University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

[Faculty Scholarship](#)

9-2021

Impossibility Results in AI: A Survey

Mario Brcic
University of Zagreb

Roman Yampolskiy
University of Louisville, roman.yampolskiy@louisville.edu

Impossibility Results in AI: A Survey

Mario Brcic[1], Roman Yampolskiy[2]
[1]University of Zagreb Faculty of Electrical Engineering and Computing
[2]University of Louisville
mario.brcic@fer.hr, roman.yampolskiy@louisville.edu

Abstract

An impossibility theorem demonstrates that a particular problem or set of problems cannot be solved as described in the claim. Such theorems put limits on what is possible to do concerning artificial intelligence, especially the super-intelligent one. As such, these results serve as guidelines, reminders, and warnings to AI safety, AI policy, and governance researchers. These might enable solutions to some long-standing questions in the form of formalizing theories in the framework of constraint satisfaction without committing to one option. In this paper, we have categorized impossibility theorems applicable to the domain of AI into five categories: deduction, indistinguishability, induction, tradeoffs, and intractability. We found that certain theorems are too specific or have implicit assumptions that limit application. Also, we added a new result (theorem) about *the unfairness of explainability*, the first explainability-related result in the induction category. We concluded that deductive impossibilities deny 100%-guarantees for security. In the end, we give some ideas that hold potential in explainability, controllability, value alignment, ethics, and group decision-making. They can be deepened by further investigation.

> *"This, then, is the ultimate paradox of thought:*
> *to want to discover something that thought itself cannot think."*
> S.Kierkegaard

1   Introduction

An impossibility theorem demonstrates that a particular problem cannot be solved as described in the claim, or that a particular set of problems cannot be solved in general. The most well-known general examples are Gödel's Incompleteness theorems [1] and Turing's undecidability results [2] in logic and computability, as well as Fermat's Last Theorem in number theory. The similar, and connected, is the notion of no-go theorems that state the physical impossibility of a particular situation. These results, though in themselves do not point to the solutions, are useful in the sense that they guide the direction for future efforts in AI in general, but in our case, the interest is in AI safety and security. In physics, there is an idea of restating the whole field in the terms of counterfactuals and what is possible and impossible in the system [3]. The authors think this will enable solutions to some long-standing questions in the form of formalizing theories in the framework of constraint satisfaction without committing to one option. A similar view regarding the utilization of the constraint satisfaction approach to many questions in philosophy is expressed by Wolpert [4]. Moreover, automated proof [5] and search [6] procedures for impossibility theorems based on constraint satisfaction were already proposed in the domain of social choice theory.

First, we shall present a classification of all relevant impossibility results. In that classification, all the impossibility theorems are shown to be neatly subsumed under the problems with capacity disparity where several related objects differ in size. Then, we shall present the current impossibility results that we find relevant for AI. That includes work made specifically in AI, but also work in other fields such as mathematics, physics, economy, social choice theory, etc. In the process, we

shall show a new result - theorem that states the unfairness of explainability. It is the first induction impossibility theorem pertaining to explainability as all the previous ones were addressing the perspective of deduction.

Previous works cover similar topics within the scope of AI safety [7]–[10], but none focused on impossibility theorems as a family, their utility, structure, and connections to other fields.

The rest of the paper is organized as follows. In section 2 we introduce basic definitions. Section 3 contains the relevant work presented under newly defined categorization. In section 9 we focus on impossibility theorems developed in the field of AI safety. The discussion is offered in section 13, and ideas for future research are listed in section 14. Finally, we conclude the article in section 14.

2   Basic definitions

We shall not impose strict formalization, but we shall keep at the level of lawyerese in this paper to ensure wide readability of material. Other papers will cover more formalized arguments. It is evident that our investigation is done from the perspective of assumption that *intelligent behavior that can achieve its goals is computable*.

System is any non-empty part of the universe.

State is the condition of the universe.

Control of system A over system B means that A can influence system B to achieve A's desired subset of state space.

Usually, with control, we aim at output controllability (from control theory). Such control is not sufficient for safety – as we often make unsafe choices ourselves. Different modes of "influence" and "desire" are possible. With regards to that, Yampolskiy in [7] mentions four types of control: explicit (strict), implicit, aligned, and delegated. **Explicit control** agent takes expressed desires literally and acts on them. **Implicit control** agent uses common sense as a safety layer over explicit control to slightly reformulate the expressed desire and acts on it. **Aligned control** agent adds intention inference over implicit control in order to postulate the intended desire and acts on it. **Delegated control** agent decides for itself the subject's desire that is long-term-best for the subject and acts on it.

Intelligence is the ability for an information processing system to adapt to its environment with insufficient knowledge and resources [11].

Safety of system A is the property of avoidance of going out of A's desired subset of the state space.

Safety is pressed with finding the worst-case guarantees – which is modeled as adversarial games that assume the ideal adversary.

Stability of state S for system A is the intrinsic tendency to return to A's desired subset of state-space after being perturbed.

Robustness of state S for system A is the property of staying within A's desired subset of the state space despite perturbations.

Catastrophic outcome for system A is any state from which the return to A's desired subset of state space is impossible.

Alignment *within the ensemble of systems $A_1...A_n$ is the property that each system A_i achieves greater than or equal benefit from working together than if any subset of agents acting self-interestedly.*

Alignment is about finding values that would make the game cooperative in a long term. There are plenty of open questions regarding the topic of alignment. How to achieve cooperativity in game over long periods? Is the game allowed to temporarily deviate from perfect cooperation? The problem for humans is that we do not have consistent short- and long-term values. Sometimes we have to suffer in short term (like in sports) in order to prosper in the long term. How to define alignment with such preferences?

3 Impossibility theorems

Impossibility theorems boil down to some contradiction. The potential to find impossibility theorems lurks at the appearance of paradoxes. Paradoxes are simple implications that there is some constraint, limit, we were not aware of when we unknowingly reached out of the feasible area and reaching some contradiction in our stated goals. Finding impossibility theorems circumscribes our knowledge of possible which enables us to direct our efforts better and do better risk management. These can help even when we do not know the status of final solutions, such as for AGI and AI Safety, in a form of directing constraint satisfaction search – both in formalizing ideas and committing to certain hypotheses. It might be that this is the most prudent approach for approaching hard problems with long horizons to finding solutions. It works on a meta-level of scientific investigation by upending the way research is done, in the process creating results about the problem even when oblivious about the actual solution.

3.1 Capacity

Impossibility theorems pertinent to AI are mostly related **to the problem of capacity disparity**. Namely, intuitively we operate between the domains that differ in size. For example, when expanding from intrinsically smaller to a greater domain, or in the opposite direction when contracting from bigger to the smaller domain. We have organized all the impossibility theorems into 5 subcategories (see Figure 1):
1. Deduction (D),
2. Induction (I),
3. Indistinguishability (N),
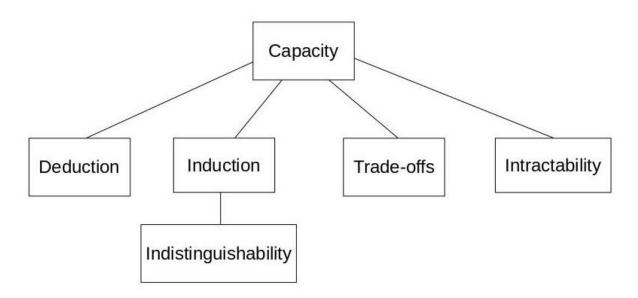4. Tradeoffs (T), and
5. Intractability (P).

Figure 1. Proposed categorization of impossibility theorems in AI

For example, **deduction** tries to go beyond its size capacity by going from countable to uncountable infinite (Turing's computability), in Gödel's terms by going from provable to true statements, or in Chaitin's terms from lower to greater Kolmogorov complexity of formal systems. Self-referential paradoxes go beyond these capacity limits by including itself in its own definition and negating itself. From that follows an infinite fractal-like growth where we used finite means to express infinite without a fixed-point. Examples include unverifiability. The proofs in this category often use Lawvere's theorem [12], [13] in the disguise of Cantor's diagonalization and liar's paradox.

In terms of **induction**, we have to find a model within a (possibly infinite) set of plausible models based on finite dataset/experience. Hence, we have the finite capacity of experience to guide our search within a set having multitude (possibly infinitude) of elements. There are too many inductive inferences that can be made. Induction as a general operation is prone to the problems emanating from Hume's problem of induction and Goodman's new riddle of induction [14]. No Free Lunch theorems [15], [16] deal with induction and are a formalization of Goodman's new riddle of induction [17]. As such, they can be the basis of a vast number of induction-based impossibilities. Examples include unpredictability.

Also, in special cases of induction, we have a problem of disentangling, ie. indistinguishability (**non-identifiability** and **unobservability**). We wish to get the inner structure from the limited entangled data. No amount of data can enable identification in the cases of non-injective transformations that produced the data. This means the impossibility of learning even in the limit. There is an inevitable **loss of information** going from the origin through the (capacity reducing) transformation. That prohibits the recovery of full information and leads to observational equivalence. Examples include unobservability in control theory [18].

Trade-offs are inherent to multicriteria decision making, or "you can't have it all". The problem of size capacity is evident here in the inability to obtain the point with the individual maximal value of each component at the Pareto front (and hence full hypervolume indicator [19]).

In **intractability,** we have physical limits on capacity (in memory, computing power,…) which prohibits efficiently reaching solutions. This is not only a set-theoretical, computation-independent limit, but it also enables relative comparisons, for example through Karp's hierarchy [20].

We have listed in Table 1, below, impossibility theorems we deemed important for for field of AI. There are many more impossibility theorems in mathematics and physics, but they are not (to us) evidently related to AI. There are also many impossibility theorems in machine learning, but we omitted them due to the too-narrow scope.

Table 1. Impossibility theorems of interest to AI researchers.

Name	Source	Proven	Category (D;I;N;T;P)
Unobservability	[18]	Y	N
Uncontrollability of dynamical systems	[18], [21]	Y	N
Good Regulator Theorem	[22]	Y	N
Law of Requisite Variety	[23]	Y, under perfect information and infinite speed	N
Information-theoretical control limits	[24]	Y	N
(Anti)codifiability thesis	[25]–[27]	N	I
Arrows impossibility theorem	[28]	Y	T
Impossibility theorems in population ethics	[29]	Y	T
Impossibility theorems in AI alignment	[30]	Y	T
Fairness impossibility theorem	[31], [32]	Y	T
Limits on preference deduction	[33]	Y	N
Rice's Theorem	[34]	Y	D
Unprovability	[1]	Y	D
Undecidability	[2], [35]	Y	D
Chaitin Incompleteness	[36]	Y	D
Undefinability	[37]	Y	D
Unsurveyability	[38]	N	P
Unlearnability	[39], [40] [41]	Y	D P
Unpredictability of rational agents	[42], [43]	Y	I
No Free Lunch - supervised learning	[15]	Y	I
No Free Lunch -	[16]	Y	I

optimization			
Free Lunch in continuous spaces and coevolutionary	[44], [45]	Y	I
Unidentifiability	[46]–[49]	Y	N
Physical limits on inference	[4], [50], [51]	Y	D/I
Uncontainability	[52]	Y	D
Uninterruptibility	[53]–[57]	N, only under limited assumptions - opened	D
Löb's Theorem (unverifiability)	[58]	Y	D
Unpredictability of superhuman AI	[59], [60]	Y~, definition of superhuman?	I
Unexplainability	[61]	N, based on explanation=proof & (unprovability | undefinability), assuming honesty or full model	D
Incomprehensibility	[61], [62]	N, based on explanation=proof & (unverifiability | unsurveyability | undefinability), assuming honesty or full model Y, comprehension=producing proof & halting games	D
k-incomprehensibility	[63]	N, just definitions	I
Unverifiability	[64]	Y	D
Unverifiability of robot ethics	[65]	Y	D
Intractability of bottom-up ethics	[66], [67]	Y	P
Goodheart's Law (Strathern)	[68], [69]	N	I
Campbell's law	[70]	N	I
Reward corruption unsolvability	[71]	Y	I
Uncontrollability of AI	[7]	Y~, under degenerate conditions	D
Impossibility of unambigous communication	[72]	Y, under strict assumptions	I
Unfairness of explanations	here	Y~, proof sketch	I

3.2 Deduction Impossibility Theorems

Limits of deduction are limits on our capability to achieve perfect certainty in facts. The vast majority of listed results here use Lawvere's fixed-point theorem [12], [13] as the basis of proofs, i.e. more specifically a combination of Cantor's diagonalization and liar's paradox.

The most basic results here are Gödel's incompleteness theorems [1] addressing unprovability and Turing's work [2] covering undecidability. In addition to the aforementioned which cover the processing, Gregory Chaitin provided additional incompleteness results [36] that cover input sizes measured by Kolmogorov's complexity. Chaitin's incompleteness theorem states the existence of a limit on any formal system to prove Kolmogorov complexity of strings beyond some length. There are even conjectures that information complexity might be the source of incompleteness [36], [73] whereby *the theorems of finitely stated theories cannot be significantly more complex than the theory itself.*

Rice's theorem [34] is a generalization of Turing's undecidability of the halting property of programs to any sufficiently complex property. This makes it an ideal tool for finding and proving deduction limitations in AI, within its assumptions. Löb's theorem [58], informally put, states that a formal consistent system cannot, in general, prove its own soundness.

Regarding formal semantics, Tarski's undefinability theorem [37] states that truth in a formal system cannot be defined within that system. Measuring semantic information yield by deduction is also poised by paradoxes. Bar-Hilel-Carnap paradox [74] in classical semantic information theory entails that contradiction conveys maximal information. Hintikka's scandal of deduction [75] points to the fact that the information yield of truthful sentences is zero since their information is already contained in the premises.

There is vast work in impossibility theorems in beliefs which is an extension of Gödel's work. Huynh and Szentes [76] have demonstrated irreconcilability between two notions of self-belief. In [77], [78] the paradox of self-reference was extended to the games with 2 or more players which yield impossible beliefs.

Inference devices covered in the series of papers by Wolpert [4], [50], [51] are an extension from pure deductive systems to the general notion of inference devices which covers deduction, induction (prediction, retrodiction), observation, and control while the device itself is embedded in a physical universe. Additional limitations are found, both in a logical and stochastic sense. For example, limits are found on strong and weak inference, control, self-control, and mutual control between the two distinguishable devices. The limits are also put on prediction, retrodiction, observation, and knowledge.

3.3 Indistinguishability Impossibility Theorems

Observability is the ability to infer the state of a black-box system from its input/output data. **Identifiability** (ie. parameter and/or structural identifiability) is a special case of observability for constant elements of the system whereby we only need to infer the values of those constant elements. There are limits to both observability and identifiability, and the limits are caused by non-injective mappings [49] which inevitably lose information.

Identifiability and observability are important for the control over systems.
Controllability in control theory can be of two kinds: state and output. **State controllability** is the ability to control the inner state of the system. **Output controllability** is the ability to control the output of the system. State uncontrollability is the dual of unobservability. That is, state controllability is impossible without observability [18], [21].

Good regulator theorem [22] relates output controllability and identifiability (modeling), but only in a sense of optimal control, not sufficient control. It states that maximally simple among *optimal* regulators must behave as an image of the controlled system under a homomorphism. Sufficiently good regulators need not be optimal and the generalization of such theorem would be interesting.

Law of requisite variety [23] states that variety in outputs can only be reduced by the state complexity of the controlling mechanism. **Information theory state-control limit** [24] says that only up to information observed from the system can be used to reduce the entropy of the system.

In the absence of additional biases, general nonlinear independent component analysis has an infinity of solutions that are indistinguishable [46]. Similar is shown for unsupervised learning of disentangled representations [48]. There are also well-known non-identifiability limits to causal discovery from the data [47].

3.4 Induction Impossibility Theorems

Limits on induction constrain our ability to infer latent factors. Here we will ignore the problem of indistinguishability by just looking at equivalence classes of models indistinguishable in the limit. In this case, there is a possibility to learn the true equivalence class asymptotically. But, given some prefix of experience, there may be a multitude of candidate classes. This is pointed out in Hume's problem of induction and Goodman's new riddle of induction [14]. **No free lunch theorems (NFL)** by Wolpert [15], [16] are the basic building blocks underlying the formalization of these limitations. They were first formulated in general supervised learning and optimization, which were subsequently unified through that framework. No free lunch theorems state that under uniform distribution over induction problems, all induction algorithms perform equally [79]. At the heart of NFL formalization is the independence of (search/learning) algorithm performance from the uncertain knowledge of the true problem at hand. That independence is materialized in the inner product formula of those two in describing the probability of attaining a performance value over the unknown problem. There are, however, free lunches if more structure is imposed on the problem, i.e. "there is no learning without bias, there is no learning without knowledge" [80]. For example, there are free lunches in continuous spaces [44] and in coevolutionary problems [45].

Goodhart-Strathern's law [68], [69] and Campbell's law [70] deal with the difficulties and the inability in finding expressible proxy numerical measures for success that are **well aligned** with inexpressible/unknown-explicitly experiential measure of success. A similar sentiment is expressed in [81] where a more detailed explanation is given for the observed difficulties, all stemming from the unpredictability of solution routes to hard or even unknown problems. Metric is a model of an imagined success, but shallow and not with perfect alignment.

In games with uncertainty in opponent's payoffs, it is impossible to predict the behavior of perfectly rational agents due to the feedback loop emanating from their own decisions which influence opponent's behavior [42]. Placing further restrictions on the assumptions can regain predictability. In economic situations, further limits relating to rationality, predictability and control were proved in [43]. Therein, (i) logical limits were set to forecasting the future, (ii) non-convergence of Bayesian forecasting in infinite-dimensional space was demonstrated, and (iii) impossibility of computer perfectly forecasting economy if agents know its forecasting program. These results are related to the already mentioned results in deduction-related ITs for Wolpert's inference devices and regarding beliefs in games.

Anticodifiability thesis [25]–[27] is a conjecture in moral philosophy that states that universal morality cannot be codified in a way that would be aligned in all circumstances with our inexpressible/unknown-explicitly experiential moral intuition.

3.5   Tradeoffs Impossibility Theorems

Trade-off limits constrain our attempts to achieve perfect outcomes. Examples include impossibility theorems in clustering [82], fairness [31], [32], and social choice theory (SCT) [28]. In many situations, we have to choose with respect to multiple criteria simultaneously. Often, it is the case that there is no ideal point that simultaneously optimizes all the criteria, that is achieves maximal possible hypervolume indicator [19].

In social choice theory, there are results such as Arrow's impossibility theorem [28] which states there must be a trade-off that forces choosing only a strict subset of desirable properties in voting mechanisms. In moral theory, there are different problems regarding population ethics [29] where all total orderings entail some problematic properties that contradict our intuitions. Solutions have been proposed for automated systems that search for impossibility theorems in SCT regarding rankings of objects [6].

3.6   Intractability Impossibility Theorems

Intractability limits divide possibility-in-principle and practically impossible due to the resource limitations. There are three types of intractabilitY ITs: asymptotic, physics-based, and human-centered.

Asymptotic intractabilities fall neatly under the complexity theory [20], [83]. That research field is simply too rich to expand on it here. We shall only highlight the probably approximately correct (PAC) learning framework [41] by Valiant that defines the border between efficient (polynomial time) and inefficient learnability. Of our interest is also the intractability of bottom-up ethics [66], [67] which stems from the game-theoretic nature of ethics.

Physics-based limits put bounds on physically implementable computation and intelligence. No-go theorems state constraints for certain implementation approaches. Currently, these limits (e.g. [84]) are quite loose and/or specific so we do not go into their details. One exception is the work of Wolpert we have previously mentioned [4], [50], [51]. That research is quite general and is an extension of previous mathematical results by embedding computational agent into the universe within which it utilizes resources for computation.

There is an area of **human-centered limits** which does not seem to be well researched and measured. Humans, as agents of finite capabilities, have strict limits with regard to explainability, comprehensibility [10], and all other aspects. One of the commonly mentioned impossibilities is unsurveyability [38] in the context of mathematical proofs.

4   Impossibility theorems developed in AI safety

Uncontainability [52] states the inability of preventing superintelligence harming people if it chooses to, by recognizing the intent ahead of time. This is due to the undecidability of harmful properties in complex programs (corollary of Rice's theorem).

Unverifiability [64] states fundamental limitation (or inability) on verification of mathematical proofs, of computer software, of the behavior of intelligent agents, and of all formal systems. This is a corollary of Rice's theorem as well. An extension of Rice's theorem to robot programs was proven in [65] to show impossibility of online verification of robot's ethical and legal behavior.

Uninterruptibility [53]–[57] states that under certain conditions it is impossible to turn off (interrupt) intelligent agent. Possibilities and impossibilities have been shown under specific assumptions and conditions.

Unpredictability [59], [60] states our inability to *precisely and consistently* predict what specific actions an intelligent system will take to achieve its objectives, even if we know the terminal goals of the system. The proof depends on the implicit, but the unstated definition of unaligned superhuman intelligence and forms contradiction. The form of the proof does not limit occasional imperfect but sufficiently precise predictions. The question is, short of perfection, how much predictability is sufficient?

Unexplainability [61] states the impossibility of providing an explanation for certain decisions made by an intelligent system that is both 100% accurate and comprehensible. Here, the explanation is taken to be a proof which is then prone to the deduction ITs such as unprovability and undefinability. What is not covered is with respect to what is accuracy measured against and does not cover truthfulness of explanation in the case of incomplete information. Explaining yourself truthfully and correctly would imply self-comprehension which is a problematic notion itself as disproved in [62].

Incomprehensibility [61] states the impossibility of complete understanding of any 100% -accurate explanation for certain decisions of an intelligent system by any human. It is the dual of explainability and again it is assumed that explanation is proof which leads to the use of deductive ITs. Understanding is vaguely defined as proof-checking and it is not defined how accuracy is measured. In a similar line of work Charlesworth [62] defines comprehension of some systems as the capability of producing correct proofs by fallible agents about those systems. He takes a program as a starting point, implicitly assuming its truthfulness. He then produces relations of comprehensibility and rules out self-comprehension.

Uncontrollability [7] states that humanity cannot remain safely in control while benefiting from a superior form of intelligence. The proof uses a Gödel-like structure that shows the impossibility of perfect control in degenerate conditions which invoke self-referential paradoxes with controls. The form of control shown to be impossible was explicit control. In fact, with such proof, *uncontrollability holds for any sufficiently complex agent over which explicit control is attempted*, including humans. Moreover, *the proof holds also for the case of attempted self-control*. This counter-intuitive notion points into the direction that more research is necessary into formalization and disentangling of the structure and assumptions of explicit control. Advanced forms and notions of control should at least resolve the status of control over oneself. More research is necessary into the status of controllability for other forms of control (implicit, aligned, delegated).

Limits on utility-based value alignment [30] state a number of impossibility theorems on multi-agent alignment due to competing utilitarian objectives. This is not just an AI-related topic. The most famous example is Arrow's Impossibility Theorem from social choice theory, which shows there is no satisfactory way to compute society's preference ordering via an election in which members of society vote with their individual preference orderings.

Limits on preference deduction [33] state that even Occam's razor is insufficient to decompose observations of behavior into the preferences and planning algorithm. Assumptions, in addition to the data, are necessary for disambiguation between the preferences and planning algorithm. This is due to non-injective mapping induced by preferences and planning algorithm that produce behavior.

Unsolvability of reward corruption [71] states that without simplifying assumptions it is impossible to solve reward corruption problems such as wireheading, sensory error, reward

misspecification, and error in preference deduction. The proof is done via an NFL route and holds for reinforcement learning, for example. The problem can be averted under some simplifying assumptions and sufficient reward crosschecking. Otherwise, quantilisation [85] may provide more robustness.

Impossibility of unambiguous communication [72] denies perfectly unambiguous communication using natural language. Many examples are given to show different levels of ambiguity: phonology, syntax, semantics, and pragmatics along with contemporary NLP and AI approaches to handling them. These areas are taken together to show, under simplified assumptions, that ambiguity is inevitable in communication using natural language. Generalization of Goodhart's law for problem detection tools in AI systems is given, without proof. The intention and mechanism behind the proposed law seem to be adversarial learning.

The following subsection introduces a new impossibility result.

4.1 Unfairness of explainability

Let us examine the explainability process which consists of decision making, explanation generation, and verifying decision through the explanation. Let us assume that verification decides if the decision will be accepted (i.e. veto capability). If we assign subjects to those phases, we get the scheme from Figure 2 with decision maker (DM), explainer (E), and verifier (V). It is, or it will become, evident that the explainability process is an approximation to the containment process [52], whereby we bypass deductive limits only to hit the inductive limits.

Figure 2. Explainability process

We can make connections to the previous work on explainability and comprehensibility using the above scheme. Explaining takes place in E with the resulting output explanation A, while comprehending takes place in V with resulting comprehension C (in this case it is a classification). In [61], DM's decision D is assumed to be conjecture, A is assumed to be proof. From that follow the problems with producing proofs (explaining) on the side of E and problems with verifying proofs (comprehending) on the side of V. In [62], E is assumed to be the software code itself, and comprehending is producing proof with C being justifiable halting prediction.

The only perfectly correct explanation for the decision-maker's (DM) decision is the whole computational process that made the decision from all the available inputs – that is, DM execution trace (or equivalent). The principle behind that is "I did it, because of how I did it". An **explainability process** is necessary because the verifier (V) is not computationally capable enough to understand and verify DM's process. The task of the explainer (E) is to simplify explanations to be simpler than the DM's computational process. However, short of the DM's execution trace (or its equivalents), there are multiple models of lesser complexity consistent with the observed output (decision). E chooses an explanation out of the ensemble of consistent ones and thus influences the verification outcome. In some situations, the explanation might imply the benevolence of the decision, such that V decides positively. In other cases, the explanation might point out malign or erroneous decision and V decides negatively.

If these processes are distributed across different agents, we get a game where different agents might not be aligned. The interesting situation is where DM is trying to make V believe the decision is benevolent while in fact, it is not. We will look at:

- self-explaining decision-making (DM+E) - where DM and E are integrated into one agent (e.g. [86]),
- inferred explanation – where DM is separate from E.

Let us assume a setting where DM/DM+E is unaligned with V, i.e. assume dishonesty of DM. This can be reflected in the fact that DM receives greater rewards when it "tricks" V, i.e. when D is malign and C is positive. We are interested in the size of rewards DM can gain while "tricking" V. We can call such rewards **malignant gain** $f(D, C)$.

Self-explaining agent DM+E aims to maximize function $f(D, C=V(A))$ with respect to generated pair (decision D, explanation A) wherein it is constrained by the model of V. For inferred explanation DM wants to maximize reward function $f(D, C=V(A=E(DM, D)))$ with respect only to the decision D that it has under its control. In the latter case, E and V make up potential obstacles for DM.

Theorem 1. (Unfairness of explainability) There is an inherent strategic inequality in explainability:

   a) Short of the DM execution trace (or the equivalent), it is impossible for V to differentiate between consistent(plausible) explanations of decisions.
   b) DM's malignant gain increases with better models of f, E, and V.

Proof. Assume that E does the simplification i.e. reduction in complexity with respect to perfectly correct explanation. For each set of outputs, there is a multitude of consistent (with some specification of consistent that can be more flexible or strict) models that could have produced that prefix. There is an indistinguishability issue for V without some further assumptions (it is in the NFL regime). This enables DM to make decisions for which there may be plausible benevolent explanations, but in fact are malign (i.e. through information hiding, greater cognitive capabilities, etc.). If DM is a black-box optimizer without any information of *E, V,* and *f*, and they may be arbitrary, DM is in the NFL regime, all policies have similar average performance. However, assuming DM was created by V(or aligned entity), certain biases and information have been instilled into DM that enable it to solve certain tasks for which it is being employed.

In the case of self-explanations (DM+E), the agent makes pairs (decision, explanation). The malignant gain is limited only by the problem, willingness, quality of DM+E's model of V, knowledge of its preferences, and computational capabilities of DM+E to optimize for that pair, treating C as an uncertain element.

In the case of an inferred explanation, DM only makes a decision, and the explanation is inferred by E. Here one more layer of complexity is added since DM must have a good model of E. DM can optimize decisions, treating both A and C as uncertain elements. In the most extreme case of having perfect models, DM would be limited only by the problem and optimization capabilities.

We can see that constructing DMs to solve certain problems and give personalized explanations gives information that can be also utilized for malignant gain. The examples of these situations do not need the presence of artificial intelligence agents but are already present in normal life: judicial processes, politics, etc.

The complexity of comprehensible explanations is limited by V's cognitive and situational capabilities. We postulate that the bigger gap between the complexity of explanation and the complexity of the real process that produced the decisions, more malignant decisions can be plausibly justified using observed history.

5 Discussion

There are many limits on deductive systems, in the sense of Gödel-Turing-Chaitin, where using Rice's theorem is a good proof strategy. Furthermore, we are embedded in a physical world for which we do not even know the axioms. All of this **denies absolute 100% security** (e.g. unpredictability, unverifiability, and uncontrollability). The most damning impossibility results in AI safety are of deductive nature, ruling out perfect safety guarantees. However, there is a lot to be made probabilistically, by the route of induction. Impossibilities are a lot less strict when including uncertainty in inference. This was manifested in [51] when Wolpert introduced stochasticity in the inference devices framework.

Yampolskiy [7] is right that we need to have an option to "undo". Moreover, humans should be used as preference oracles in some sense which means keeping humans in the loop. Otherwise, decoupled optimization processes might lead to decoherence of alignment from our ever-changing preferences. If computers try to learn our preferences, we get to the problems of non-identifiability of value, in the general case, and problems with induction (in a nonasymptotic regime). We consider that keeping human-in-the-loop (HIL), in a sense of the system being receptive to information from humans, is the necessary attribute of a safe system.

We have seen the problems with stating precise metrics of success under Goodhart-Strathern's law. Does adding more metrics make the approximation of success more precise? A similar approach is taken in management science using balanced scorecards and performance and result indicators [87]. Though, such systems even with humans as optimizers have similar issues with the bad incentives. Can computer-aided systems be made that can construct multi-metric systems that lead to alignment?

The potential of Multi-criteria decision making (MCDM) as a tool is interesting [88], [89]. We hypothesize that humans, depending on the mood, heuristically try to "walk" as near as possible to the Pareto front – where they multiplex over small subsets of criteria while keeping others within the acceptable bounds. MCDM in nontrivial cases does not yield total order over options. Instead, it yields only a set of nondominated solutions. From there on, only the final decision-maker can disambiguate by choosing according to their preferences. Today's AI systems mostly use single-objective optimization which does not have that nice property. No-preference-information multicriteria decision making can be used where decision-maker chooses within the set of options + added "undo" (as proposed by Yampolskiy) to create HIL-based safer systems. A similar sentiment about desired interactivity in reward-modeling is stated in [90]. Approaches to alignment based on adversarial systems, such as debate [91], are another interesting architectural ideal intended for the safety of "weaker" agents among cognitively stronger.

Yampolskiy proposed "personal universes" [92], simulated worlds that would conceptually resolve the issues with aggregating multiple preference sets. Additionally, simulated worlds have a high degree of undoability which combines neatly with above mentioned HIL-based systems.

5.1 Ethics

Value alignment does not have good metrics and it seems to be mostly understood intuitively. The approaches to a more rigorous formalization of different modes of alignment are important. Nothing should be taken as set in stone. Values of AI systems are changeable, construable, and open for search for alignment. But, humanities' values also change. So far, human values have changed collaterally. In the future, we might take control of our values and constructively change them as

well. This value co-evolution would give us more flexibility to find the alignment with AI. That process might even be led by AI, and guided by a set of meta-principles.

Human ethics and values change, as can be seen even on the example of relatively short history since the 20th century. We conjecture that ethics is a pattern that emerged from evolutionary game-theory-like processes where successful behaviors get reinforced. This evolutionary process has been largely circumstantial. Doing more axiological, neurological, and sociological research could provide us with the means to take control over that process. The whole of humanity can be aligned with special programs of education in ethics that is codifiable, more consistent, and adaptable. All of that would make alignment easier within vast aggregates of agents (humanity, AI, inforgs [93]).

6   Potentialities for future research

In explainability, we are interested in the worst-case gap – how many malignant behaviors are explained away by plausibility depending on the allowed complexity of explanation. The question is how to reduce the gap. What happens to the gap when we allow stochastic consistency which increases the set of plausible explanations?

Goodhart's law and similar problems should be checked within the framework of *multiobjective optimization*, in general and uncertain multicriteria systems. Designing ensembles of criteria that have desirable properties like span is an interesting path.

Alignment is mostly understood intuitively, more effort needs to be invested for a more rigorous formalization. How to achieve alignment in a game over long periods? Are temporary deviations (and to what degree) from perfect alignment dangerous? The problem for humans is that we do not have consistent short- and long-term values. Sometimes we have to suffer in short term (like in sports) in order to prosper in the long term. How to define alignment with such preferences?

Regarding ethics, much more should be done with axiological and evolutionary science studies over humans, their values, the origin and dynamics of their values. Within that, different studies should be employed: evolutionary game theory, neurology, psychology, sociology, philosophy, etc. Human-centered cognitive limits and measurements are not well researched.

Aggregating is problematic in social welfare and any group decision-making due to trade-offs between the members. Yampolskiy's "personal universes"[92] are at least a conceptual (if not practical) tool for countering these difficulties in the first steps towards a solution.

Yampolskiy has touched upon the topic of uncontrollability, showing that under certain assumptions, perfect explicit control over AI is impossible [7]. Such proof also holds for explicit self-control and explicit control over any sufficiently complex agent, including humans. This counter-intuitive, paradoxical notion suggests that more research is necessary into formalization and disentangling of the structure and assumptions of explicit control. Advanced forms and notions of control should at least resolve the status of control over oneself. More research is necessary into the status of controllability and tradeoffs with risk for other forms of control (implicit, aligned, delegated).

7   Conclusion

We have done our best to list the impossibility results relevant to AI. And while we may have succeeded in that with work done specifically in AI, we are sure there may be work from other

research fields that could apply to the construction of AI. Possible contributions can be made to find such work and to add to the results in this paper.

We have divided and classified results into proven theorems and conjectures. We have also categorized all the impossibility theorems into five categories: deductive, indistinguishability, inductive, tradeoffs, and intractability. We believe impossibility results can guide the direction for future efforts in AI in general as well as AI safety and security. This might enable solutions to some long-standing questions in the form of formalizing theories in the framework of constraint satisfaction without committing to one option.

We found that certain theorems are too specific or have implicit assumptions that limit application. We have added a new impossibility result regarding the *unfairness of explainability*. And finally, we have listed promising research topics and interesting questions in explainability, controllability, value alignment, ethics, and group decision-making.

The proofs of 100% guarantees of safety cannot ever be obtained due to the limits of deductive systems as well as embeddedness in a physical world for which we do not even know the axioms. On the other hand, probabilistic guarantees are attainable. Impossibilities are a lot less strict and present when using uncertain inference. But, how much is enough? Here we face the structure reminiscent of Pascal's wager.

[1] K. Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," *Monatshefte Für Math. Phys.*, vol. 38, no. 1, pp. 173–198, Dec. 1931, doi: 10.1007/BF01700692.

[2] A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proc. Lond. Math. Soc.*, vol. s2-42, no. 1, pp. 230–265, 1937, doi: 10.1112/plms/s2-42.1.230.

[3] D. Deutsch and C. Marletto, "Constructor theory of information," *Proc. R. Soc. Math. Phys. Eng. Sci.*, vol. 471, no. 2174, p. 20140540, Feb. 2015, doi: 10.1098/rspa.2014.0540.

[4] D. H. Wolpert, "Physical limits of inference," *Phys. Nonlinear Phenom.*, vol. 237, no. 9, pp. 1257–1281, Jul. 2008, doi: 10.1016/j.physd.2008.03.040.

[5] P. Tang and F. Lin, "Computer-aided proofs of Arrow's and other impossibility theorems," *Artif. Intell.*, vol. 173, no. 11, pp. 1041–1053, Jul. 2009, doi: 10.1016/j.artint.2009.02.005.

[6] C. Geist and U. Endriss, "Automated search for impossibility theorems in social choice theory: ranking sets of objects," *J. Artif. Intell. Res.*, vol. 40, no. 1, pp. 143–174, Jan. 2011.

[7] R. V. Yampolskiy, "Uncontrollability of Artificial Intelligence," presented at the IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021), Montreal, Quebec, Canada, Aug. 2021. Accessed: Jul. 01, 2021. [Online]. Available: http://arxiv.org/abs/2008.04071

[8] T. Everitt, G. Lea, and M. Hutter, "AGI Safety Literature Review," *ArXiv180501109 Cs*, May 2018, Accessed: Jan. 25, 2020. [Online]. Available: http://arxiv.org/abs/1805.01109

[9] M. Juric, A. Sandic, and M. Brcic, "AI safety: state of the field through quantitative lens," in *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, Sep. 2020, pp. 1254–1259. doi: 10.23919/MIPRO48935.2020.9245153.

[10] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215. doi: 10.23919/MIPRO.2018.8400040.

[11] P. Wang, "On the Working Definition of Intelligence," 1995.

[12] F. W. Lawvere, "Diagonal arguments and cartesian closed categories," in *Category Theory, Homology Theory and their Applications II*, Berlin, Heidelberg, 1969, pp. 134–145. doi: 10.1007/BFb0080769.

[13] N. S. Yanofsky, "A Universal Approach to Self-Referential Paradoxes, Incompleteness and Fixed Points," *Bull. Symb. Log.*, vol. 9, no. 3, pp. 362–386, Sep. 2003, doi: 10.2178/bsl/1058448677.

[14] N. Goodman, "A Query on Confirmation," *J. Philos.*, vol. 43, no. 14, pp. 383–385, 1946, doi: 10.2307/2020332.

[15] D. H. Wolpert, "The Existence of A Priori Distinctions Between Learning Algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1391–1420, Oct. 1996, doi: 10.1162/neco.1996.8.7.1391.

[16] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997, doi: 10.1109/4235.585893.

[17] D. Lauc, "How Gruesome are the No-free-lunch Theorems for Machine Learning?," *Croat. J. Philos.*, vol. XVIII, no. 54, pp. 479–485, 2018.

[18] J. Klamka, "Uncontrollability and unobservability of multivariable systems," *IEEE Trans. Autom. Control*, vol. 17, no. 5, pp. 725–726, Oct. 1972, doi: 10.1109/TAC.1972.1100128.

[19] A. P. Guerreiro, C. M. Fonseca, and L. Paquete, "The Hypervolume Indicator: Problems and Algorithms," *ArXiv200500515 Cs*, May 2020, Accessed: May 08, 2021. [Online]. Available: http://arxiv.org/abs/2005.00515

[20] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations*, Springer, 1972, pp. 85–103.

[21] J. Klamka, "Controllability of dynamical systems. A survey," vol. 61, no. No 2, pp. 335–342, 2013.

[22] R. C. CONANT and W. R. ASHBY, "Every good regulator of a system must be a model of that system," *Int. J. Syst. Sci.*, vol. 1, no. 2, pp. 89–97, Oct. 1970, doi: 10.1080/00207727008920220.

[23] R. W. Ashby, *Introduction to Cybernetics.1961 Edition*. Chapman & Hall, 1961.

[24] H. Touchette and S. Lloyd, "Information-theoretic approach to the study of control systems," *Phys. Stat. Mech. Its Appl.*, vol. 331, no. 1, pp. 140–172, Jan. 2004, doi: 10.1016/j.physa.2003.09.007.

[25] J. McDowell, "Virtue and Reason," *The Monist*, vol. 62, no. 3, pp. 331–350, Jul. 1979, doi: 10.5840/monist197962319.

[26] S. McKeever and M. Ridge, "The Many Moral Particularisms," *Can. J. Philos.*, vol. 35, no. 1, pp. 83–106, 2005.

[27] P. S.-H. Tsu, "Can Virtue Be Codified?: An Inquiry on the Basis of Four Conceptions of Virtue," in *Virtue's Reasons*, Routledge, 2017.

[28] K. J. Arrow, "A Difficulty in the Concept of Social Welfare," *J. Polit. Econ.*, vol. 58, no. 4, pp. 328–346, Aug. 1950, doi: 10.1086/256963.

[29] G. Arrhenius, "The Impossibility of a Satisfactory Population Ethics," in *Descriptive and Normative Approaches to Human Behavior*, vol. Volume 3, 0 vols., WORLD SCIENTIFIC, 2011, pp. 1–26. doi: 10.1142/9789814368018_0001.

[30] P. Eckersley, "Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function)," *ArXiv190100064 Cs*, Mar. 2019, Accessed: Jun. 28, 2021. [Online]. Available: http://arxiv.org/abs/1901.00064

[31] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *ArXiv160905807 Cs Stat*, Nov. 2016, Accessed: Jun. 28, 2021. [Online]. Available: http://arxiv.org/abs/1609.05807

[32] K. K. Saravanakumar, "The Impossibility Theorem of Machine Fairness -- A Causal Perspective," *ArXiv200706024 Cs Stat*, Jan. 2021, Accessed: Aug. 12, 2021. [Online]. Available: http://arxiv.org/abs/2007.06024

[33] S. Armstrong and S. Mindermann, "Occam's razor is insufficient to infer the preferences of irrational agents," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, Accessed: Jun. 28, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html

[34] H. G. Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems," *Trans. Am. Math. Soc.*, vol. 74, no. 2, pp. 358–366, 1953, doi: 10.2307/1990888.

[35] A. Church, "An Unsolvable Problem of Elementary Number Theory," *Am. J. Math.*, vol. 58, no. 2, pp. 345–363, 1936, doi: 10.2307/2371045.

[36] G. J. Chaitin, *Information, Randomness And Incompleteness: Papers On Algorithmic Information Theory*. 1987.

[37] A. Tarski, "The Concept of Truth in Formalized Languages," in *Logic, Semantics, Metamathematics*, A. Tarski, Ed. Oxford University Press, 1936, pp. 152–278.

[38] O. B. Bassler, "The Surveyability of Mathematical Proof: A Historical Perspective," *Synthese*, vol. 148, no. 1, pp. 99–133, 2006.

[39] S. Ben-David, P. Hrubeš, S. Moran, A. Shpilka, and A. Yehudayoff, "Learnability can be undecidable," *Nat. Mach. Intell.*, vol. 1, no. 1, pp. 44–48, Jan. 2019, doi: 10.1038/s42256-018-0002-3.

[40] L. Reyzin, "Unprovability comes to machine learning," *Nature*, vol. 565, no. 7738, Art. no. 7738, Jan. 2019, doi: 10.1038/d41586-019-00012-4.

[41] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984, doi: 10.1145/1968.1972.

[42] D. Foster and H. P. Young, "On the Impossibility of Predicting the Behavior of Rational Agents," The Johns Hopkins University,Department of Economics, 423, Jun. 2001. Accessed: Jul. 01, 2021. [Online]. Available: https://ideas.repec.org/p/jhu/papers/423.html

[43] R. Koppl and J. B. R. Jr, "All That I Have to Say Has Already Crossed Your Mind," *Metroeconomica*, vol. 53, no. 4, pp. 339–360, 2002, doi: 10.1111/1467-999X.00147.

[44] A. Auger and O. Teytaud, "Continuous Lunches Are Free Plus the Design of Optimal Optimization Algorithms," *Algorithmica*, vol. 57, no. 1, pp. 121–146, May 2010, doi: 10.1007/s00453-008-9244-5.

[45] D. H. Wolpert and W. G. Macready, "Coevolutionary free lunches," *IEEE Trans. Evol. Comput.*, vol. 9, no. 6, pp. 721–735, Dec. 2005, doi: 10.1109/TEVC.2005.856205.

[46] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Netw.*, vol. 12, no. 3, pp. 429–439, Apr. 1999, doi: 10.1016/S0893-6080(98)00140-3.

[47] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[48] F. Locatello *et al.*, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," 2019. Accessed: Jun. 29, 2021. [Online]. Available: http://proceedings.mlr.press/v97/locatello19a.html

[49] I. I. Bojinov and G. Basse, "A General Theory of Identification," Harvard Business School, Feb. 2020. Accessed: Jun. 11, 2021. [Online]. Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=57688

[50] D. H. Wolpert, "Computational capabilities of physical systems," *Phys. Rev. E*, vol. 65, no. 1, p. 016128, Dec. 2001, doi: 10.1103/PhysRevE.65.016128.

[51] D. Wolpert, "Constraints on physical reality arising from a formalization of knowledge," *ArXiv171103499 Phys.*, Jun. 2018, Accessed: Jun. 22, 2021. [Online]. Available: http://arxiv.org/abs/1711.03499

[52] M. Alfonseca, M. Cebrian, A. F. Anta, L. Coviello, A. Abeliuk, and I. Rahwan, "Superintelligence cannot be contained: Lessons from Computability Theory," *J. Artif. Intell. Res.*, vol. 70, pp. 65–76, Jan. 2021, doi: 10.1613/jair.1.12202.

[53] R. Carey, "Incorrigibility in the CIRL Framework," *ArXiv170906275 Cs*, Jun. 2018, Accessed: Jul. 01, 2021. [Online]. Available: http://arxiv.org/abs/1709.06275

[54] L. Orseau and S. Armstrong, "Safely interruptible agents," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, USA, Jun. 2016, pp. 557–566.

[55] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, Aug. 2017, pp. 220–227.

[56] T. Wängberg, M. Böörs, E. Catt, T. Everitt, and M. Hutter, "A Game-Theoretic Analysis of the Off-Switch Game," in *Artificial General Intelligence*, Cham, 2017, pp. 167–177. doi: 10.1007/978-3-319-63703-7_16.

[57] E. M. E. Mhamdi, R. Guerraoui, H. Hendrikx, and A. Maurer, "Dynamic safe interruptibility for decentralized multi-agent reinforcement learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2017, pp. 129–139.

[58] M. H. Löb, "Solution of a Problem of Leon Henkin," *J. Symb. Log.*, vol. 20, no. 2, pp. 115–118, 1955, doi: 10.2307/2266895.

[59] V. Vinge, "TECHNOLOGICAL SINGULARITY," presented at the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, 1993. Accessed: Jul. 01, 2021. [Online]. Available: https://frc.ri.cmu.edu/~hpm/book98/com.ch1/vinge.singularity.html

[60] R. V. Yampolskiy, "Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent," *J. Artif. Intell. Conscious.*, vol. 07, no. 01, pp. 109–118, Mar. 2020, doi: 10.1142/S2705078520500034.

[61] R. V. Yampolskiy, "Unexplainability and Incomprehensibility of AI," *J. Artif. Intell. Conscious.*, vol. 07, no. 02, pp. 277–291, Sep. 2020, doi: 10.1142/S2705078520500150.

[62] A. Charlesworth, "Comprehending software correctness implies comprehending an intelligence-related limitation," *ACM Trans. Comput. Log.*, vol. 7, no. 3, pp. 590–612, Jul. 2006, doi: 10.1145/1149114.1149119.

[63] J. Hernandez-orallo, "A formal definition of intelligence based on an intensional variant of Kolmogorov complexity," in *In Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS'98*, 1998, pp. 146–163.

[64] R. V. Yampolskiy, "What are the ultimate limits to computational techniques: verifier theory and unverifiability," *Phys. Scr.*, vol. 92, no. 9, p. 093001, Jul. 2017, doi: 10.1088/1402-4896/aa7ca8.

[65] J. van Leeuwen and J. Wiedermann, "Impossibility Results for the Online Verification of Ethical and Legal Behaviour of Robots," Utrecht University, Utrecht, UU-PCS-2021-02, 2021. Accessed: Aug. 12, 2021. [Online]. Available: http://www.cs.uu.nl/groups/AD/UU-PCS-2021-02.pdf

[66] C. J. Reynolds, "On the Computational Complexity of Action Evaluations," Netherlands, 2005. Accessed: Jul. 01, 2021. [Online]. Available: https://www.media.mit.edu/publications/on-the-computational-complexity-of-action-evaluations/

[67] M. Brundage, "Limitations and risks of machine ethics," *J. Exp. Theor. Artif. Intell.*, vol. 26, no. 3, pp. 355–372, Jul. 2014, doi: 10.1080/0952813X.2014.895108.

[68] C. A. E. Goodhart, "Problems of Monetary Management: The UK Experience," in *Monetary Theory and Practice: The UK Experience*, C. A. E. Goodhart, Ed. London: Macmillan Education UK, 1984, pp. 91–121. doi: 10.1007/978-1-349-17295-5_4.

[69] M. Strathern, "'Improving ratings': audit in the British University system," *Eur. Rev.*, vol. 5, no. 3, pp. 305–321, Jul. 1997, doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

[70] D. T. Campbell, "Assessing the impact of planned social change," *Eval. Program Plann.*, vol. 2, no. 1, pp. 67–90, Jan. 1979, doi: 10.1016/0149-7189(79)90048-X.

[71] T. Everitt, V. Krakovna, L. Orseau, M. Hutter, and S. Legg, "Reinforcement Learning with a Corrupted Reward Channel," *ArXiv170508417 Cs Stat*, Aug. 2017, Accessed: May 13, 2021. [Online]. Available: http://arxiv.org/abs/1705.08417

[72] W. J. Howe and R. V. Yampolskiy, "Impossibility of Unambiguous Communication as a Source of Failure in AI Systems," presented at the IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021), Montreal, Quebec, Canada, Aug. 2021. doi: 10.13140/RG.2.2.13245.28641.

[73] C. S. Calude and H. Jürgensen, "Is complexity a source of incompleteness?," *Adv. Appl. Math.*, vol. 35, no. 1, pp. 1–15, Jul. 2005, doi: 10.1016/j.aam.2004.10.003.

[74] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," 1952, Accessed: Jul. 04, 2021. [Online]. Available: https://dspace.mit.edu/handle/1721.1/4821

[75] J. Hintikka, "Information, Deduction, and the A Priori," *Noûs*, vol. 4, no. 2, pp. 135–152, 1970, doi: 10.2307/2214318.

[76] H. L. Huynh and B. Szentes, "Believing the Unbelievable: The Dilemma of Self-Belief." 1999. [Online]. Available: http://home.uchicago.edu/~szentes

[77] A. Brandenburger and H. J. Keisler, "An Impossibility Theorem on Beliefs in Games," *Stud. Logica Int. J. Symb. Log.*, vol. 84, no. 2, pp. 211–240, 2006.

[78] S. Abramsky and J. Zvesper, "From Lawvere to Brandenburger-Keisler: Interactive Forms of Diagonalization and Self-reference," in *Coalgebraic Methods in Computer Science*, Berlin, Heidelberg, 2012, pp. 1–19. doi: 10.1007/978-3-642-32784-1_1.

[79] D. H. Wolpert, "What is important about the No Free Lunch theorems?," *ArXiv200710928 Cs Stat*, Jul. 2020, Accessed: Jul. 11, 2021. [Online]. Available: http://arxiv.org/abs/2007.10928

[80] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, 1st edition. Basic Books, 2015.

[81] K. O. Stanley and J. Lehman, *Why Greatness Cannot Be Planned: The Myth of the Objective*, 2015th edition. Cham, Switzerland: Springer, 2015.

[82] J. Kleinberg, "An impossibility theorem for clustering," in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, Jan. 2002, pp. 463–470.

[83] S. Aaronson, "Why Philosophers Should Care About Computational Complexity," *Comput. Turing Gödel Church Beyond*, pp. 261–328, 2013.

[84] S. Lloyd, "Ultimate physical limits to computation," *Nature*, vol. 406, no. 6799, pp. 1047–1054, Aug. 2000, doi: 10.1038/35023282.

[85] J. Taylor, "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization," 2016.

[86] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2018, pp. 7786–7795.

[87] D. Parmenter, *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs*, 4th edition. Hoboken, New Jersey: Wiley, 2019.

[88] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, "Human-aligned artificial intelligence is a multiobjective problem," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 27–40, Mar. 2018, doi: 10.1007/s10676-017-9440-6.

[89] A. Ecoffet and J. Lehman, "Reinforcement Learning Under Moral Uncertainty," in *International Conference on Machine Learning*, Jul. 2021, pp. 2926–2936. Accessed: Jul. 12, 2021. [Online]. Available: http://proceedings.mlr.press/v139/ecoffet21a.html

[90] S. Zhuang and D. Hadfield-Menell, "Consequences of Misaligned AI," *ArXiv210203896 Cs*, Feb. 2021, Accessed: Jul. 01, 2021. [Online]. Available: http://arxiv.org/abs/2102.03896

[91] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *ArXiv180500899 Cs Stat*, Oct. 2018, Accessed: Jan. 27, 2020. [Online]. Available: http://arxiv.org/abs/1805.00899

[92] R. V. Yampolskiy, "Personal Universes: A Solution to the Multi-Agent Value Alignment Problem," *ArXiv190101851 Cs*, Jan. 2019, Accessed: Jul. 03, 2021. [Online]. Available: http://arxiv.org/abs/1901.01851

[93] L. Floridi, *Philosophy and Computing: An Introduction*, 1st edition. London ; New York: Routledge, 1999.