Expanding Observability via Human-Machine Cooperation

Petr Spelda[1] and Vit Stritecky

Department of Security Studies, Institute of Political Studies, Faculty of Social Sciences, Charles University

Accepted Manuscript, *Axiomathes* 32, 819-32, https://doi.org/10.1007/s10516-022-09636-0

Abstract

We ask how to use machine learning to expand observability, which presently depends on human learning that informs conceivability. The issue is engaged by considering the question of correspondence between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, states of affairs. A possible answer lies in importing out of reference frame content which could provide means for conceiving further observability counterfactuals. They allow us to define high-fidelity observability, increasing the level of correspondence in question. To achieve high-fidelity observability, we propose to use generative machine learning models as the providers of the out of reference frame content. From an applied point of view, such a role of generative machine learning models shows an emerging dimension of human-machine cooperation.

Keywords: observability; machine learning; conceivability; human-machine cooperation

1. Introduction

Can unaided human conceivability negotiate an agreement between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena? The intricate works on conceivability provided a host of anthropocentric perspectives on the question (e.g., Yablo 1993; Tidman 1994; Hill 1997; Szabo and Hawthorne 2002; Chalmers

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

2002; Kung 2010; Rescher 2020) and some even challenged the consistency of conceivability (Campbell et al. 2017; Fiocco 2020). There is an underexplored connection between conceivability and generalisations from observed states of affairs that can shed a new light on the relation between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena. Conceivability dominated the debate because its relation to generalisations was treated as *unidirectional*. That is, conceiving interventions that lead to new observations from which it is possible to generalise and produce observability counterfactuals whose fidelity to the observable, yet so far unobserved or unconceived, phenomena is at stake.

We propose to treat the relation between conceivability and generalisations as *bidirectional*. Bidirectionality involves non-anthropic cognitive devices, generative machine learning models, attaining generalisations independent of the human cognitive baseline, and thus providing new content for inferences about so far unobserved or unconceived states of affairs. Within the anthropocentric epistemology, the relation remains unidirectional, with conceivability providing generalisations. However, if a kind of syncretic epistemology is considered, then by synthesising between human and *artificial* representation learning the relation becomes bidirectional. A generalisation-capable generative machine learning model produces samples enabling access to further states of affairs. Machine learning can thus provide fresh content for further observability counterfactuals. Although we might never observe the entirety of phenomena implied by our generalisations, which prompts the question of correspondence between conceived observability counterfactuals and the actual observations of yet unobserved or unconceived phenomena, a partial remedy could be found in the samples from generative machine learning models. These samples provide fresh content for conceiving much broader sets of observability counterfactuals.

In case it falls within the bounds of established generalisations, fresh content acquired by artificial means might vindicate our observability counterfactuals, easing the dilemma over the fidelity between observations and observability counterfactuals. If falling outside, fresh content should lead to their revision, contributing to maintaining a correspondence between observations and observability counterfactuals. The question of what would transpire if an observable, yet so far unobserved or unconceived, phenomenon presented itself to us could be made more manageable by high-fidelity observability, synthesising between human and machine representation learning.

The rest of the paper is structured as follows. First, we review the way in which conceivability is applied to learn generalisations that underpin the reachable horizon of observability. Second, we show how competitive coevolution establishes adversarial generative machine learning models (Goodfellow et al. 2014). These models learn generalisations capable of producing unobserved or unconceived yet observable samples that provide fresh content for observability counterfactuals. Third, building on the bidirectional relation between conceivability and generalisations, we propose how to utilise epistemic gains that follow from high-fidelity observability posited on a combination of human and machine learning.

2. Conceivability Providing Generalisations

To elaborate on the anthropocentric case where conceivability provides generalisations, we relate observability to invariance. Invariance plays the central role in acquiring content for observability counterfactuals. Invariances underpin generalisations that in turn support explanans in terms of counterfactual dependencies, thus answering questions of what would be observed had the things been different. Such a function of invariances is captured by

James Woodward's interventionist account of explanation (2003; Hitchcock and Woodward 2003; Woodward 2000). It was further developed by Alexander Reutlinger into a monist approach, integrating both causal as well as non-causal explanations (2016, 2018). Although there are other attempts at such an integration, e.g., Saatsi and Pexton (2012), and Woodward himself considered non-causal explanations as well (cf. Hitchcock and Woodward 2003, pp. 191-92; Woodward 2018), Reutlinger's theory is best suited for our present purposes. His monist counterfactual theory of explanation (CTE) comprises a framework of structure, veridical, inferential, and dependency conditions, subsuming the counterfactual dependence of an explanandum on explanans regardless of whether we consider the explanans causal or non-causal (2018, pp. 78-81). CTE involves a set of generalisations, initial conditions, and statements about observability (the structure condition) and shows that had the initial conditions been different the generalisations would still support counterfactual instances of the phenomena in question (reflecting the dependency condition of Reutlinger's framework, cf. ibid.).

Considering CTE in the light of observability, first it is necessary to conceive patterns of phenomenal occurrences which furnish generalisations and underpin the structure condition (together with a set of varying initial conditions). Otherwise, it will not be possible to resolve the consequent dependency condition which determines the supported observability counterfactuals that establish explanations. Such conceived patterns reflect agents' experiences involving phenomena that interacted with the segments of an environment counterfactually determining instances of the explananda (mere segments are considered to prevent explosive admissions of irrelevant facts which do not appear to be counterfactually involved with the explananda and would thus dilute the explanatory relevance of explanans, cf. Reutlinger 2016). In the anthropocentric setting, the unidirectional relation between

conceivability and generalisations is foundational for observability. It provides a framework for resolving the compatibility of conceived observability counterfactuals with the underlying frame of reference, thus permitting to form consistent explanations of the phenomena which populate it.

The issue of whether the frame of reference latches onto the objective features of the world recedes and the attention shifts to the fidelity between conceived observability counterfactuals and so far unobserved or unconceived phenomena that might occur within the reference frame. Since high-fidelity observability combines human and machine learning, a question remains how to relate it to anthropocentric theories of explanation and to explanations. Both depend on the rigidity of reference frames, regardless of whether it stems from the objectivity of non-modal observability or from latching onto the objective features of the world (cf. Monton and van Fraassen 2003, p. 411; Ladyman 2004, p. 762 respectively). Under constructive empiricism, even a provisionally fixed reference frame defines the relevant epistemic context and imparts objectivity to observability (Monton and van Fraassen 2003, p. 411). What lies beyond the reference frame becomes for the time being irrelevant and attempts to access it entail the risk of relying on inadequate metaphysics. On the other hand, even if the risk is selectively taken, the resulting reference frame ends up equally rigid, with the fidelity of observability derived from the modality used to determine the objective features of the world.

CTE is considered a useful reference point among the anthropocentric approaches because it integrates causal as well non-causal explanations. Its structure condition is prior to the dependency condition and thus reflects the unidirectionality between conceivability and generalisations.

2.1 The Rigidity of Reference Frames and High-Fidelity Observability

To achieve high-fidelity observability by combining human and machine learning, the rigidity of anthropocentric reference frames must be relinquished. Making our reference frames flexible would allow us to accommodate further observability counterfactuals stemming from bidirectionality between conceivability and generalisations. It is natural to suppose that the fidelity of observability increases with the rigidity of reference frames. We argue the *opposite* because the generalisations providing further (fresh) content for conceiving observability counterfactuals come from outside (machine learning models acquiring phenomenal representations) and thus *relax* the dependence of observability on the reference frame. As a result, the rigidity of reference frames seems unnecessary. It might be retorted that high-fidelity observability merely replaces the epistemic anxiety over the correspondence between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena with a worse tension of not knowing what comprises the reference frame.

This neglects a possibility that the phenomenal samples, coming from outside of the reference frame as fresh content for conceiving further observability counterfactuals, could be evaluated within the CTE framework. The evaluation would either confirm or correct the generalisations which constitute the reference frame. The dependency condition would be used as a resolver of the generalisations' support for further observability counterfactuals (cf. Reutlinger 2018, p. 79) based on the out of reference frame content sampled from a generative machine learning model. We would gain new information about imperfections of our reference frames, i.e., cognitive biases causing unaided human conceivability to become insufficient, and an opportunity to achieve high-fidelity observability by *synthesis*.

By being able to generalise, generative machine learning models provide fresh content for conceiving further observability counterfactuals, offering additional information on the very generalisations that establish our frame of reference. Even though synthesised by human conceivability, these counterfactuals originate truly outside of the respective reference frame. Such a disposition is interesting because it does not require that the epistemic community integrates agents or devices cognitively different from humans (and their reference frames) into an extended, rigid reference frame to stabilize observability (cf. Monton and van Fraassen 2003, p. 411). Similarly, while selectively committing to metaphysical inflations by engaging modality (cf. Chakravartty 2017), a succession of reference frames can evolve, evermore precisely latching onto the objective features of the world. Using different means, both positions intend to provide a rigid reference frame, establishing observability as incontestable as possible.

High-fidelity observability sharpens our present frame of reference by opening it to artificially produced content which makes it less rigid. Although the further observability counterfactuals come from outside of the reference frame, they are *conceived* by humans. Regarding the first of above positions, no other epistemic agency, and thus conceivability, is involved in the synthesis. As a result, there is no need to reengineer the reference frame to maintain its rigidity while basing observability counterfactuals on content from outside. Concerning the second position, out of reference frame samples engage modality only indirectly, in terms of what is possible yet so far unobserved or unconceived, providing a fresh foundation for conceiving further observability counterfactuals. Again, reengineering of the reference frame would become unnecessary.

The synthesis would require a new reference frame only if artificial representation learning became confounded with a sort of hypothesised artificial intelligence aligned with human agents (at least on some minimal level, cf. Wilks 2017). Apart from the question of the viability of such a phenomenon (cf. ibid.), it is unclear whether it would provide a better prospect for high-fidelity observability than mostly unconstrained explorations by plain representation learning. Artificial representation learning via generative machine learning models does not involve an epistemic agency. It merely provides out of reference frame generalisations used to produce phenomenal samples (i.e., fresh content for conceiving further observability counterfactuals) which can augment human conceivability.

The key to high-fidelity observability lies in acquiring some content outside of our frame of reference. The content then provides better prospects for achieving the correspondence between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena. A decreasing dependence of observability on the reference frame lowers the reference frame's rigidity. This in turn permits us to consider further observability counterfactuals and re-evaluate the generalisations (within the CTE framework) which underlie our reference frame, a process that leads towards high-fidelity observability.

3. Generalisations Providing Conceivability

High-fidelity observability rests on the assumption that the out of reference frame generalisations provide fresh content for conceiving further observability counterfactuals. In this sense, some of the generalisations underpinning less rigid reference frames come not only from outside but they are also *prior to exercising* human conceivability. This in turn gives rise to the notion of *bidirectionality*. As the generalisations learnt by generative machine learning models produce so far unobserved or unconceived phenomenal content

(see Spelda and Stritecky 2021 for applications of generative machine learning models in various scientific fields), our reference frame ceases to depend exclusively on the human epistemic agency. To enrich our reference frame in this way requires that a generative machine learning model learns to approximate the probability distribution underlying an empirical area where there is an interest to conceive further observability counterfactuals. Learning out of reference frame generalisations requires a mechanism that by processing samples drawn from the underlying distribution produces its approximation, allowing generation of novel samples consistent with the original distribution (cf. Goodfellow et al. 2016, p. 645). Given an empirical domain, this mechanism allows the generative machine learning model to produce fresh content for conceiving observability counterfactuals. They provide further guidance in 'observable yet so far unobserved or unconceived' epistemic situations because the content comes from outside of the reference frame. High-fidelity observability depends on how accurately the model learns to approximate the distribution over the domain of samples from which the model generalises to produce fresh, out of reference frame content. The study of artificial generative models suggests that an efficacious distribution learning comes from locking two populations of reference frames, or a population and an environment, into competitive coevolution (cf. respectively Olsson et al. 2018; Wang et al. 2018, for the underlying generative adversarial model refer to Goodfellow et al. 2014).

Generalising generative models, able to interpolate from observed samples to unobserved or unconceived samples (considering the anthropocentric perspective), provide fresh content which can be used to conceive further observability counterfactuals. Considering them as possible explananda occurrences, involving 'observable yet so far unobserved or unconceived' epistemic situations (a step justified by approximating the underlying

distribution), it is possible to examine the validity of generalisations underpinning our frame of reference. In line with the CTE's dependency condition, we can assess whether the existing generalisations support further counterfactuals or a revision is due, either way advancing towards high-fidelity observability. The opportunity to review the generalisations is made possible by assuming that some of CTE's initial conditions, and content they prompt, originate outside of our frame of reference. As none of the CTE's conditions precludes such an import, the synthesis between human and machine learning strengthens invariances founding our reference frame, since they begin to reflect observability counterfactuals produced using non-anthropic sources. The import decreases the rigidity of our reference frame and increases the explanatory power of CTE by overcoming the unidirectional understanding of conceivability and generalisations. It also utilises the monist nature of CTE, insomuch as the most successful present artificial representation learning relies on neural networks and achieves generalisations by utilising the neural network's mathematical property of *non-causal* universal approximation (cf. Hornik et al. 1989). To realise the full epistemic potential of high-fidelity observability, a monist counterfactual theory of explanation is necessary, because it remains agnostic about the sources of counterfactual dependencies. Its superiority stems not only from covering causal as well as non-causal explanations (Reutlinger 2018, p. 78) but also from its potential to cover human as well as machine sources of observability counterfactuals. Although explananda remain counterfactually connected to explanans, the rigidity of the underlying dependency stays low because it accommodates a multitude of principles. Such a plurality might help to explore the idea of epistemically non-uniform approaches, i.e., *syntheses* between cognitive devices and agents such as the proposed one, which do not inherit their justification from the reference frames' rigidity. Exemplified by CTE, the monist approach to counterfactual

dependence thus provides a fitting theory of explanation able to utilise the full epistemic potential of reference frames whose rigidity decreases with the introduction of high-fidelity observability.

3.1 Inducing Competitive Coevolution Among Artificial Reference Frames to Achieve Generalisations

Competitive coevolution originates from considering the evolutionary change an 'arms race' between attacker and defender lineages, iteratively developing increasingly complex strategies of adversarial interactions (cf. Dawkins and Krebs 1979, here referring to the asymmetric interspecific version of the arms race metaphor). Apart from the pressures of inanimate environments, the lineages coevolve due to the coupled fitness which drives the 'arms race' forward and causes extinctions of non-adapting lineages (cf. 'Red Queen Hypothesis', Van Valen 1973). For participating individuals, the underlying zero-sum game provides a shared learning framework which facilitates a cross generations arms race and controls natural selection among the competing lineages (cf. Dawkins and Krebs 1979, p. 495).

The original experiments with artificial learning by competitive coevolution concerned games, usually a version of Prisoner's Dilemma (Darwen and Yao 1995; Yao and Darwen 1995; Yao 1997). It was established that the population arms races prove to be successful in dynamic settings, where the coupled fitness provides a principled guidance in the search for solutions, i.e., successful game strategies, which would be otherwise hard to secure given the difficulty or impossibility to devise an absolute measure (Chong et al. 2012, p. 70). Artificial reference frames comprising the competing populations usually represent participants of *n*-player Iterated Prisoner's Dilemma games (cf. Chong et al. 2008, 2009,

2012). In such cases, generalisations obtained by coevolutionary learning correspond to the strategies that outperformed the highest number of test strategies which were not encountered during the evolution (ibid.). By assuming the process of strategy acquisition a kind of machine learning, i.e., competitive coevolution as the players' training and a confrontation with unseen strategies as the players' testing (ibid.), it is possible to posit that competitive coevolution facilitates learning which leads to generalisations (cf. ibid.). In these experiments, the reference frames remain mere 'wire frames' (their expressivity is limited to representing strategies of the symbolic game play). As a result, they do not provide any insight into artificial representation learning, besides showing that population 'arms races' open a prospective avenue for attaining generalisations.

The avenue provides an opportunity to connect generative machine learning models with arms races among populations of artificial reference frames. The former can be represented by a generative adversarial model, comprising two competing artificial neural networks labelled according to their role 'generator' and 'discriminator' (Goodfellow et al. 2014). To learn generalisations, which allow production of unobserved or unconceived samples, the generator enters a game of deception against the discriminator, which, having access to the observational evidence, provides the learning signal that facilitates an approximation of the target distribution. The competition, unfolding between the generator and discriminator network, is similar to the arms race interpretation of the relationship between cuckoos and their hosts, especially considering mimetic eggs (cf. Dawkins and Krebs 1979: 61-64). Similar to a cuckoo, attempting to lay sufficiently mimetic eggs, the generator network attempts to transform an input (a vector of noise) into a phenomenal sample that would be considered as originating from the probability distribution underlying the observational evidence (cf. Goodfellow et al. 2014). As with the cuckoo egg met by a scrutinising host, the discriminator

network processes the generator's sample to predict whether it comes from the observational evidence or was sampled by the generator (cf. ibid.). The competition unfolds in steps. In every iteration, the discriminator is exposed not only to the generated samples but also to further observational evidence to ensure that the learning signal provided to the generator evolves over time (i.e., the discriminator's feedback regarding the presumed authenticity of the samples). The competition halts when the discriminator can no longer discern generated samples from the observational evidence. Reaching such an equilibrium indicates that the generator achieved a generalisation which allows interpolation from observational evidence to unobserved or unconceived phenomenal samples, presumably coming from the same probability distribution that underlies the empirical domain. In this sense, the generator, at least partially, succeeded in recovering the evidence generating distribution, since it managed to minimise the divergence of its approximation of the target distribution (ibid.). Insomuch as the generator codevelops with the discriminator, both sides would benefit from competitions between populations rather than individuals.

Thus far, two approaches to competitive coevolution of generative adversarial networks were proposed. The first one pits an individual discriminator, assuming the function of an environment, against a population of generators seeking to adapt to it (Wang et al. 2018, i.e., seeking to evade its pattern recognition countermeasures). Each individual (generator) from the population reflects a distinct measure of divergence between the approximated and evidence generating distributions (ibid.). During each evolutionary round, the generators' fitness derives from the integrity *and* diversity of their samples measured by a relative strength of the discriminator's present countermeasures (ibid.). The fittest generator then provides a foundation for a progeny that develops the lineage further, until

reaching an equilibrium where the discriminator can no longer separate the observational evidence from the generated phenomenal samples (ibid.).

The second approach proposes a tournament designed to rank individuals from a population of generators according to their performance against a population of discriminators (Olsson et al. 2018). Two versions of the tournament are considered – first, where the two populations consist of developmental snapshots of the generator and discriminator pertaining to a single model (i.e., the snapshots correspond to checkpoints from the learning trajectory of a single model), with the tournament players thus effectively engaging either past or future states of their adversary (ibid.). The second version of the tournament involves two populations of generators and discriminators supplied from a set of different models diverging in their initial settings, hyperparameters controlling the co-development, or architectures of the underlying artificial neural networks (ibid.). The players thus engage adversaries that are encountered for the first time, since the generators entering the tournament matches co-developed with different discriminators than they get to face during the competition. From the evolutionary 'arms race' point of view, the tournament results provide information about a relative fitness of the models' constitution and parameter settings which control the generator/discriminator co-development and influence their generalisation capability. Holding the tournament repeatedly then simulates a kind of selection process. The competition results provide directions on how to evolve the models' architectures and properties to gradually achieve a better generalisation performance in the upcoming populations, either within the context of a single or several learning trajectories.

Introducing out of reference frame content to sustain high-fidelity observability is *recursive*. As we utilise the generator's interpolation to acquire content for conceiving further

observability counterfactuals, the generator achieves its interpolation-enabling generalisation by engaging an out of reference frame learning signal provided by the discriminator. In these epistemic circumstances, some of the generalisations underpinning our reference frame come from an artificial reference frame (a generator), itself underpinned by an out of reference frame generalisation providing the learning signal (a discriminator). While in theory the recursion might reach arbitrary depths, competitive coevolution provides a principled guidance for the ongoing selection of the best candidates for reference frames that could provide outside generalisations. Unfolding in a recursive manner, every step comprises an import instead of self-reference. With every iteration, some generalisations stem from outside of the manifest frame of reference. The trajectory of a generalisation can be unpacked into a branching tree, with each node stabilised by reaching the following equilibrium. The moment the discriminator begins consistently identifying samples produced by the generator as observational evidence, *albeit they represent hitherto unobserved or unconceived states of affairs*, the level of convergence between the approximated and evidence generating distributions permits us to import the samples as content for conceiving further observability counterfactuals. The guarantee that each iteration reaches a near-optimal equilibrium is provided by competitive coevolution, selecting the fittest generator and discriminator in terms of their generalisation performance at the downstream task.

Recursive imports from outside of the manifest frame of reference sustain high-fidelity observability. It does not suffer from the same degree of underdetermination by available evidence as observability derived from rigid frames of reference. The case for synthesis between human and machine learning leads to decreases of the reference frames' rigidity and thereby to a progress on the issue of underdetermination by available evidence. The

synthesis impacts observability regardless of whether it is considered a non-modal objective property or a result of latching onto the objective features of the world. The more observability counterfactuals we have at our disposal the less underdetermined by available evidence our reference frame becomes. *Growing the zone of observability* by a synthesis between human and machine learning evolves our generalisations as well as explanations constructed on them.

4. Growing the Zone of Observability: Forays into the Twilight Zone

By way of high-fidelity observability, we aim to grow the zone of observability of our reference frame. This should increase the likelihood of at least a partial correspondence between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, states of affairs. The notion of zones draws on the relation between conceivability and possibility (Chalmers 2002). The question is what we can learn about possibility by relying on positive conceivability. The notion of positive conceivability reflects the unidirectional epistemic disposition of conceivability providing generalisations. The unidirectionality then leads the conceived observability counterfactuals to a precarious state where we cannot rule out that observable, yet so far unobserved or unconceived, phenomena will undermine the generalisations underpinning our frame of reference. The unidirectionality, and the rigidity of reference frames which it produces, underdetermines observability in a similar way as Chalmers's negative conceivability, concerning phenomena that cannot be conceived positively, however, we fail to rule them out on an a priori basis (2002, pp. 149-50). Insomuch as we assume X negatively conceivable, i.e., we are supposedly unable to *rule it out or conceive*, X might be just the case of an observable, yet so far *unobserved or unconceived*, state of affairs (respectively). An underdetermined reference

frame confounds two epistemic conditions. First, the genuine negative conceivability and, second, yet to be revealed discrepancies between conceived observability counterfactuals and content from outside of the reference frame. Facing such epistemic difficulties, high-fidelity observability offers a principled guidance. The synthesis between human and machine learning provides a way to distinguish mere discrepancies from the cases of negative conceivability that concerns genuine inconceivabilities.

Indicating their epistemic unavailability, Chalmers counts inconceivabilities among the twilight zone inhabitants (2002, pp. 186-88). Importing some of them as fresh content from outside of the reference frame incorporates accessible regions of the twilight zone into the observability zone. The bidirectional flow of content between generalisations and conceivability facilitates converting some of the twilight zone's regions into observability counterfactuals. These would in turn grow the observability zone of our reference frame. As a result, high-fidelity observability could expand the range of positive conceivability beyond the point reachable by human cognition and unidirectionality. Entering the twilight zone provides an opportunity to reassess generalisations underpinning our frame of reference and shed a new light on the relation between unobserved/unconceived and unobservable/inconceivable states of affairs. In some cases, it would be possible to distinguish the latter from the former by sampling the unobserved/unconceived states of affairs (i.e., phenomenal samples, see Spelda and Stritecky 2021 for applications in some scientific fields) from the generator's latent space. Underpinned by the generalisation capability learnt during competitive coevolution (i.e., arms races between generator and discriminator populations), the generator's latent space captures some of the twilight zone's regions. If we consider out of reference frame content as a learning signal, the further

observability counterfactuals conceived by humans interpolate beyond observational evidence and provide high-fidelity observability without making the reference frame rigid.

Every reference frame is surrounded by a twilight zone, which renders the reference frame's epistemic horizon dim and the correspondence between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena uncertain. This uncertainty is usually addressed by making reference frames rigid, shaping observability either into a non-modal, objective property or into an outcome of using modality to determine the objective features of the world. In any case, the horizon remains dusky and the twilight zone's unobserved/unconceived and unobservable/inconceivable inhabitants continue to undermine observability.

James Ladyman notes *"The predicate "unobservable" does not feature in scientific theories about unobservable things, so if we cannot deduce observability counterfactuals that are false from them, we cannot ever use science to tell us whether something is objectively unobservable or just as yet unobserved."* (2004, p. 762). However, by considering out of reference frame content, the twilight zone becomes less obscure. Therefore, we might meaningfully speak about conceiving further observability counterfactuals based on observable, yet so far unobserved or unconceived, phenomenal content imported from outside. By embracing Ladyman's worry, while *considering the anthropic perspective extended by machine learning*, we are in fact deducing observability counterfactuals for yet unobserved or unconceived phenomena, moving the edge of the zone of observability forward, and thus solving a part of the dilemma of whether we are facing '*objectively unobservable or just yet unobserved*' states of affairs.

4.1 A Note on Observability in Scientific Realism and Constructive Empiricism

Related to our discussion on observability is a dispute over observability between some scientific realists and constructive empiricists (cf. Ladyman 2000; Monton and van Fraassen 2003; Ladyman 2004; Muller 2005; Dicken 2009, pp. 194-97). Constructive empiricism treats observability as a non-modal objective property to avoid inflationary metaphysics (Monton and van Fraassen 2003). Some scientific realists argue against this kind of observability. The epistemic trouble of 'X is not ever actually observed' nonetheless considered objectively observable despite the lack of objective truth conditions for the observability counterfactual (Ladyman 2004, pp. 761-63) is at the centre of the dispute.

The example of measuring properties of gun flashes provided by Monton and van Fraassen shows that observability can be derived from phenomena observed during experimentation and from the generalities agents have about themselves and their environment (2003, pp. 410-11, 413). Invariances among the observed phenomena enable the agents to generalise, *fix* their reference frame, and sustain the objective nature of observability. Although observability counterfactuals depending on a reference frame lack objective truth conditions (there are many possible reference frames entailing different observability counterfactuals), fixing the reference frame objectifies observability for a given epistemic community (cf. Monton and van Fraassen 2003, p. 411). Some realists responded, referring to problems with: (1) fixing the reference frame to fit an epistemic community, (2) an arbitrary selection of background generalities, (3) the situation where a phenomenon X is never present to anyone while considered observable (cf. Ladyman 2004, pp. 760-61). This situation problematises the objectivity of observability based on the correspondence between an observation and a conceived observability counterfactual (ibid.).

Even before Ladyman voiced his concerns, various proposals attempted to solve the problems 1-3 listed above. For example, structural empiricism applied the idea of simplicity ordering (cf. Putnam 1975, pp. 301-2) to create degrees of empirical adequacy (Bueno 1997). The degrees correspond to partial models from a hierarchy capturing gradual increments in information acquired about the phenomena, where the hierarchy constitutes the overall structure considered by constructive empiricists (ibid.). Direct responses to Ladyman used modal agnosticism to remain neutral and avoid commitments to modal statements that involve a distinction between observable and unobservable (cf. Muller 2005; Dicken 2007).

We think that if the relation between conceivability and generalisations becomes bidirectional thanks to fresh content from outside of our reference frame, then the third problem posed by Ladyman can be made less pressing. The third problem lies in the fact that constructive empiricists cannot be sure what would happen had an actual observation of X been made as they do not admit invariances that latch onto the objective features of the world (cf. Ladyman 2004, p. 762). Supposedly, empiricists may then lack objective truth conditions for the corresponding observability counterfactual and cannot maintain the objective nature of observability (ibid.). We expect that using fresh content from generative machine learning models to conceive further observability counterfactuals increases the likelihood that a more robust agreement could be negotiated between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena. Interestingly, fresh content from outside of their reference frame may help some scientific realists as well, for example, while defending realist theories against unconceived alternatives (Spelda and Stritecky 2021). It seems that if one is not opposed to relaxing the rigidity of their reference frame, observability can be turned into high-fidelity observability and the bidirectionality between conceivability and generalisations could make

observability counterfactuals more robust. A separate issue is the epistemological study of machine learning enhancers that should delimit conditions under which humans are justified to use the models' outputs as fresh content for conceiving further observability counterfactuals (cf. Leonelli 2020). This question is outside the scope of the paper. It should be also noted that the use of generative machine learning models suggested here, and already practiced in some scientific fields (cf. Spelda and Stritecky 2021), builds on earlier investigations of computational science (e.g., Humphreys 2004).

5. Conclusion

The synthesis between human and machine learning could help to negotiate a more robust relation between conceived observability counterfactuals and observable, yet so far unobserved or unconceived, phenomena. Using machine learning proxies to explore the twilight zone decreases the rigidity of our reference frame and expands its epistemic horizon. This transforms observability into high-fidelity observability and provides us with an opportunity to review the core generalisations that underpin our reference frame and particular explanations. High-fidelity observability remains agnostic in the monist sense. It does not prefer one source of generalizations over other because the relation between conceivability and generalisations becomes bidirectional.

This output was supported by the NPO 'Systemic Risk Institute' no. [LX22NPO5101] funded by European Union – Next Generation EU (Ministry of Education, Youth and Sports, NPO: EXCELES).

References

Bueno, O. (1997). Empirical adequacy: A partial structures approach. *Studies in History and Philosophy of Science Part A, 28*(4), 585-610.

Campbell, D., Copeland, J., Deng, Z-R. (2017). The Inconceivable Popularity of Conceivability Arguments. *The Philosophical Quarterly 67*(267), 223-240.

Chakravartty, A. (2017). Reflections on new thinking about scientific realism. *Synthese, 194*(9), 3379-3392.

Chalmers, D., J. (2002). Does conceivability entail possibility? In T. S. Gendler & J. Hawthorne (Eds.), *Conceivability and Possibility* (pp. 145-200). New York: Oxford University Press.

Chong, S., Y., Tino, P., & Yao, X. (2008). Measuring Generalization Performance in Coevolutionary Learning. *IEEE Transactions on Evolutionary Computation 12*(4), 479-505.

Chong, S., Y., Tino, P., & Yao, X. (2009). Relationship Between Generalization and Diversity in Coevolutionary Learning. *IEEE Transactions on Computational Intelligence and AI in Games 1*(3), 214-232.

Chong, S., Y., Tino, P., Ku, D., C., & Yao, X. (2012). Improving Generalization Performance in Co-Evolutionary Learning. *IEEE Transactions on Evolutionary Computation 16*(1), 70-85.

Darwen, P., J., & Yao, X. (1995). On Evolving Robust Strategies for Iterated Prisoner's Dilemma. In X. Yao (Ed.), *Progress in Evolutionary Computation. AI '93 and AI '94 Workshops on Evolutionary Computation Melbourne, Victoria, Australia, November*

16, 1993 Armidale, NSW, Australia, November 21–22, 1994 Selected Papers. Berlin: Springer.

Dawkins, R., & Krebs, J., R. (1979). Arms races between and within species. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 205*(1161), 489-511.

Dicken, P. (2007). Constructive Empiricism and the Metaphysics of Modality. *The British Journal for the Philosophy of Science, 58*(3), 605–612.

Dicken, P. (2009). Constructive Empiricism and the Vices of Voluntarism. *International Journal of Philosophical Studies, 17*, 189-201.

Fiocco, M., O. (2020). The epistemic idleness of conceivability. In O. Bueno, S. A. Shalkowski (Eds.), *The Routledge Handbook of Modality.* London: Routledge.

Gendler, T., S., Hawthorne, J. (2002). *Conceivability and Possibility*. New York: Oxford University Press.

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*. Cambridge, MA: The MIT Press.

Hill, C., S., (1997). Imaginability, Conceivability, Possibility and the Mind-Body Problem. *Philosophical Studies 87*(1), 61-85.

Hitchcock, C., & Woodward, J. (2003). Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Noûs 37*(2), 181-199.

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

Humphreys, P. (2004). *Extending ourselves: computational science, empiricism, and scientific method*. New York, NY: Oxford University Press.

Kung, P. (2010). Imagining as a Guide to Possibility. *Philosophy and Phenomenological Research 81*(3), 620-663.

Ladyman, J. (2000). What's Really Wrong with Constructive Empiricism? Van Fraassen and the Metaphysics of Modality. *The British Journal for the Philosophy of Science 51*(4), 837-856.

Ladyman, J. (2004). Constructive Empiricism and Modal Metaphysics: A Reply to Monton and van Fraassen. *The British Journal for the Philosophy of Science, 55*(4), 755–765.

Leonelli, S (2020). Scientific Research and Big Data. *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*. In E. N. Zalta (ed.) https://plato.stanford.edu/archives/sum2020/entries/science-big-data/.

Monton, B., & van Fraassen, B. C., Constructive Empiricism and Modal Nominalism. *The British Journal for the Philosophy of Science, 54*(3), 405–422.

Muller, F. A. (2005). The Deep Black Sea: Observability and Modality Afloat. *The British Journal for the Philosophy of Science, 56*(1), 61–99.

Olsson, C., Bhupatiraju, S., Brown, T., Odena, A., & Goodfellow, I. (2018). Skill Rating for Generative Models, arXiv:1808.04888v1 [stat.ML].

Putnam, H. (1975). *Mathematics, Matter and Method (Philosophical Papers, Vol. 1)*. Cambridge: Cambridge University Press.

Rescher, N. (2020). *Knowledge at the Boundaries*. Cham: Springer.

Reutlinger, A. (2016). Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science 83*(5), 733-745.

Reutlinger, A. (2018). Extending the Counterfactual Theory of Explanation. In A. Reutlinger, J. Saatsi (Eds.), *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations* (pp. 74-95). New York: Oxford University Press.

Saatsi, J., & Pexton, M. (2012). Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations. *Philosophy of Science 80*(5), 613-624.

Spelda, P., Stritecky, V. (2021). What Can Artificial Intelligence Do for Scientific Realism? *Axiomathes 31*, 85-104.

Tidman, P. (1994). Conceivability as a Test for Possibility. *American Philosophical Quarterly 31*(4), 297-309.

Van Valen, L. (1973). A New Evolutionary Law. *Evolutionary Theory, 1*, 1-30.

Wang, C., Xu, C., Yao, X., & Tao, D. (2018). Evolutionary Generative Adversarial Networks, arXiv:1803.00657v1 [cs.LG].

Wilks, Y. (2017). Will There Be Superintelligence and Would It Hate Us? *AI Magazine, 38*(4), 65-70.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science 51*(2), 197-254.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2018). Some Varieties of Non-Causal Explanation. In A. Reutlinger, J. Saatsi (Eds.), *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations* (pp. 117-140). New York: Oxford University Press.

Yablo, S. (1993). Is Conceivability a Guide to Possibility? *Philosophy and Phenomenological Research 53*(1), 1-42.

Yao, X., & Darwen, P., J. (1995). An Experimental Study of N-Person Iterated Prisoner's Dilemma Games. In X. Yao (Ed.) *Progress in Evolutionary Computation. AI '93 and AI '94 Workshops on Evolutionary Computation Melbourne, Victoria, Australia, November 16, 1993 Armidale, NSW, Australia, November 21–22, 1994 Selected Papers.* Berlin: Springer.

Yao, X. (1997). Automatic Acquisition of Strategies by Co-evolutionary Learning. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA '97)*.