What Can Artificial Intelligence Do for Scientific Realism?

Petr Spelda[1] and Vit Stritecky
Faculty of Social Sciences, Charles University

(Preprint, forthcoming in *Axiomathes*[2])

ABSTRACT

The paper proposes a synthesis between human scientists and artificial representation learning models as a way of augmenting epistemic warrants of realist theories against various anti-realist attempts. Towards this end, the paper fleshes out unconceived alternatives not as a critique of scientific realism but rather a reinforcement, as it rejects the retrospective interpretations of scientific progress, which brought about the problem of alternatives in the first place. By utilising adversarial machine learning, the synthesis explores possibility spaces of available evidence for unconceived alternatives providing modal knowledge of what is possible therein. As a result, the epistemic warrant of synthesised realist theories should emerge bolstered as the underdetermination by available evidence gets reduced. While shifting the realist commitment away from theoretical artefacts towards modalities of the possibility spaces, the synthesis comes out as a kind of perspectival modelling.

[1] petr.spelda@fsv.cuni.cz

[2] https://doi.org/10.1007/s10516-020-09480-0

1. Introduction

Perhaps an equally relevant question could run as follows. What can artificial intelligence *do to* scientific realism? Admittedly, this latter one carries a certain adversarial sentiment, or even maybe an intent, to mount yet another attack showing what is wrong with scientific realism in either the epistemological, semantic, or metaphysical dimension (cf. Chakravartty, 2017a). On the contrary, this paper intends to offer an answer to a very tangible argument against scientific realism and turn it, with an assistance of artificial intelligence, into an argument supporting the epistemic warrant of scientific realism. The defence against the counterargument comprising the case of the present paper will not entail vague prospects considering artificial intelligence a universal solution to every conceivable problem. Therefore, the paper proposes an engagement at a level allowing for applications of specific artificial representation learning models towards building a more resilient foundation for the epistemic warrants of realist scientific theories. In this regard, the paper aims at the issue of underdetermination of scientific theories and the related matter of unconceived alternatives together constituting the analysed argument against scientific realism.

In a broader sense, the paper intends to contribute to the debate on computational methods acting as epistemic enhancers that extend the natural inferential abilities of human scientists (cf. Humphreys 2004; 2011; 2020). In his pioneering analysis, Humphreys (2004) showed a way in which we can reason about computational science, emphasising that the hybrid epistemic regime, combining human and machine elements, signals the rise of a new kind of epistemology. A major concern associated with computational methods lies in their epistemic opacity, manifested as representational opacity in the case of machine learning models, challenging their low-level understanding by humans (Humphreys 2020). With

growing generalisation capabilities of the state-of-the-art models, this is in no way surprising and does not prevent the hybridisation of science from going forward, as suggested by recent developments in several fields (cf. Radovic et al., 2018; Carleo et al., 2019). Yet the hybrid epistemic regime, synthesising the results of human and machine learning, affords new perspectives on long standing disputes in general philosophy of science, such as the debate between scientific realism and its challengers (e.g. Wray 2018). The paper uses this opportunity to put forward a different view on Stanford's unconceived alternatives, which shows a way of reversing their effects on realist theories by extending human inferential abilities with adversarial machine learning.

Towards this goal, the argument proceeds in the following manner. First, the paper specifies which kinds of underdetermination, and thus of unconceived alternatives, are amenable to the proposed synthesis between human scientists and artificial representation learning models. Second, instead of conceiving unconceived alternatives through retrospective interpretations of the scientific progress, the paper proposes a counterintuitive move, conceiving unconceived alternatives as the results of exploring possibility spaces of available evidence by utilising artificial representation learning. Third, the paper argues that adversarial machine learning produces samples from the left out regions of the possibility spaces thus yielding modal knowledge regarding what is possible therein. Finally, by consulting nascent applications in astrophysics, cosmology, and high energy physics, the paper relates the argument to a recently proposed program of perspectival modelling (Massimi, 2018).

2. Underdetermination of Scientific Theories and the Problem of Unconceived Alternatives

Then, in consequence of the offered introduction, the following narrowing of the subject matter is due. In regard to the phenomenon of underdetermination, we will mostly refrain from any comments concerning its holist version as developed by Quine and later subsumed under the umbrella of the Duhem-Quine thesis (Quine, 1951, pp. 39-43). Since our proposal consists of a practical amendment to the process of theory building, so as to bolster the resulting epistemic warrants, it would still be susceptible to holist underdetermination. The reason for this lies in the nature of the amendment. It is designed to face the underdetermination by evidence bringing about a set of contending theories equally supported by the available observations inputting a process of theory building. Then, in the case of holist underdetermination occurring within the context of the totality of our knowledge, this amendment remains indeed toothless. In selecting among possible revisions of an existing theory facing recalcitrant experience, it won't be able to help because holist underdetermination invites chain re-evaluations of the respective *total* system of beliefs (Quine, 1951, pp. 39-40). It thus goes beyond of what Quine later (1970; 1975) considered to be the case of empirical underdetermination by all possible evidence yielding infinitely many observation conditionals because his argument for the holist case, however naturally related, implies also a prospect of revising the rules comprising the scientific method itself. As Stanford (2017) defends this position against attempts to constrain its pertinence (cf. Laudan, 1990), he argues that the possibility of revision applies not only to ampliative rules of the scientific method but to deductive principles as well. Although this radical version of

holist underdetermination remains contentious and outside of the amendment's scope, we will use its properties to further delineate the make-up of our proposal.

The core of the proposal consists in revising the ampliative rules of the scientific method in order to address the case of underdetermination by available evidence. The amendment is not intended to resolve the case of empirical equivalents; that is distinct theories with identical empirical consequences (either arising by the natural progress of the scientific endeavour, as in van Fraassen, 1980, or constructed artificially with the intention to show invariability in finding the equivalents as in Kukla, 1996). Our proposal aims at an arguably more severe case. It comprises the theories which, while integrating all the available evidence, predict different yet to be observed phenomena. We consider this latter case more serious because it affects not only scientific realism but also the supposed solution of the equivalents stalemate in terms of empirical adequacy proposed by constructive empiricism (van Fraassen, 1980, pp. 11-12). From this perspective, a theory remains safe as long as we know the alternative accounts beforehand or know the future theories would be exclusively empirical equivalents. Only then one can, based on some voluntary epistemic attitude, attribute empirical adequacy accordingly. This clarity begins to deteriorate if we concede that there is a pool of yet to be conceived theories whose nature as well as volume is presently unknown, however, congruent with the available evidence (cf. Stanford, 2006). As Stanford suggests, the progression of science should be thus considered a history of displacement of the status quo theories by unconceived alternatives equally well confirmed by the then available evidence (ibid.). By offering a historical account of the displacement, the argument is construed as recurrent, supposedly affecting the present as well the future of the scientific endeavour (Stanford, 2006, pp. 17-18). The epistemic warrant of a theory,

regardless of whether aiming at truth or empirical adequacy, then shifts to a mere instrumentality of achieving some practical predictive goals (cf. Stanford, 2006, pp. 24-25).

Not unlike the previous attempts, this anti-realist position gains its viability by interpreting the past of the scientific endeavour, which permits it to make assertions about the future. It thus lends itself to direct counterarguments building on the deficiencies of the selected interpretation (for Stanford's New Induction f.e. Saatsi, 2015; Mizrahi, 2016; Mizrahi, 2017). Various claims about selectivity and/or overreaching of the interpretation then lead to a conclusion that the unconceived alternatives do not affect the scientific endeavour indiscriminately (f.e. Magnus, 2010 for an argument about the limited impact of unconceived alternatives due to the equally limited use of the affected eliminative inferences in theory building). Realists responded by reconsidering the epistemic attitude towards unobservables that would enable a more discerning selectivity or minimalism in their commitments (Chakravartty, 2017b; Saatsi, 2015 respectively). Such a shift in the program's core should better delineate the epistemic warrant of realist theories making the program less of a catch-all strawman for the rivals to wield freely. Stratifying the commitment to unobservables according to some rule (f.e. principled continuity in terms of causally efficacious properties as in Chakravartty, 2008 or Egg, 2014) should then safeguard realism against the unconceived alternatives. As some minimal core of a theory, to which realists ascribe the highest degree of belief, latches onto the reality firmly enough, even the unconceived alternatives falsifying most of the landscape would leave the core and scientific realism itself unshaken (cf. Chakravartty, 2008; Saatsi, 2015).

3. Reversing the Dynamics: Probing the Unconceived

In sum, with respect to unconceived alternatives a realist can assume that "*What you don't know can't hurt you; what matters is how we assess what we think we know now*" (Chakravartty, 2008, pp. 157-158). The suggested solution towards the content of even our best theories consists in a selective commitment superseding the indefensible global position (Chakravartty, 2017b, p. 3391). Facing the challenge of unconceived alternatives, the adoption of this strategy, however prudent, represents a retreat. By this we don't mean one clearing out the field for anti-realist claiming because the selectivity in commitment opens a way for peaceful cohabitations. The retreat lies in the assessment of what realists assume they don't know, or more precisely of what they are unsure of to the point of withdrawing the commitment to it. Swapping the global position for many selective ones, while dismissing unconceived alternatives, narrows the resulting theories' epistemic warrants. As the present and future theories should share an approximately true core, scientific realism prescribes parsimony in theoretical commitments, which translates into narrowing of the epistemic warrants. Such a discerning establishes the continuous predictive success of science and the 'No Miracles Argument' as its explanation in terms of scientific realism itself (cf. Putnam, 1975, p. 73). However, if a theory should latch onto the reality firmly enough, maintaining a steady predictive success, it has to be unique in the sense that the appearance of unconceived alternatives remains improbable, or even better impossible. Leaving the global position for selective commitments represents one way of delivering such uniqueness. Adopting this strategy introduces a subtle tension to the theory building process. Discerning a genuine parsimony in commitment is necessary for securing the theory's steady predictive success. Following this strategy further, an additional

selectivity in commitment in order to avoid unconceived alternatives could place the theory uncomfortably close to the edge of *artificial parsimony*. As long as uniqueness and the steady predictive success remain tied together without assessing the underdetermination by unconceived alternatives, the tension could push the theories towards artificial parsimony skewing the resulting epistemic warrants. Without a way to assess theories' exposure to the underdetermination by unconceived alternatives, the narrowing of the commitments becomes a rational strategy of theory building. The risk of artificial parsimony could then enter the picture permanently. One way such an outcome might actualise is if the 'No Miracles Argument' in terms of scientific realism remains the supposed cause of the predictive success of science without considering a framework assessing its exposure to unconceived alternatives (cf. Dawid, 2013, pp. 172-173).

Pointing out the risk of artificial parsimony is only a part of the picture, as a strong counterargument will always entail the fall-back to a sort of causal realism (cf. Psillos, 1999, Chapter 12; Egg, 2014), justifying the selectivity possibly approaching artificial parsimony. There is, however, also the second part because, as Stanford observes, the failure to conceive alternatives concerns the theory building processes as enacted by human scientists (2006). Admitting this observation suggests that any case of underdetermination by unconceived alternatives emerges within some framework of ampliative rules comprising a particular model of the scientific method (cf. Dawid, 2013, p. 60). It might be further argued that any set of ampliative rules represents a systemic elaboration of human cognitive defaults. Considering this situation from the anthropocentric perspective, there are at any given time alternatives which cannot be conceived due to the underdetermination of scientific method by human cognitive defaults. If it was the case that we had no way of weakening the influence of our cognitive defaults on the 'ladder' of underdetermination

(from ampliative rules to unconceived alternatives), then selectivity in theoretical commitments would prevail as the only meaningful strategy. However, assuming the feasibility of a generative process which, while learning the available evidence, could *produce phenomena unconceived within anthropocentric frameworks,* an opportunity would emerge permitting to probe the possibility spaces. Stemming from other than human cognitive defaults, such a process would have a potential to reverse the impact of unconceived alternatives on the realist theory building enterprise. Instead of being pushed into selectivity by the past flaws of exploring the possibility spaces, an expansion or a further selectivity in commitment would become a deliberate choice informed by the ability to search for the unconceived phenomena at the theory forming stage. Unconceived alternatives, in the anthropocentric sense inaccessible due to human cognitive defaults, would be made scrutable depending on the type of applied alternatives generating processes. By consulting the produced alternative phenomena with a human conception of the corresponding possibility space, epistemic warrants would emerge consolidated, achieving a lesser degree of underdetermination by the available evidence. The resulting theoretical commitments, as well as the epistemic warrants, would reflect a *synthesis* of the generated alternative phenomena with human representations that emerges from an enhanced understanding of the possibility space from which originates the available evidence. Augmenting the set of ampliative rules in this way would then remake a portion of the total of unconceived alternatives into an opportunity, one reversing the retreat following selectivity into a possibility of theoretical commitments' expansion. They would latch onto the reality more firmly, since by accounting for an additional part of the possibility space, the decreases in underdetermination of the resulting theories would acquire a more distance from the edge of artificial parsimony.

4. Synthesising Between Human Scientists and Artificial Representation Learning Models

In a nutshell, this is what can artificial intelligence do for scientific realism. By introducing other than human cognitive defaults, and with them also different modalities of probing the possibility spaces, synthesising between human scientists and artificial representation learning models could decrease the level of underdetermination by available evidence. Producing by other means phenomena which we are unable to conceive when confronted with the available evidence affords a transformation of the alternatives beyond their anti-realist interpretation. If we were able to achieve the synthesis, it would overturn the reductive perception of unconceived and instead accomplish solidifying of the epistemic warrants of realist theories against the very threat of unconceived alternatives. The feasibility of enhancing ampliative rules by such a synthesis co-founding generative process depends on whether there are artificial representation learning models with the desired properties.

The key desired property is a control over empirical underdetermination of the model contributing to the synthesis. As a result of aiming at unconceived alternatives, the synthesis will remain underdetermined by the following choices. First, it will be susceptible to what Quine considered indeterminacy of translation (1970), as only one part of the dyad, in the present case a human scientist, weaves both veins of the conceived phenomena into a single theory. Having a precise control over empirical underdetermination of the artificial model, in Quine's terms having the ability to fix its (foreigner's) observational sentences (cf. 1970, pp. 179-180), doesn't rule out that there are several ways of combining both veins of knowledge into a coherent theory. The second choice causing underdetermination of the

synthesis consists in the architecture and settings of the applied artificial representation learning model. A proper choice from the joint parameter space of models' architectures and their settings determines the key desired property, the control over models' empirical underdetermination.

As to the indeterminacy of translation, unless we abandon the notion of synthesis for autonomous computational discovery of knowledge producing full-fledged scientific theories, it will persist to cause underdetermination of the theory building. (Semi-) autonomous computational discovery of scientific knowledge was historically considered the avenue which would lead to artificial intelligence revolutionising the scientific endeavour and its philosophy alike (f.e. Gillies, 1996; Thagard, 1988). At first, the then computational state of the art merely sufficed to experimental rediscoveries of historical results, later leading to what can be considered novel, i.e. publishable, discoveries (Langley, 2000). However, as far as concerning the probing of possibility spaces for unconceived phenomena by artificial means, it is contentious whether the field of (semi-) autonomous computational discovery might contribute in any substantial way. Although professing a human-computer cooperation as well, which might be assumed for another kind of synthesis, its cornerstone consists of casting knowledge in terms of anthropocentric formalisms typical in individual disciplines (Džeroski et al., 2007). Combined with the initial emphasis on rediscovering the theories conceived by human scientists in the past, the methodological frameworks tend chiefly toward human cognitive defaults. Augmenting the theory building by computational processes necessitating the communication of knowledge in pre-established theoretical terms renders it impractical towards the issue of underdetermination by unconceived alternatives.

This deficiency was already identified by Alai (2004) while discussing the issues which disqualify earlier attempts, including the approach of Simon's group and work done in 'Turing' tradition, from making genuine discoveries. Concerned with the kind of scientific discoveries that encourage realism, Alai considers the (computational) processes of induction, as proposed by Holland et al. (1986), the only viable option due to their natural alignment with the model-based representation of reality benefiting realism (Alai, 2004). Although the present-day machine learning models still do not satisfy Alai's requirements for autonomous discovery, i.e. *unaided* goal discovery and model building in the human sense (cf. Alai, 2004, pp. 34-37), thanks to advances in Deep Learning we made significant progress in the end-to-end generalisation learning (cf. LeCun et al., 2015). The proposed synthesis between human and machine learning shows that extending human conceivability by samples from the left out regions of possibility spaces of available evidence essentially enriches realist theories. Finding new ways to generalise about evidence, while decreasing human involvement in the process (end-to-end machine learning), offers a reinforcement to realism even if we cannot rely on full-blown autonomous discovery machines. The hybrid epistemic regime of human-machine learning also agrees with the recently proposed functional novelty of predictions (Alai 2014), showing that despite the discovery machines as conceived by Holland et al. (1986) and Alai (2004) have not arrived yet, scientific realism can make use of the existing adversarial machine learning to pre-empt unconceived alternatives.

Returning back to the issue of pre-established anthropocentric formalisms, it is the supposed black-box character of artificial representational learning, i.e. of machine learning and *Deep Learning currently in particular*, which makes it less appealing to scientists preferring clear- or grey-box modelling of the traditional automated discovery systems (cf.

Stolle and Bradley, 2007). Although such a reservation has its merits in general settings, in the context of exploring possibility spaces represents a missed opportunity to reconsider the unconceived alternatives' reductive interpretation[3]. Admittedly, there are endemic concerns over the interpretation of the state-of-the-art Deep Learning models, as their complexity interferes with building a precise theoretical picture of their inner workings (cf. Zhang et al., 2017). Acknowledging this nature of contemporary artificial representation learning, however, doesn't impede its application towards probing the possibility spaces for unconceived phenomena. As the difficulties of interpretation make its integration into the anthropocentric frameworks of automated discovery harder, the viability of the proposed synthesis depends merely on an efficacious control over empirical underdetermination of the contributing model.

In sum, regarding theory building processes, the synthesis between human scientists and artificial representation learning models should deliver the following. Given a possibility space from which originates the available evidence, a generative process, while learning the probability distribution underlying this space, samples phenomena unconceived by human scientists dealing with the evidence. Accounting for these phenomena in the constructed theory should lessen underdetermination of the result by so far unconceived theories. As the possibility space of available evidence becomes better mapped, the synthesis raises confidence in the resulting theory, making it less prone to underdetermination by unconceived alternatives. In other words, the subsequent emergence of a different theory

[3] It's also perhaps a bit unfortunate that contemporary reviews of automated discovery adhere to somewhat dated typologies of artificial intelligence applications within science (f.e. Giza, 2017). Reflecting for the most part earlier results entirely omits recent successes delivered by Deep Learning, while underappreciating the influence of artificial representation learning on science in general (cf. ibid.).

fitting the available evidence while predicting novel phenomena becomes less probable. A part of the synthesised theory's predictive success would then derive from the epistemic warrant whose realist nature gets bolstered by incorporating the phenomena sampled from regions of the possibility space left out by human scientists.

4.1 Acquiring Material for the Synthesis: Sampling from the Possibility Space of Available Evidence

Crucial for the success of the synthesis is that the generative process sampling from a possibility space can be tuned to explore its truly left out regions. It is not an entirely straightforward task, as the model (generative process) learning to generate phenomena from the possibility space easily slides to sampling from an incorrect probability distribution mistaking it for the one truly generating the observational evidence. Although the available evidence represents merely a finite sample from the true distribution, its approximation learnt by the model needs to evade empirical underdetermination as much as possible. Otherwise, generating from an inadequately fitted model, failing to approximate the evidence producing distribution, ceases to explore the possibility space in a useful way. Achieving a good approximation is difficult, since apart from synthetic evidential data, we don't know the true distribution, and it is the point of the theory to hypothesise about it so as to reliably account for yet to be observed phenomena. Maintaining an efficacious control over the model's empirical underdetermination is necessary to avoid sampling phenomena vindicating underdetermination of the synthesised theory. Essential to such an end is the ability to recognise an underdetermined generative model. As the artificial part of the synthesis aims at the left out regions, it is vital to identify which kind of the produced

phenomena doesn't originate from them, since their presence implies an underdetermined generative model.

A generative model fails at sampling from the left out regions of a possibility space if it merely recreates the phenomena comprising the available evidence. In such a case, this behaviour can be considered an extremum producing biased low variance samples converging in faithfulness on the original observational evidence. In some settings, such as learning the most salient features of the available evidence, it constitutes a sound strategy, as it reliably produces faithful recreations while lowering the complexity of the learnt representations (f.e. Kingma and Welling, 2014). However, as a method of acquiring material for the synthesis, which accounts for the phenomena from the left out regions, it comprises a self-defeating option. Without a way of assessing to what degree the samples' fidelity approaches a recreation of the evidence, the synthesis would most likely yield further instances of empirically equivalent theories. In this sense, the generative model needs to correctly step beyond the available evidence to capture modalities underlying the possibility space thus acquiring the ability to sample unconceived phenomena from its left out regions. Put differently, the model must attempt to learn an approximation of the true distribution generating both the observational evidence as well as any phenomena congruent with the modalities determining the respective possibility space. A vital component establishing fruitful syntheses, and also the most formidable puzzle, consists in pushing the generative model beyond the fidelity of recreations towards high diversity samples while learning from a merely finite set of observational evidence.

A possible path leading beyond synthesising empirical equivalents entails arranging the artificial representation learning model in an adversarial manner. Traditionally, an artificial

representation learning model consists of a parameter set obtained by minimising a cost function capturing some learning objective (cf. Goodfellow et al., 2016, pp. 149-150). Such a set-up doesn't offer a straightforward solution to the puzzle of how to push the generative model to sample phenomena from the left out regions of a possibility space. To this end, it has been recently suggested that a viable strategy of acquiring samples from an approximation of the true distribution incorporating the left out regions involves the notion of *adversarial learning*. In its simplest form, the process of generating unconceived phenomena, i.e. samples from the left out regions, stems from an adversarial interaction between two players in terms of a minimax game (Goodfellow et al., 2014). Considering such a setting, each player conceived as a representation learning model attempts to minimise its cost function entailing both parameter sets while having a direct control only over its own parameters (ibid.). Theoretically, finding a Nash equilibrium of this zero-sum game during training of the model induces minimisation of the divergence between a learnt distribution and the true data generating distribution underlying the possibility space (cf. Goodfellow et al., 2014; Fedus et al., 2018).

In practice, implementing such a design are two Deep Learning models, i.e. artificial neural networks, forming a generative adversarial model comprised of a generator and discriminator network assuming the role of competing adversaries (Goodfellow et al., 2014). The adversarial learning of the distribution approximating parameters proceeds by introducing the generator network to a vector of random noise which it transforms into a sample supposedly coming from the same distribution as the observational evidence (Goodfellow et al., 2014). The sample gets in turn scrutinised by the discriminator network estimating the probability that it originates from the generator rather than from the observational evidence (Goodfellow et al., 2014). As the game develops, the discriminator

improves its ability to distinguish between the artificially created and observed samples, while the generator produces increasingly convincing novel phenomena as its approximation of the distribution improves by further interacting with its foe, the discriminator. The game continues until the discriminator can no longer correctly decide the origin of incoming samples. Reaching this state, the generator thus *fools* the discriminator into believing that its samples come from the observational evidence rather than from its approximation of the true distribution underlying the corresponding possibility space. In such an adversarial scenario conceived as a minimax game, the generator therefore iteratively minimises the probability of the discriminator correctly classifying the incoming samples, thus supposedly converging towards the theoretical equilibrium[4].

4.2 Underfitting of Adversarial Learning Models: Yet Another Case of Underdetermination

If carried out correctly, adversarial learning might assist in accessing the left out regions of possibility spaces so far accessible merely through narrow vistas of the presently available evidence. Furthermore, if such a generative model develops at least an approximately correct account of the underlying distribution, it would open a way for synthesising theories better withstanding the anti-realist charges referring to unconceived alternatives. In this

[4] As Goodfellow notes, since the players are neural networks, and their parameters acquired by back-propagation of error, heuristically, to secure a non-vanishing gradient it is better to consider the generator as maximising the probability of the discriminator being mistaken (Goodfellow et al., 2014). This slightly changes the nature of the game, since it can no longer be described in terms of a single value function (ibid.). Although this represents a shift from describing the scenario in terms of a minimax game, it doesn't lessen the game's relevance towards the theoretical analysis of adversarial artificial representation learning.

respect, the yield is twofold. First, as the generator produces samples comprising the left out regions, it compensates for human cognitive defaults by introducing different modalities of probing the corresponding possibility space. Second, building on this newly gained exploratory capability, the synthesis also achieves a new level of understanding of the modalities determining *what is possible* within that particular space. As a result, acquiring at least an approximate awareness of the left out regions and the phenomena therein provides an opportunity to conceive theories less prone to the underdetermination caused by unconceived phenomena supported by the available evidence, however, disagreeing with the state-of-the-art theories. A part of the realist commitment could be thus invested into the way of getting a more comprehensive picture of the possibility spaces by sampling from artificially learnt approximations of the underlying probability distributions. In other words, in a bid to compensate for human cognitive defaults, a fragment of the commitment could be taken and deposited not within theoretical artefacts or acquired modalities of the possibility spaces but in a different approach of exploring what lies in the neighbourhood of available observations as delivered, for instance, by adversarial representation learning.

In theory, then, replacing a single cost function with a collection leads to novel insights regarding the learning objective, as such a generative model constitutes a system of adversaries bound to compete while being exposed to observational evidence. Despite a seemingly straightforward exchange between game theory and representation learning, the benefits pushing the acquired samples beyond recreations should be considered the game's side effects. Since the properties of the counterparts' interaction resemble almost an Escherian strange loop, the discriminator shaping the generator which, feeding its results back, attempts to change the discriminator's conception of observed and generated, it is necessary to watch out for signs of the model's empirical underdetermination.

In the machine learning context, the most pertinent kind of empirical underdetermination corresponds to the notion of *underfitting*. It occurs if a learning model, arranged according to a certain architecture possessing a fixed representational capacity, fails to correctly account for the entirety of the structural pattern entailed in the data which serve as the model's training input. In other words, the underfitted model fails to recover the true data generating distribution and instead obtains a bogus approximation that can account for only an arbitrary portion of the data (cf. Goodfellow et al., 2014, pp. 108-110). Such a model then cannot *reliably generalise* beyond the training input, as it failed to learn a close enough approximation of the underlying distribution. A generative model, which is not specifically designed to produce recreations, underfits the modalities of a possibility space when the majority of its samples manifest a low diversity gravitating towards recreations of the observational evidence and/or self-repetition. With respect to the adversarial learning of the underlying distribution, such a model suffers from mode collapses/drops stemming from a failure to reach the equilibrium at which the generator learns and sustains all the distribution's modes implied within the available evidence (cf. Arora et al., 2017; 2018). Such an underfitted generative model thus cannot reliably generalise beyond the evidence, which hampers its sustained production of highly diverse samples coming from all the left out regions of the respective possibility space (cf. ibid.).

Besides theoretical analyses, in empirical settings, as the process of reaching the equilibrium remains a side effect of an adversarial exchange, a certain degree of mode collapse, and thus of the model's underfitting, is always present. Considering the landscape of artificial representation learning in general, there aren't yet any practical methods offering universal guarantees of the optimal generalisation performance regarding arbitrarily large and complex empirical datasets. Facing real world observational evidence, the model is thus

expected to learn an approximation of the distribution which, however subject to a degree of underdetermination, proves *instrumental* for solving the task at hand[5]. Since mode collapses and/or drops are never total, permitting the model to consistently reach a degree of generalisation, it is always possible to obtain samples from some of the left out regions. Further experimentation with different initialisations, architectural patterns or representational capacity of the model's components, in the present case of the two competing neural networks, then yields samples from other left out regions. In other words, it is nearly impossible to acquire a complete map of the possibility space at one go if it pertains to non-trivial observational evidence. Instead, the process is considered exploratory, gradually informing the theoretical synthesis through incoming samples, itself driven by an adversarial exchange pushing the model towards novel insights. If by then, as a group, the instances of the model successfully achieve generalisation regarding the existing evidence, theories resulting from the synthesis gain resistance against unconceived alternatives. Crucially, this occurs even without a retreat of theoretical commitments to what is possibly artificial parsimony, since at worst the samples from the left out regions corroborate the state-of-the-art theories. Conversely, at best, the commitment might be advanced, as the samples contribute to ruling out yet to be conceived theories predicting phenomena which would become incongruous with the state-of-the-art theories. Both gains

[5] The model can get also stranded in an overfitted state, arising from what is usually described as memorisation of the training data (observational evidence), likewise hampering its capability to generalise beyond the evidence. However, as it is underfitting which mostly imperils the current generative adversarial models, and the subject at hand comprises mainly underdetermination, a discussion of overfitting would diverge from the goal of the paper.

would derive from epistemic warrants building on the increased confidence in what is possible considering the recovered modalities of the respective possibility space.

4.3 Limits of Software Intensive Science

Even though adversarial machine learning can offer a window into the left out regions of possibility spaces, apart from suffering mode collapses and/or mode drops, which cause it to underfit the evidence, it is also subject to constraints associated with *software intensive science* (cf. Symons and Horner, 2014; Symons and Horner, 2017; Symons and Horner, 2019). Arguably, finite knowers will exploit every opportunity to extend their cognitive reach. The proposed synthesis between human and machine learning expands available evidence by samples from the left out regions of possibility spaces. Evidence expanding inductive inferences provided by machine learning change the nature of human cognitive finitude. Machine learning-based ampliative inferences produce unconceived facts which we have not been able to consider due to their contingent, non law-like nature (thus pushing the limits of knowledge in the empirical sense, cf. Rescher, 2006, pp. 95-104).

However, epistemic justifiability of such ampliative inferences depends not only on how well the generative model learns to generalise beyond the evidence, but crucially on reliability of the underlying software platform. Symons and Horner (2014; 2019) showed that if the underlying software exhibits high conditionality, its error distribution cannot be characterised, which leaves no room for principled reasoning about the program's (software's) reliability. Their claim relies on practical impossibility of testing a sufficient number of the program's execution paths (ibid.). As a result, it is *a priori* out of question to reach a confidence level that would justify any assumption about reliability of the software

at hand (ibid.). The impossibility of realising a satisfactory test coverage distinguishes software intensive from non-software intensive science (ibid.).

The synthesis between human and machine learning falls in the former category. It is thus an open question by how much we can improve epistemic warrant of realist theories while considering samples from generative adversarial models. Following this line of reasoning might even lead to a disappointing conclusion that software intensive science, utilising machine learning models, cannot support a convergence to the truth account associated with scientific inquiry, and its realist philosophy in particular[6]. This would diminish the thrust of the proposal, making it effective only as a remedy for underdetermination by available evidence. Crucially, the proposal could be then used by realists as well as empiricists to ward off the instrumentalists' attack with unconceived alternatives (even if for empiricists this kind of underdetermination does not play a significant role). By being equally relevant to the realist and empiricist philosophy of science, the proposal would lose its exclusive support for the realist side of the debate.

The paper argues that such a reading would not be entirely correct, because it omits an (important) qualification to the impossibility of characterising the error distribution of software exhibiting high conditionality. To achieve generality and objectivity, and thus to constitute an upper-bound on testability, the error distribution of a piece of software (program) depends on *all* inputs that can invoke its full path complexity, i.e. observing all, or almost all, depending on a sought for confidence interval, of the program's execution paths. In their breakthrough result, and its later elaboration, Symons and Horner demonstrated

[6] We are indebted to John Symons for pointing out this limit of the human-machine learning synthetisation of (realist) scientific theories.

that any procedure that would reach a sufficient test coverage is provably intractable, given any computer program of a non-trivial conditionality. This outcome, however, has an interesting corollary: Beyond a trivial level of conditionality there are no a priori distinctions between reliability of computer programs. In theory, software intensive science, including the subset utilising machine learning, then succumbs to Hume's Problem (1739/1978), because, a priori, we cannot justify any inductive inference about the reliability of a computer program exhibiting high conditionality. If there are no *a priori* distinctions regarding reliability, humans do not have any *a priori* guidance on whether to follow or avoid unconceived phenomena sampled from generative machine learning models. Our prior knowledge notwithstanding, we cannot a priori rule out the possibility that an otherwise well-behaved model will for a certain input produce a phenomenon which does not belong the possibility space of available evidence. It is simply because the machine learning model, i.e. a computer program, cannot distinguish among the consequences of all possible inputs, and so cannot the human developer/user due to the impossibility to achieve a sufficient test coverage. Even though for the model there is no difference between the a priori and a posteriori assumption of reliability, there is one for the humans.

To reach generality and objectivity constitutive of an upper bound on software error distribution, the 'no distinctions' argument dismisses fortuitous conditions where inputs and the machine learning model generate legitimate results. Fortuitous conditions can be ascertained only a posteriori and never without justification provided by human scientists. Fortuitous conditions occur when a model, i.e. a piece of high conditionality software, fits the training data (i.e. evidence, in the present case fitting the learning signal provided by the discriminator) well enough so that it reliably interpolates beyond the evidence to sample unconceived phenomena from the left out regions of the possibility space. Generalisation,

i.e. in-distribution interpolation, depends on reliable software and can thus serve as a proxy to determine whether we reached fortuitous conditions or not. Therefore, if we seek to a posteriori dispel the 'no distinctions' argument, the human scientists, synthesising a (realist) theory based on their insights and machine samples, need to reflect on the training data (evidence), model, and their mutual fit with the assumptions about the target domain (possibility space). Without this step, the machine learning model becomes subject not only to difficulties identified by Symons and Horner but also to 'No Free Lunch' theorem (Wolpert, 1996). The latter applies because the weighting over targets (possibility spaces) is based on the distribution of erroneous inferences which cannot be a priori specified unless we can prove that a supposed shape of the possibility space is well-matched by a particular model. Due to Symons and Horner any such inferences will be contested and, moreover, also suffering from wrong assumptions about the uniformity between a model and the possibility space of available evidence. Hence there are no a priori distinctions among the range of applicable machine learning models, only a posteriori insights which lead to the following conclusion.

In less technical terms, without human scientists, experimenting on possible variants of the training data (evidence), model, and their mutual fit with the assumptions about the target possibility space (see Section 4.2), the synthesis will be most likely unsuccessful (i.e. depending on epistemic luck). In this sense, Symons and Horner showed that machine learning models, i.e. high conditionality software, cannot a priori guarantee a fully autonomous convergence to the truth account associated with scientific inquiry, and its realist philosophy in particular. However, their argument does not preclude a version of software intensive science where humans intensively experiment on evidence and machine learning models to find uniformity with the possibility space at hand and its underlying

probability distribution. In other words, despite a priori objectivity and generality of the 'no distinctions' argument, it is still very much possible to find a posteriori fortuitous conditions. They can provide a window into possibility spaces of available evidence and extend our understanding by so far unconceived phenomena.

5. Beyond Hypothetical Musings: Practical Prospects of the Synthesis

Regarding practical realisations of the synthesis, in several fields, there are recently emerged applications of generative adversarial models (f.e. cf. Mustafa et al., 2017; Paganini et al., 2018; Albert et al., 2018), implying that its hypothesised epistemological benefits find their real counterparts when scientists integrate artificial representation learning into their methodological toolboxes. However, the synthesis' full potential is yet to be appreciated, as the initial impetus for introducing generative models didn't come from a concern for the left out regions but rather from an interest seeking to acquire a cheap way of simulating the studied phenomena. The cheapness delivered by machine learning comes in two forms of which the second paves the way for exploring the possibility spaces of observational evidence. In its first form the cheapness relates to often prohibitive computational costs of numerical simulations impeding experimental and theoretical developments alike (ibid.). As a generative model produces, at a relatively low computational cost, samples of phenomena difficult to observe and/or expensive to faithfully simulate, it lays the groundwork for the second kind of cheapness. This latter form disposes of the necessity to conceive an antecedent mathematical model underlying any numerical simulation. Since the multilayer feed-forward neural network, which constitutes adversarial models, proves to be a *universal approximator* regarding arbitrary continuous functions (Hornik et al., 1989), such prior model is no longer necessary. This property of neural networks increases the degree of

freedom from imposing an artificially parsimonious theoretical commitment at the very beginning of the enterprise, which would stem from choosing an ill-fitting antecedent model. Learning approximations of evidence generating distributions without imprinting the biases of preconceived models onto the resulting theory contributes to exposing the left out regions of possibility spaces. Avoiding antecedent models, while using adversarial learning to push generative models in the direction of the left out regions, might thus aid to deliver realist theories from the threat of unconceived alternatives. Consequently, the synthesis' exploratory stage remains mostly unbiased, as the human cognitive inputs enter the picture only later on. Leaving one's options open by relying on agnostic, i.e. model-free, universal approximators would postpone the commitment until scientists inspect the phenomena sampled from the left out regions in an effort to pre-empt unconceived alternatives.

Arguably, quite close to a full-fledged synthesis is a recent methodological prototype concerned with generating weak lensing convergence maps for an instance of the ΛCDM cosmological model (Mustafa et al., 2017). As the model's testing and inferring its correct parameters vis-à-vis our universe involves consulting generated maps pertaining to variously initialised instances of the standard model, the original concern was with computational economy allowing for agile simulations (ibid.). However, as it turned out generative adversarial models can in fact produce *new maps* congruent with an instance of the standard model without being ever introduced to its summary statistics apart from an observational exposure in the form of a limited sample of maps coming from the numerical simulation (ibid.). Considering the study of dark matter, energy and related phenomena, the generative model provides an avenue for unbiased exploration of the possibility space comprising observations of a universe described by the corresponding instance of the standard model. As the samples from such a model convey perspectives on the virtual

universe, the generative model delivers foremost modal knowledge about what kinds of cosmological structures are possible given the ΛCDM model and its instance parameters.

Motivated by the same goal to select the best fitting physics model of the Universe, Zamudio-Fernandez et al., 2019 proposed to use a generative adversarial network to produce 3D distributions of cosmic neutral hydrogen (HI). Compared to hydrodynamic simulations, adversarial models can generate distributions of HI five orders of magnitude faster (ibid.). The increased efficiency, provided by the generalisation capability of a model trained on samples from the simulations, allows to survey a much larger portion of the possibility space of available evidence, including its left out regions. With more samples of HI distributions generated, the actual 21cm emissions from HI captured by radio telescopes can be compared to a wider range of theoretical predictions, i.e. 'synthetic' observables generated by the adversarial model (ibid.). This manoeuvre allows better utilisation of data coming from cosmological surveys (ibid.). However, by working with phenomena from the left out regions, it also adjusts the theory/model under construction to modalities of the possibility space of available evidence, thus lowering its exposure to unconceived alternatives.

Pursuing a similar goal, Rodríguez et al., 2018 used a generative adversarial network to approximate distributions of matter that can be used to sample synthetic cosmic webs, complex networks of cosmic structures and interactions which can provide insights into dark matter, dark energy or laws of gravity (ibid.). Building on the work of Mustafa et al., 2017 mentioned above, Rodríguez et al., 2018 trained the adversarial model on examples of cosmic webs produced by classical N-body simulations. The motivation was once again to remove the computational bottleneck of simulations which might prevent the full

realisation of cosmological surveys (ibid.). The generalisation capability of adversarial models can alleviate the bottleneck by making it possible to produce rich sets of theoretical predictions, i.e. synthetic cosmic webs, that can be compared to empirical data from the surveys. As in the previous experiments, cosmic webs generated by the adversarial models do not exhibit correlations with training data (ibid.), which indicates that they come from the left out regions and can be used to align the theory/model under construction with modalities of the possibility space of available evidence.

Apart from astrophysics and cosmology, identically motivated attempts to reduce the costs of simulations are emerging in high-energy particle experiments at the Large Hadron Collider (LHC, de Oliveira et al., 2017; Paganini et al., 2018; Hashemi et al., 2019; Di Sipio et al., 2019). Paganini et al., 2018 showed that it is possible to use adversarial models to generate, or in the traditional sense simulate, synthetic particle showers in electromagnetic calorimeters. In the context of LHC's ATLAS or CMS experiments, alleviating the computational bottleneck allows to encompass a wider range of theoretical assumptions reflected as different subatomic particle collisions and interactions (ibid.). The generalisation capability of adversarial models enables to sample synthetic energy depositions of particle showers whose diversity suggests an expanded reach into the possibility space of available evidence (cf. ibid.). Similarly to the cosmological experiments, adversarial models can be used to extend our reach to regions of the possibility space of available evidence that have been left out so far, ask about its modalities, and thus at least partially diminish the likelihood of unconceived alternatives emerging in the future.

Finally, attempts are made to use adversarial models to produce new effective field theories (Erbin and Krippendorf, 2018). Erbin and Krippendorf constructed a proof-of-concept

adversarial model able generate new samples from a class of supersymmetry models. The experiment has an epistemological significance because it implies that adversarial models can be applied to survey the solution space of string theory and generate new predictions (ibid., p. 5). Similarly to the previous empirical cases, the epistemic concern is with modalities of the possibility space and methods that can help to tease them out in a bid to pre-empt unconceived alternatives.

Even though generative adversarial models are successfully used in other areas of science, with examples including generation of materials (e.g. Kim et al., 2020) or drug discovery (e.g. Méndez-Lucio et al., 2020), these applications do not seek modal surveys of possibility spaces for theoretical purposes. Their goal lies in generating new samples from a priori delimited regions supposedly holding new viable materials or molecules. The surveys are thus conditioned to stay only in the known regions of possibility spaces and serve to practical rather than theoretical purposes. Therefore, the parts of astrophysics, cosmology, and high energy physics, which start to experiment with adversarial models, are worth observing, for they hold a promise to begin synthesising the new breed of realist theories based on an extended epistemic reach of the human-machine learning nexus. Practically, the above outlined cases approach full syntheses, since they are only a step from applying the acquired knowledge to asses and possibly revise the theories' commitments so as to reflect the modalities determining the possibility spaces of available evidence. In so doing, the theories' exposure to unconceived alternatives would remain limited, as the modal knowledge of the possibility spaces helps to pre-empt yet to be observed phenomena possibly in conflict with the synthesised theories.

Considering ramifications of such a synthesis at the meta-theoretical level, the exploration of possibility spaces by adversarial representation learning comes out as a kind of *perspectival modelling* (Massimi, 2018). This kinship derives from the emphasis attached to modal knowledge, delimiting the range of possibilities, rather than to individual phenomena acquired by the exploration, however important they might be for a theory at hand. Going for the modal dimension derived from the obtained representational content permits pre-emptive discerning between the kinds of phenomena which might be observed in the future and those ruled out by a recovered modality (cf. Massimi, 2018, pp. 338-339). As a result, metaphysically delicate realist commitments (apart from unconceived alternatives also bearing in mind the issue of conceived inconsistent rivals) might be now secured even without the recourse to undue selectivity, as the acquired pictures of possibility spaces provide a framework for assessing theories' exposure to unconceived alternatives as well their standing with conceived rivals. By consulting this modal dimension, while being engaged in realist theory building, the anti-realist's job of hunting for the unconceived becomes a more demanding affair (cf. ibid.) than referring to the past flaws of exploring the possibility spaces. The realist could act accordingly, and instead of the backwards orientation embark on a forward looking quest which would, carving out the modal dimensions of possibility spaces, provide a new more resilient kind of the selective realist commitment (cf. Massimi, 2018, pp. 348-349). Finally, the question of "*What can artificial intelligence do for scientific realism?*" finds its answer in helping to facilitate the shift of realist commitments towards modal knowledge of possibility spaces, which would balance the retreats of past selectivity caused by anti-realist pressures of both kinds, those blaming prior missteps as well as the ones prophesying inevitability of future breakdowns. As to the subject of underdetermination, synthesising between human scientists and artificial

representation learning models would then yield theories whose epistemic warrants enjoy a lesser degree of underdetermination by available evidence.

References

Alai M (2004) A.I., Scientific Discovery and Realism. Minds Mach 14:21-42

Alai M (2014) Novel Predictions and the No Miracle Argument. Erkenntnis 79:297-326

Albert A, Strano E, Kaur J, Gonzáles M (2018) Modeling urbanization patterns with generative adversarial networks. arXiv:1801.02710v1 [cs.LG]

Arora S, Ge R, Liang Y, Ma T, Zhang Y (2017) Generalization and Equilibrium in Generative Adversarial Nets (GANs). In: Proceedings of International Conference on Machine Learning (ICML). August 6-11 Sydney, Australia, pp 224-232

Arora S, Zhang Y, Risteski A (2018) Do GANs learn the distribution? Some theory and empirics. In: Proceedings of 6th International Conference on Learning Representations (ICLR). April 30 - May 03 Vancouver, Canada

Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L, Zdeborová L (2019) Machine learning and the physical sciences. Rev Mod Phys 91(4)

Chakravartty A (2008) What you don't know can't hurt you: realism and the unconceived. Philos Stud 137:149-158

Chakravartty A (2017a) Scientific Realism. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Summer 2017 Edition) https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.

Chakravartty A (2017b) Reflections on new thinking about scientific realism. Synthese S.I. New Thinking about Scientific Realism:3379-3392

Dawid R (2013) String Theory and the Scientific Method. Cambridge University Press, Cambridge

Di Sipio R, Giannelli MF, Haghighat SK, Palazzo S (2019) DijetGAN: a Generative-Adversarial Network approach for the simulation of QCD dijet events at the LHC. J High Energ Phys

Džeroski S, Langley P, Todorovski L (2007) Computational Discovery of Scientific Knowledge. In: Džeroski S, Todorovski L (eds) Computational Discovery of Scientific Knowledge: Introduction, Techniques, and Applications in Environmental and Life Sciences. Springer, Berlin, pp 1-14

Egg M (2014) Expanding Our Grasp: Causal Knowledge and the Problem of Unconceived Alternatives. Br J Philos Sci 67(1):115-141

Erbin H, Krippendorf S (2018) GANs for generating EFT models. arXiv:1809.02612v1 [cs.LG]

Fedus W, Rosca M, Lakshminarayanan B, Dai AM, Mohamed S, Goodfellow I (2018) Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step. In: Proceedings of 6th International Conference on Learning Representations (ICLR). April 30 - May 03 Vancouver, Canada

van Fraassen BC (1980) The Scientific Image. Clarendon Press, Oxford

Gillies D (1996) Artificial Intelligence and Scientific Method. Oxford University Press, Oxford

Giza P (2017) Automated discovery systems and the inductivist controversy. J Exp Theor Artif Intell 29(5):1053-1069

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Networks. In: Proceedings of 27th Advances in Neural Information Processing Systems (NIPS). December 8-13 Montreal, Canada

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, Cambridge, MA

Hashemi A, Amin N, Datta K, Olivito D, Pierini M (2019) LHC analysis-specific datasets with Generative Adversarial Networks. arXiv:1901.05282v1 [hep-ex]

Holland JH, Holyoak KJ, Nisbett RE, Thagard P (1986) Induction. Processes of Inference, Learning, and Discovery. MIT Press, Cambridge, MA

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359-366

Hume D (1739/1978) A Treatise on Human Nature. Book I: On Human Understanding. Oxford University Press, Oxford

Humphreys P (2004) Extending Ourselves: Computational Science, Empiricism, and Scientific Method. Oxford University Press, New York

Humphreys P (2011) Computational Science and Its Effects. In: Carrier M, Nordmann A (eds) Science in the Context of Application. Springer, Dordrecht, pp 131-142

Humphreys (2020) Why Automated Science Should Be Cautiously Welcomed. In: Bertolaso M, Sterpetti F (eds) A Critical Reflection on Automated Science. Will Science Remain Human? Springer, Cham, pp 11-26

Kim B, Lee S, Kim J (2020) Inverse design of porous materials using artificial neural networks. Science Advances 6(1)

Kingma DP, Welling M (2014) Auto-Encoding Variational Bayes. In: Proceedings of 2nd International Conference on Learning Representations (ICLR). April 14-16 Banff, Canada

Kukla A (1996) Does every theory have empirically equivalent rivals? Erkenntnis 44(2):137-166

Langley P (2000) The computational support for scientific discovery. International Journal of Human-Computer Interaction 53:393-410

Laudan L (1990) Demystifying underdetermination. In: Wade Savage C (ed) Scientific theories. University of Minnesota Press, Minneapolis, MN, pp 267-297

LeCun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521:436-44

Magnus PD (2010) Inductions, Red Herrings, and the Best Explanation for the Mixed Record of Science. Br J Philos Sci 61(4):803-819

Massimi M (2018) Perspectival Modeling. Philos Sci 85(3):335-359

Méndez-Lucio O, Baillif B, Clevert DA, Rouquié D, Wichard J (2020) De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature Communications 11

Mizrahi M (2016) Historical Inductions: New Cherries, Same Old Cherry-picking. International Studies in the Philosophy of Science 29(2):129-148

Mizrahi M (2017) The History of Science as a Graveyard of Theories: A Philosophers' Myth? International Studies in the Philosophy of Science 30(3):263-278

Mustafa M, Bard D, Bhimji W, Al-Rfou R, Lukić Z (2017) Creating Virtual Universes Using Generative Adversarial Networks. arXiv:1706.02390v1 [astro-ph.IM]

de Oliveira L, Paganini M, Nachman B (2017) Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis. Comput Softw Big Sci 1

Paganini M, de Oliveira L, Nachman B (2018) Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. Physical Review Letters 120:042003-1-042003-6

Psillos S (1999) Scientific Realism: How Science Tracks Truth. Routledge, London

Putnam H (1975) Mathematics, Matter and Method, Philosophical Papers, Volume I. Cambridge University Press, Cambridge

Quine WVO (1951) Two Dogmas of Empiricism. Philos Rev 60(1):20-43

Quine WVO (1970) On the Reasons for Indeterminacy of Translation. J Philos 67(6):178-183

Quine WVO (1975) On Empirically Equivalent Systems of The World. Erkenntnis 9:313-328

Radovic A, Williams M, Rousseau D, Kagan M, Bonacorsi D, Himmel A, Aurisano A, Terao K, Wongjirad T (2018) Machine learning at the energy and intensity frontiers of particle physics. Nature 560:41-48

Rescher N (2006) Epistemetrics. Cambridge University Press, Cambridge

Rodríguez AC, Kacprzak T, Lucchi A, Amara A, Sgier R, Fluri J, Hofmann T, Réfrégier A (2018) Fast Cosmic Web Simulations with Generative Adversarial Networks. Comput Astrophys Cosmology 5

Saatsi J (2015) Historical inductions, Old and New. Synthese S. I. Conceived Alternatives:1-15

Stanford PK (2006) Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford University Press, New York, NY

Stanford PK (2017) Underdetermination of Scientific Theory. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Winter 2017 Edition). https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/

Stolle R, Bradley E (2007) Communicable Knowledge in Automated System Identification. In: Džeroski S, Todorovski L (eds) Computational Discovery of Scientific Knowledge: Introduction, Techniques, and Applications in Environmental and Life Sciences. Springer, Berlin, pp 17-43

Symons J, Horner J (2014) Software Intensive Science. Philos Technol 27(3):461-477

Symons J, Horner J (2017) Software Error as a Limit to Inquiry for Finite Agents: Challenges for the Post-human Scientist. In: Powers TM (ed) Philosophy and Computing. Springer, Cham

Symons J, Horner J (2019) Why There is no General Solution to the Problem of Software Verification. Found Sci DOI:10.1007/s10699-019-09611-w

Thagard P (1988) Computational Philosophy of Science. MIT Press, Cambridge, MA

Wolpert D (1996) The Lack of A Priori Distinctions Between Learning Algorithms. Neural Comput 8:1341-1390

Wray KB (2018) Resisting Scientific Realism. Cambridge University Press, Cambridge

Zamudio-Fernandez J, Okan A, Villaescusa-Navarro F, Bilaloglu S, Cengiz AD, He S, Levasseur LP, Ho S (2019) HIGAN: Cosmic Neutral Hydrogen with Generative Adversarial Networks. In: Machine Learning and the Physical Sciences Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS)

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2017) Understanding deep learning requires rethinking generalization. In: Proceedings of 5th International Conference on Learning Representations (ICLR). April 24-26 Toulon, France