothing to do with the *atheoreticity* of the way the algorithm learns or the model (that is, p2); it is about the relations between the purpose of a ML system (which can and should be made entirely transparent) and the domain knowledge and domain norms of the context of implementation. Whether there is p2 is irrelevant to these considerations, as they are irrelevant concerns about opacity (that is, p3), which, should they be overcome, do not add anything to how the purpose function is evaluated with respect to the first two dimensions.

5 CONCLUSION

In this article, we have provided a philosophical interpretation of the parallels between AI and medicine, and have used these parallels for developing a reliabilist framework for ML. We have interpreted the parallels as analogies (section 2). In particular, we have shown that the process of clinical translation can be understood as a process of building an IE, which can then be used to assess ensemble- and use-reliability (section 3). Interpreting the analogy with AI as generative, we use it to suggest the construction of an equivalent ensemble (i.e., a LE) to assess ensemble- and use-reliability of ML systems (section 4). Thereby, we have distinguished three dimensions of

LEs: boundaries of reliability (D1); performance (D2); and functionality (D3). Our account of ML reliability is compatible with other accounts present in the literature, such as computational reliabilism (Duran and Formanek 2018; Duran 2024). In this paper, we have only discussed some examples of relevant components of LEs. As Grote et al (2024) notice, building a robust and reliable ML system "typically involves a trial-and-error process, requiring a combination of domain knowledge, external evaluation with out-of-distribution data, data augmentation, explainable AI techniques, and continuous retraining" (p 5). Much more needs to be said about the reliability-conferring properties (both for ensemble-reliability and use-reliability) of all components of the three dimensions of LEs that we have identified. This article, though, paves the way to do this, and provides a framework that can guide such a process.

REFERENCES

Alvarado, R., & Humphreys, P. (2017). Big data, thick mediation, and representational opacity. *New Literary History*, *48*(4), 729–749. https://doi.org/10.1353/nlh.2017.0037

Aronson, J. K., la Caze, A., Kelly, M. P., Parkkinen, V. P., & Williamson, J. (2018). The use of mechanistic evidence in drug approval. *Journal of Evaluation in Clinical Practice*, *24*(5), 1166–1176. https://doi.org/10.1111/jep.12960

Bartha, P. (2022). Analogy and analogical reasoning, *Stanford Encyclopedia of Philosophy*

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity - Decomposition and Localization as Strategies in Scientific Research*. The MIT Press.

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, *32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Chan, A.-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jeric, K., Hró bjartsson, A., Mann, H., Dickersin, K., Berlin, J. A., Doré, C. J., Parulekar, W. R., Summerskill, W. S., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D., & Moher, D. (2013). SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. In *Ann Intern Med* (Vol. 158). www.annals.org

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. https://doi.org/10.1136/bmj-2023-078378

Craver, C., & Darden, L. (2013). *In search of Mechanisms*. The University of Chicago Press.

Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Darzi, A., Holmes, C., Yau, C., Moher, D., Ashrafian, H., Deeks, J. J., Ferrante di Ruffano, L., Faes, L., Keane, P. A., Vollmer, S. J., Lee, A. Y., Jonas, A., Esteva, A., Beam, A. L., … Rowley, S. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Djulbegovic, B. (2007). Articulating and responding to uncertainties in clinical research. *Journal of Medicine and Philosophy*, *32*(2), 79–98. https://doi.org/10.1080/03605310701255719

Duran, J. (2024). Beyond transparency: computational reliabilism as an externalist epistemology of algorithms.

FDA. (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. *U.S Food & Drug Administration*, 1–20.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, *202*(4). https://doi.org/10.1007/s11229-023-04334-9

Genin, K., & Grote, T. (2021). Randomized Controlled Trials in Medical AI. *Philosophy of Medicine*, *2*(1). https://doi.org/10.5195/pom.2021.27

Goldman, A.I. (1979). What is Justified Belief?. In: Pappas, G.S. (eds) Justification and Knowledge. Philosophical Studies Series in Philosophy, vol 17. Springer, Dordrecht. https://doi.org/10.1007/978-94-009-9493-5_1

Hesse, M. (1966). *Models and Analogies in Science*. Notre Dame University Press.

Howick, J. (2011). *The Philosophy of Evidence-based Medicine*. John Wiley & Sons.

Humphreys, P. (2011). Computational science and its effects. In M. Carrier & A. Nordmann (Eds.), *Science in the Context of Application* (Boston Stu). Springer.

Kay, L. (2000). *Who wrote the book of life? A History of the Genetic Code*. Stanford University Press.

Kimmelman, J. (2012). A theoretical framework for early human studies: Uncertainty, intervention ensembles, and boundaries. In *Trials* (Vol. 13). https://doi.org/10.1186/1745-6215-13-173

Kimmelman, J., & London, A. J. (2015). The Structure of Clinical Translation: Efficiency, Information, and Ethics. *Hastings Center Report*, *45*(2), 27–39. https://doi.org/10.1002/hast.433

Lång, K., Josefsson, V., Larsson, A. M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936–944. https://doi.org/10.1016/S1470-2045(23)00298-X

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., Chan, A. W., Darzi, A., Holmes, C., Yau, C., Ashrafian, H., Deeks, J. J., Ferrante di Ruffano, L., Faes, L., Keane, P. A., Vollmer, S. J., Lee, A. Y., Jonas, A., Esteva, A., Beam, A. L., … Rowley, S. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21. https://doi.org/10.1002/hast.973

Nicholson, D. J. (2013). Organisms ≠ Machines. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *44*(4), 669–678. https://doi.org/10.1016/j.shpsc.2013.05.014

Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M. P., Norell, C., Russo, F., Shaw, B., & Williamson, J. (2018). *Evaluating Evidence of Mechanisms in Medicine Principles and Procedures*, Springer

Schulz K F, Altman D G, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials, *BMJ* 2010; 340 :c332 doi:10.1136/bmj.c332

Stegmann, U. E. (2016). "Genetic Coding" Reconsidered: An Analysis of Actual Usage. In *Source: The British Journal for the Philosophy of Science* (Vol. 67, Issue 3).