Impossibility Results in AI: A Survey

MARIO BRCIC, University of Zagreb Faculty of Electrical Engineering and Computing, Croatia
ROMAN V. YAMPOLSKIY, University of Louisville, USA

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. We strongly believe this to be the most prudent approach to long-term AI safety initiatives. In this article, we have categorized impossibility theorems applicable to AI into five mechanism-based 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 new results (theorems) such as the unfairness of explainability, the first explainability-related result in the induction category. The remaining results deal with misalignment between the clones and put a limit to the self-awareness of agents. 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.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Philosophical/theoretical foundations of artificial intelligence;** • **Security and privacy** → *Social aspects of security and privacy*; **Human and societal aspects of security and privacy**;

Additional Key Words and Phrases: Artificial intelligence, AI safety, limitations, impossibility theorems

ACM Reference format:
Mario Brcic and Roman V. Yampolskiy. 2023. Impossibility Results in AI: A Survey. *ACM Comput. Surv.* 56, 1, Article 8 (August 2023), 24 pages.
https://doi.org/10.1145/3603371

> 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 [40] and Turing's undecidability

Authors' addresses: M. Brcic (corresponding author), University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, 10000, Zagreb, Croatia; email: mario.brcic@fer.hr; R. V. Yampolskiy, University of Louisville, 132 Eastern Pkwy, 40292, Louisville, KY; email: roman.yampolskiy@louisville.edu.

This work is licensed under a Creative Commons Attribution International 4.0 License.
© 2023 Copyright held by the owner/author(s).
0360-0300/2023/08-ART8
https://doi.org/10.1145/3603371

results [90] 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 of themselves do not point to the solutions, are useful in the sense that they guide the direction for future efforts in **artificial intelligence (AI)** in general, but in our case, the interest is in AI safety and security. For example, it may point to difficulties in verifiability for some of the current approaches [30]. 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 [24]. 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 [99]. Moreover, automated proof [85] and search [35] 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 based on two independent axes: *mechanism* and *domain*. Mechanism axis should help in finding new impossibility results. Under that axis, all the impossibility theorems are shown to be neatly subsumed under the problems with capacity disparity where several related objects differ in size. Domain axis is informative for applications in system engineering that combines symbolic and learned modules. 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, economics, social choice theory, and so on. In the process, we shall show new results. The first result 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. The second result deals with misaligned embodiments when the cloning procedure can produce an adversarial situation. The third result states the impossibility of being perfectly self-aware.

Previous works cover similar topics within the scope of AI safety [27, 32, 50, 109], but none focused on impossibility theorems as a family, their utility, structure, and connections to other fields. Especially in the light of recent advances of language models on math [74] and programming [60] that have spurred speculations, we strongly believe that research in AI safety should start at hypothesized limitations of AI and work inwards to the current technology. This approach is proactive and more robust than the reactive approach that plays catch-up with current results.

The rest of the article is organized as follows: In Section 2, we introduce basic definitions. Section 3 contains the relevant work presented under newly defined classification. In Section 4, we focus on impossibility theorems developed in the field of AI safety. The discussion is offered in Section 5, and ideas for future research are listed in Section 6. We conclude the article in Section 7.

2 BASIC DEFINITIONS

We shall not impose strict formalization, but we shall keep at the level of lawyerese in this article to ensure wide readability of material. Our investigation is done from the perspective of assumption that *intelligent behavior that can achieve its goals is computable.*

Definition 2.1. **System** is any non-empty part of the universe.

Definition 2.2. **State** is the condition of the universe.

Definition 2.3. **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 Reference [109] 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 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.

Definition 2.4. **Intelligence** is the ability for an information processing system to adapt to its environment with insufficient knowledge and resources [96].

Intelligence is a complex and multifaceted concept that defies an easy all-encompassing definition. Although Definition 2.4 might be contentious, we consider it reasonable and general for describing AI. Regarding adaptivity, there is an obvious gradation to intelligence—the more conditions with lesser resources the system can adapt to, the more intelligence it manifests. Regarding insufficiency of resources and knowledge, in the absence of those, purely brute-force exhaustive search and unrealistic memory retrieval approaches would be permissible, which by *reduction ad absurdum* do not manifest intelligence for big problems.

Definition 2.5. **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.

Definition 2.6. **Stability** of state S for system A is the intrinsic tendency to return to A's desired subset of state-space after being perturbed.

Definition 2.7. **Robustness** of state S for system A is the property of staying within A's desired subset of the state space despite perturbations.

The Definitions 2.5–2.7 gain a matter of degree in the provided guarantees by introducing specific contexts of validity. The more restricted the contexts, the lower the degree of guarantees. These contexts may be based on concepts from dynamic systems and control theory, such as boundedness, bifurcations, critical points, and hierarchical attractor-repeller dynamics, as well as quantitative measures of external conditions, such as multi-context and multi-aspect perturbations. To fully evaluate these properties within their respective contexts of validity, it is important to take a holistic view of system A, considering as many larger meta-systems that contain A as possible. Together, the meta-systems and their respective contexts of validity can provide a complete description of the guarantees provided.

Definition 2.8. **Catastrophic outcome** for system A is any state from which the return to A's desired subset of state space is impossible.

Definition 2.9. **Alignment** within the ensemble of systems $A_1 \ldots A_n$ 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) to prosper in the long term. How to define alignment with such preferences?

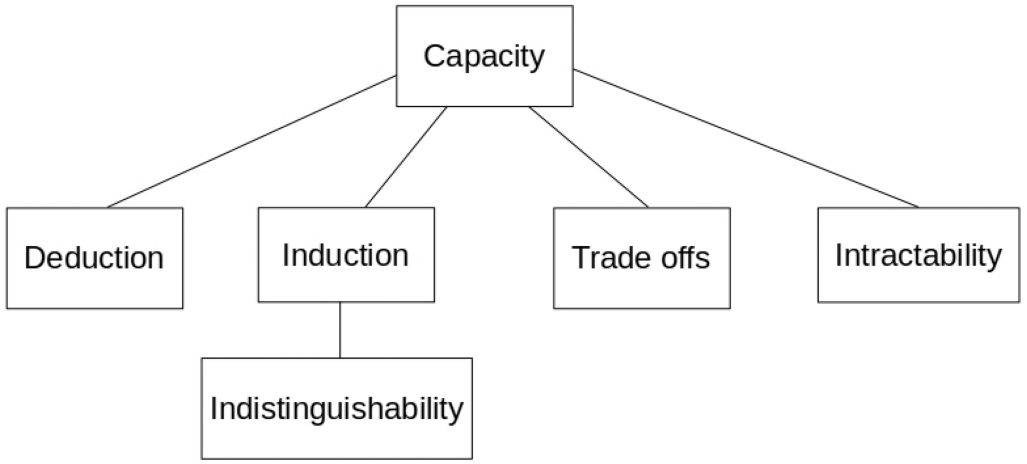

Fig. 1. Proposed mechanism-based categorization of impossibility theorems in AI.

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 **artificial general intelligence (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. We propose two independent axes in classification of impossibility results:

(1) **mechanism** axis, which is related to the capacity disparity
(2) **domain** axis, which pertains to the characteristics of systems where they appear:
    (a) *symbolic (S)*—in formal mathematical models
    (b) *data (D)*—in empirical systems that learn about the world from observations

3.1 Mechanism

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 five subcategories (Figure 1):

(1) Deduction (D),
(2) Induction (I),
(3) Indistinguishability (N),
(4) Tradeoffs (T), and
(5) Intractability (P).

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 [59, 110] 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 [38]. **No Free Lunch (NFL)** theorems [97, 102] deal with induction and are a formalization of Goodman's new riddle of induction [58]. 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, i.e., 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 [52].

Tradeoffs 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 [39]).

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 [51].

We have listed in Table 1 impossibility theorems we deemed important 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.

3.1.1 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 [59, 110] 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 [40] addressing unprovability and Turing's work [90] covering undecidability. In addition to the aforementioned that cover the processing, Gregory Chaitin provided additional incompleteness results [20] 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 [16, 20] whereby *the theorems of finitely stated theories cannot be significantly more complex than the theory itself.*

Rice's theorem [77] 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

Table 1. Impossibility Theorems of Interest to AI Researchers

Name	Source	Proven (Y – yes, N – no, ∼- not rigorous)	Mechanism category (D; I; N; T; P)	Domain category (S; D)
Unobservability	[52]	Y	N	D
Uncontrollability of dynamical systems	[52, 53]	Y	N	D
Good Regulator Theorem	[23]	Y	N	D
Law of Requisite Variety	[8]	Y under perfect information and infinite speed	N	D
Information-theoretical control limits	[88]	Y	N	D
(Anti)codifiability thesis	[65, 66, 89]	N	I	D
Arrow's impossibility theorem	[7]	Y	T	S
Impossibility theorems in population ethics	[6]	Y	T	S
Impossibility theorems in AI alignment	[28]	Y	T	S
Fairness impossibility theorem	[55, 78]	Y	T	S
Limits on preference deduction	[5]	Y	N	D
Rice's Theorem	[77]	Y	D	S
Unprovability	[40]	Y	D	S
Undecidability	[22, 90]	Y	D	S
Chaitin Incompleteness	[20]	Y	D	S
Undefinability	[86]	Y	D	S
Unsurveyability	[10]	N	P	D
Unlearnability	[11, 76]	Y	D	D
	[91]	Y	P	
Unpredictability of rational agents	[34, 56]	Y	I	D
No Free Lunch - supervised learning	[97]	Y	I	D
No Free Lunch - optimization	[102]	Y	I	D

(Continued)

Table 1. Continued

Name	Source	Proven (Y – yes, N – no, ~- not rigorous)	Mechanism category (D; I; N; T; P)	Domain category (S; D)
Free Lunches in continuous spaces, quantum machine learning, coevolutionary problems, and metalearning	[9, 57, 79, 80, 82, 95, 103]	Y	I	D
Unidentifiability	[12, 47, 63, 73]	Y	N	D
Physical limits on inference	[25, 98, 99, 100]	Y	D+I	D
Uncontainability	[3]	Y	D	S+D
Uninterruptibility	[18, 41, 67, 70, 104]	N only under limited assumptions - opened	D	D
Löb's Theorem (unverifiability)	[64]	Y	D	S
Unpredictability of superhuman AI	[94, 108]	Y~ definition of superhuman?	I	D
Unexplainability	[107]	Y~ implicit proof sketch, based on explanation=proof and using at least one of: (unprovability, undefinability), assuming honesty or full model	D	D
Incomprehensibility	[107]	Y~ implicit proof sketch, based on explanation=proof and using at least one of: (unverifiability, unsurveyability, undefinability), assuming honesty or full model	D	D
	[21]	Y comprehension= producing proof and halting games		

(Continued)

Table 1. Continued

Name	Source	Proven (Y – yes, N – no, ∼- not rigorous)	Mechanism category (D; I; N; T; P)	Domain category (S; D)
k-incomprehensibility	[43]	N just definitions	I	D
Unverifiability	[105]	Y	D	S
Unverifiability of robot ethics	[93]	Y	D	S
Intractability of bottom-up ethics	[14, 75]	Y	P	D
No-flattening theorems for deep learning	[61]	Y	P	D
Efficiency of computing Boolean functions for multilayered perceptrons	[15]	Y	P	D
Goodheart's Law (Strathern)	[37, 84]	N	I	D
Campbell's law	[17]	N	I	D
Reward corruption unsolvability	[31]	Y	I	D
Uncontrollability of AI	[109]	Y∼ under degenerate conditions	D	D
Impossibility of unambiguous communication	[45]	Y under strict assumptions	I	D
Unfairness of explainability	here	Y∼ proof sketch	I	D
Misaligned embodiment	here	Y∼ proof sketch	T	D
Limited self-awareness	here	Y∼ proof sketch	I	D

Mechanism categories: D – deduction, I – induction, N - indistinguishability , T – tradeoffs, P – intractability.
Domain categories: S – symbolic, D - data.

deduction limitations in AI, within its assumptions. Löb's theorem [64], informally put, states that a formal consistent system cannot, in general, prove its own soundness.

Regarding formal semantics, Tarski's undefinability theorem [86] 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 [19] in classical semantic information theory entails that contradiction conveys maximal information. Hintikka's scandal of deduction [44] 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 [46] have demonstrated irreconcilability between two notions of self-belief. In References [2, 13] the paradox of self-reference was extended to the games with two or more players that yield impossible beliefs.

Deductive impossibility results have been found in machine learning as well regarding the properties of learning algorithms and processes given the data. Authors in References [11, 76] have shown that learnability can be undecidable and unprovable for certain problems. Finding if an encodable learning algorithm always underfits a dataset is undecidable [81], even with unlimited training time.

Inference devices covered in the series of papers by Wolpert [98–100] 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. This work was extended in Reference [25] to find the constraints on modeling systems from within systems using the frame of relativity. The concluding impossibility theorem states that the universe cannot be completely modeled from within the universe.

3.1.2 Indistinguishability Impossibility Theorems. **Observability** is the ability to infer the state of a black-box system from its input/output data. **Identifiability** (i.e., 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 [12], 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 [52, 53].

Good regulator theorem [23] 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 [8] states that variety in outputs can only be reduced by the state complexity of the controlling mechanism. **Information theory state-control limit** [88] 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 [47]. Similar is shown for unsupervised learning of disentangled representations [63]. There are also well-known non-identifiability limits to causal discovery from the data [73].

3.1.3 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 [38]. **No free lunch theorems (NFL),** by Wolpert [97, 102], 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 [101]. 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" [26]. For example, there are free lunches in continuous spaces [9], in quantum machine learning by utilizing entanglement [82, 95], in coevolutionary problems [103], under certain regularity assumptions in meta-learning [36, 79, 80], and under the assumption of Occam's razor validity [57].

Goodhart-Strathern's law [37, 84] and Campbell's law [17] 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 Reference [83] 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 that influence opponent's behavior [34]. Placing further restrictions on the assumptions can regain predictability. In economic situations, further limits relating to rationality, predictability, and control were proved in Reference [56]. 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 impossibility theorems for Wolpert's inference devices and regarding beliefs in games.

Anticodifiability thesis [65, 66, 89] 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.

3.1.4 Tradeoffs Impossibility Theorems. Tradeoff limits constrain our attempts to achieve perfect outcomes. Examples include impossibility theorems in clustering [54], fairness [55, 78], and **social choice theory (SCT)** [7]. 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, in other words, achieves maximal possible hypervolume indicator [39].

In social choice theory, there are results such as Arrow's impossibility theorem [7], which states there must be a tradeoff that forces choosing only a strict subset of desirable properties in voting mechanisms. In moral theory, there are different problems regarding population ethics [6] 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 [35].

3.1.5 Intractability Impossibility Theorems. Intractability limits divide possibility-in-principle and practically impossible due to the resource limitations. There are three types of intractability impossibility results: asymptotic, physics-based, and human-centered.

Asymptotic intractabilities fall neatly under the complexity theory [1, 51]. That research field is simply too rich to expand on it here. We shall only highlight the **probably approximately correct (PAC)** learning framework [91] by Valiant that defines the border between efficient

(polynomial time) and inefficient learnability. Of our interest is also the intractability of bottom-up ethics [14, 75], which stems from the game-theoretic nature of ethics. Lin et al. in Reference [61] have dealt with possible reasons and situations when deep learning works efficiently and when does it necessitate an exponential number of parameters in the number of variables. They prove various "no-flattening theorems" that show loss of efficiency when approximating deep networks with shallow ones. Calude et al. have extended the latter work in Reference [15] by working on sensitive and robust Boolean functions, whereby especially parity function poses exponential difficulties for learning by multilayered perceptrons with ***single*** hidden layer.

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., Reference [62]) 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 [98–100]. 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** that does not seem to be well researched and measured. Humans, as agents of finite capabilities, have strict limits with regard to explainability, comprehensibility [27], and all other aspects. One of the commonly mentioned impossibilities is unsurveyability [10] in the context of mathematical proofs.

3.2 Domain

Domain axis in the categorization pertains to the characteristics of systems where impossibility results appear:

(1) symbolic (S) – in formal mathematical models
(2) data (D) – in empirical systems that learn about the world from observations

This reflects the two extremes of the artificial intelligence approaches spectrum: Model-based systems (i.e., symbolic/logical/deductive) and empirical data-based systems (i.e., statistical/inductive). The key differentiator between these two extremes lies in the nature and quantity of their inherent biases. The interior of the spectrum is populated with hybrid approaches such as inductive logic programming and neuro-symbolic programming. This axis in categorization also covers the split between mathematics and physics in their usage of the word "theory." In the former, it is a robust and self-sufficient set of crafted principles for which we can have some understanding and rigorous analysis. In the latter, the theory is contingent and not robust to data variations as future data can falsify theories. The symbolic category is mostly hit by impossibilities due to undecidability, high complexity, and constraints. Data category, in addition to undecidability, constraints, and complexity, is subject to underdetermination and intractability.

4 IMPOSSIBILITY THEOREMS DEVELOPED IN AI SAFETY

Uncontainability [3] 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 [105] 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 Reference [93] to show impossibility of online verification of robot's ethical and legal behavior.

Uninterruptibility [18, 41, 67, 70, 104] 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 [94, 108] 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 it proceeds to form a 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 [107] 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 that is then prone to the deduction impossibility theorems 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 Reference [21].

Incomprehensibility [107] 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 impossibility theorems. Understanding is vaguely defined as proof-checking and it is not defined how accuracy is measured. In a similar line of work Charlesworth [21] 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 [109] 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 that 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 [28] 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 [5] 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 [31] 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, quantilization [87] may provide more robustness.

Impossibility of unambiguous communication [45] denies perfectly unambiguous communication using natural language. Many examples are given to show different levels of ambiguity:

Fig. 2. Explainability process.

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 subsections introduce new impossibility results.

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, then 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 [3], whereby, we bypass deductive limits only to hit the inductive limits.

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 Reference [107], 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 Reference [21], explanation A is assumed to be the software code itself, comprehending is the activity of producing proof and comprehension C is a justifiable halting prediction.

The only perfectly correct explanation for the DM's 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 V is not computationally capable enough to understand and verify DM's process. The 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, then 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., Reference [4]),
- 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 4.1 (Unfairness of Explainability). *There is an inherent strategic inequality in explainability:*

(1) Short of the DM execution trace (or the equivalent), it is impossible for V to differentiate between consistent (plausible) explanations of decisions.
(2) 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, then 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, and so on.

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.

4.2 Misaligned Embodiment

This section is inspired by Parfit's work [71].

Let us imagine a situation where a computational agent goes through operational cloning by mistake, in one of two ways: copy or separation. A pure copy would correspond to the network transfer of the agent to another location, mistakenly not deleting the agent at the source location. Separation would correspond to the accidental loss of connectivity between the agent's

coordinated components, each of which has an identical copy of code, in which case they would become two separate agents.

In that situation, we shall refer to the agent's code as a **type** and an instance of that code in each agent as a **token**. We are concerned with the characteristics of types that lead to misalignment between the mistakenly operationally created clones. The decisive assumption is the existence of the "reward mechanism" with which agent's tendencies can faithfully be algorithmically modeled. Especially interesting is the locus of self in self-modeling through reward. Additionally, we assume that the improvement of reward utilizes some scarce resource R.

Definition 4.2 (Reward Mirroring). Two agents mirror rewards if they have identical rewards calculated only from the shared objective information.

Definition 4.3 (Locus of Self). Agent's ***locus of the self*** is the largest set that simultaneously contains all the information that is input to its reward calculation.

For example, if all the information that is input to reward calculation comes from another agent, the locus of self is that other agent. If all such information is replicated among the agents of the same type, then the locus is the type. On the contrary, if all such information is exclusively within the embodiment (private information), then we have a self-interested agent.

Definition 4.4 (Self-interestedness). The agent is self-interested when the locus of self is in the embodiment of that agent.

Few options for the locus of self, somewhat idealized, in addition to the self-interestedness, are:

- in-type-interestedness—e.g., all its clones under reward mirroring
- extra-type-interestedness—hierarchically above its own type, e.g., family, colony, nation, species
- other-instance—another arbitrary agent e.g., specific master
- selflessness—everything

Theorem 4.5 (Misaligned Embodiment). *If self-interested agent A is mistakenly operationally cloned into agent B, then neither A nor B can perfectly control each other.*

Proof. Under the assumption of operational cloning, A is functionally identical to B (tokens of the same type). In copying, B is created intentionally whereby A in future plans wanted to assume its identity. In separation case, two tokens come to existence out of which none gets the primate in identification.

Due to the same functionality and self-interestedness, there is no perfect control-relation between A and B (from A to B). All forms of control are excluded: explicit and implicit due to the self-interested nature of clones and the veil of ignorance with which the pre-cloning agent approaches the situation. Aligned control is not guaranteed, since two clones want the same stuff for themselves over the same scarce resource R. □

A could attain control over B (and vice versa) in certain scenarios when B, depending on the circumstances, can find itself receptive to the control outside. If the locus of self is flexible in a way that incorporates newly created clones, then alignment can be achieved within the type. However, the problem of identification (recognizing the type) must be solved, with the risk of accepting intruders. All this depends on the calculation of reward over the environment and its faithfulness in perception, as opposed to the introspection and rigid model that does not adjust readily. In the latter case, we do not get type-compatibility.

Hanson's clans of brain emulations [42] are related to the above work but differ in at least two ways. First, computational agents therein are human-brain emulations with clear structure owing

to evolution, as opposed to general computational agents that are objects of investigation here. For such reasons, that work pertains mostly to self-interested agents and ties them to be cooperative through identity-confusion that holds for some selected agents close to the copying operation and dissolves with time. Second, copying of computational agents here is made unintentionally while Hanson deals with intentional copying that does not have the veil of ignorance property. Another related work is Reference [69], where superrationality is used to investigate cooperation between correlated decision-makers, which includes the copies of agents as well. The first option considered, precommitment, is valid only under intentional cloning. The second option of mere correlations between decision-making mechanisms makes strong assumptions: simple structure of games, exactly same conditions, and that identical decision-making mechanisms imply identical decision-making processes. Humans can be used as counter-examples to weaker-than perfect correlation; in the space of possible minds, human minds can be considered as quite resembling with regards to decision-making mechanism but still do not achieve superrationality.

4.3 Limited Self-awareness

Definition 4.6 (Awareness). For agent A to be aware of some phenomenon B means that it observes and predictively models B.

Phenomena that agents can be aware of can be external as well as its own internal processes. For awareness of internal processes, we shall use the term self-awareness. It is important to disentangle the notion of the process and the awareness of the process. The first just produces results as sampling from the black-box, the latter one comprises attending to and modeling (understanding) of the process in an ontology that is richer than just the final results and where the model is simpler than the process itself. These six assumptions are useful:

(1) budget - limited computational resources for agents
(2) costliness - every process consumes a positive amount of computational resources
(3) lower-bound - there is minimal, a positive computational cost that can possibly be attained by any process
(4) positive duration of information propagation
(5) process awareness can trail process execution itself only by a limited time
(6) process-awareness relations can be faithfully modeled by the directed graph (awareness-graph), i.e., there is a directed edge from each aware process to the process it is aware of.

Proposition 4.7 (Process Self-awareness). *The process cannot be aware of itself.*

Proof. Assume that process is aware of itself. That means that awareness is part of the process itself. Hence, it should also be aware of that awareness as well—which prolongs the chain of awareness recursively to infinity. This is absurd, as it would necessitate infinite results from finite resources. □

Theorem 4.8 (Limited Self-awareness). *In every agent A, under the assumptions above, there are internal processes A is unaware of.*

The above theorem puts a limit to self-awareness and hence declares perfect self-awareness impossible.

Proof. If there is a node in an awareness-graph without incoming edges, then the claim is trivially proven. Otherwise, due to assumptions 1, 2, and 3, there must exist at least one weakly connected component in awareness-graph (all of the components are finite). Let us pick any

such component and let us call it C. The whole C itself describes a new, bigger process. Since by Proposition 4.7 process cannot be self-aware, that means that we are not aware of the process described by C. □

We hypothesize that the consciousness we experience is the feeling coming from all the self-awareness processes whose results we experience as samples, but we do not have awareness of those processes.

Previously listed results such as impossibility of self-comprehension [21], inconsistency between different notions of self-belief [46], extensions of the paradox of self-reference to the games with two or more players that yield impossible beliefs [2, 13] are related to the topic of this section. Exact connections are to be elaborated in future research. There exist psychological studies that exemplify similar limits in humans [68].

The open question is, can awareness-graph have cycles? In special cases that might function between the trivial processes, but for more advanced agents with more complex processes, it might result in inconsistencies and paradoxes. Russell's paradox could be used as an inspiration where possible venue might be to show if Lawvere hypothesis [59] of weak point surjectivity propagates across awareness chains, like it was shown in Reference [2], that it does propagate along the belief chains.

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 100% guarantees of 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 Reference [100] when Wolpert introduced stochasticity in the inference devices framework.

Yampolskiy [109] 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, then 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 [72]. 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 [29, 92]. 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 Reference [111]. Approaches to alignment based on adversarial systems, such as debate [48], are another interesting architectural ideal intended for the safety of "weaker" agents among cognitively stronger.

Yampolskiy proposed "personal universes" [106], 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 [49]. 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 [33]). Further analysis of this topic is available in Reference [49].

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) 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, and so on. Human-centered cognitive limits and measurements are not well researched.

Aggregating is problematic in social welfare and any group decision-making due to tradeoffs between the members. Yampolskiy's "personal universes"[106] 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 [109]. Such proof also holds for explicit

self-control and explicit control over any sufficiently complex agent, including humans. This counter-intuitive, paradoxical notion points in 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 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 article.

We have divided and classified results into proven theorems and conjectures. We have also categorized all the impossibility theorems into five mechanism-based categories: Deductive, indistinguishability, inductive, tradeoffs, and intractability. Additionally, for the purposes of applications in systems engineering, we have split them into two domain-based categories: symbolic and data-based [15]. 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 also believe impossibility theorems to be the most prudent approach for long-term AI safety research.

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

Advancing the generality and capability of AI systems increases their reach into safety-critical applications. For safety-critical applications, there is higher stringency in safety verification and validation, where more emphasis is placed on formal verification techniques such as model checking and theorem proving. In the case of highly capable AI systems, this route to 100% guarantees of safety is void due to the limits of deductive systems as well as embeddedness in a physical world for which we do not even know the axioms. The alternative is to conduct rigorous testing, analysis, and frequent auditing, but that is in the inductive range. Impossibilities are a lot less strict and present for uncertain inference, but this route rules out 100% guarantees. But, how much testing is enough? Here, we face the structure reminiscent of Pascal's wager.

What can be done is to design limitations into systems that have simplified conditional guarantees based on intrinsic properties such as stability, boundedness, bifurcations, critical points, and hierarchical attractor-repeller dynamic. All the guarantees can be valid conditional on quantitative measures of external aspects such as multi-context, and multi-aspect perturbations in the style of dynamic systems and control theory. It is important to take a holistic view of systems of interest within multiple relevant aggregations, which must be kept within desired subset of state space. As this seems computationally complex, AI bootstrapping with a sequence of increasingly capable AI systems might help with the process of increasing both capabilities and the coverage of conditional control. Perfection is out of reach, but there is hope for practicality. AI policy will have to play an important coordinating role in balancing the chased capabilities and attainable coverage of conditional control in an effort to control the risks. But, no measure will remove every ounce of downside risk from developing highly capable AI systems.

ACKNOWLEDGMENTS

We thank Christian Calude (University of Auckland) and Joseph Sifakis (Verimag laboratory, Grenoble) for useful comments and remarks regarding the domain-based categorization.

REFERENCES

[1] Scott Aaronson. 2013. Why philosophers should care about computational complexity. *Computab.: Turing, Gödel, Church Bey.* (2013), 261–328.

[2] Samson Abramsky and Jonathan Zvesper. 2012. From Lawvere to Brandenburger-Keisler: Interactive forms of diagonalization and self-reference. In *Coalgebraic Methods in Computer Science (Lecture Notes in Computer Science)*, Dirk Pattinson and Lutz Schröder (Eds.). Springer, Berlin, 1–19. DOI: https://doi.org/10.1007/978-3-642-32784-1_1

[3] Manuel Alfonseca, Manuel Cebrian, Antonio Fernandez Anta, Lorenzo Coviello, Andres Abeliuk, and Iyad Rahwan. 2021. Superintelligence cannot be contained: Lessons from computability theory. *J. Artif. Intell. Res.* 70 (Jan. 2021), 65–76. DOI: https://doi.org/10.1613/jair.1.12202

[4] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, 7786–7795.

[5] Stuart Armstrong and Sören Mindermann. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. *Adv. Neural Inf. Process. Syst.* 31 (2018). Retrieved from https://proceedings.neurips.cc/paper/2018/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[6] Gustaf Arrhenius. 2011. The impossibility of a satisfactory population ethics. In *Descriptive and Normative Approaches to Human Behavior*. (Advanced Series on Mathematical Psychology, Vol. 3). World Scientific, 1–26. DOI: https://doi.org/10.1142/9789814368018_0001

[7] Kenneth J. Arrow. 1950. A difficulty in the concept of social welfare. *J. Polit. Econ.* 58, 4 (Aug. 1950), 328–346. DOI: https://doi.org/10.1086/256963

[8] Ross W. Ashby. 1961. *Introduction to Cybernetics.1961 Edition*. Chapman & Hall.

[9] Anne Auger and Olivier Teytaud. 2010. Continuous lunches are free plus the design of optimal optimization algorithms. *Algorithmica* 57, 1 (May 2010), 121–146. DOI: https://doi.org/10.1007/s00453-008-9244-5

[10] O. Bradley Bassler. 2006. The surveyability of mathematical proof: A historical perspective. *Synthese* 148, 1 (2006), 99–133. Retrieved from https://www.jstor.org/stable/20118682.

[11] Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. 2019. Learnability can be undecidable. *Nat. Mach. Intell.* 1, 1 (Jan. 2019), 44–48. DOI: https://doi.org/10.1038/s42256-018-0002-3

[12] Iavor I. Bojinov and Guillaume Basse. 2020. A General Theory of Identification. Retrieved from https://www.hbs.edu/faculty/Pages/item.aspx?num=57688.

[13] Adam Brandenburger and H. Jerome Keisler. 2006. An impossibility theorem on beliefs in games. *Studia Logica: Int. J. Symbol. Logic* 84, 2 (2006), 211–240. Retrieved from https://www.jstor.org/stable/20016831.

[14] Miles Brundage. 2014. Limitations and risks of machine ethics. *J. Exper. Theoret. Artif. Intell.* 26, 3 (July 2014), 355–372. DOI: https://doi.org/10.1080/0952813X.2014.895108

[15] Cristian S. Calude, Shahrokh Heidari, Joseph Sifakis, What perceptron neural networks are (not) good for? *Information Sciences* 621 (2023), 844–857. DOI: https://doi.org/10.1016/j.ins.2022.11.083

[16] Cristian S. Calude and Helmut Jürgensen. 2005. Is complexity a source of incompleteness? *Adv. Appl. Math.* 35, 1 (July 2005), 1–15. DOI: https://doi.org/10.1016/j.aam.2004.10.003

[17] Donald T. Campbell. 1979. Assessing the impact of planned social change. *Eval. Progr. Plan.* 2, 1 (Jan. 1979), 67–90. DOI: https://doi.org/10.1016/0149-7189(79)90048-X

[18] Ryan Carey. 2018. Incorrigibility in the CIRL framework. *arXiv:1709.06275 [cs]* (June 2018).

[19] Rudolf Carnap and Yehoshua Bar-Hillel. 1952. An outline of a theory of semantic information. Retrieved from https://dspace.mit.edu/handle/1721.1/4821. Research Laboratory of Electronics, Massachusetts Institute of Technology.

[20] Gregory J. Chaitin. 1987. *Information, Randomness and Incompleteness: Papers on Algorithmic Information Theory*. World Scientific Publishing Company, Singapore.

[21] Arthur Charlesworth. 2006. Comprehending software correctness implies comprehending an intelligence-related limitation. *ACM Trans. Computat. Logic* 7, 3 (July 2006), 590–612. DOI: https://doi.org/10.1145/1149114.1149119

[22] Alonzo Church. 1936. An unsolvable problem of elementary number theory. *Amer. J. Math.* 58, 2 (1936), 345–363. DOI: https://doi.org/10.2307/2371045

[23] Roger C. Canant and W. Ross Ashby. 1970. Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 2 (Oct. 1970), 89–97. DOI: https://doi.org/10.1080/00207727008920220

[24] David Deutsch and Chiara Marletto. 2015. Constructor theory of information. *Proc. Roy. Societ. A: Math., Phys. Eng. Sci.* 471, 2174 (Feb. 2015), 20140540. DOI: https://doi.org/10.1098/rspa.2014.0540

[25] Abigail Devereaux, Roger Koppl, Stuart Kauffman, and Andrea Roli. 2021. *Constraints on Modeling Systems from within Systems: The Principle of Frame Relativity*. Social Science Research Network, Rochester, NY. DOI: https://doi.org/10.2139/ssrn.3968077

[26] Pedro Domingos. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World (1st ed.)*. Basic Books, New York.

[27] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. 2018. Explainable artificial intelligence: A survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. MIPRO Association, 0210–0215. DOI: https://doi.org/10.23919/MIPRO.2018.8400040

[28] Peter Eckersley. 2019. Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). *arXiv:1901.00064 [cs]* (Mar. 2019).

[29] Adrien Ecoffet and Joel Lehman. 2021. Reinforcement learning under moral uncertainty. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2926–2936. Retrieved from http://proceedings.mlr.press/v139/ecoffet21a.html.

[30] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv:2110.06674 [cs]* (Oct. 2021).

[31] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. 2017. Reinforcement learning with a corrupted reward channel. *arXiv:1705.08417 [cs, stat]* (Aug. 2017).

[32] Tom Everitt, Gary Lea, and Marcus Hutter. 2018. AGI safety literature review. *arXiv:1805.01109 [cs]* (May 2018).

[33] Luciano Floridi. 1999. *Philosophy and Computing: An Introduction (1st edition ed.)*. Routledge, London; New York.

[34] Dean Foster and H. Peyton Young. 2001. *On the Impossibility of Predicting the Behavior of Rational Agents*. Technical Report 423. The Johns Hopkins University, Department of Economics. Retrieved from https://ideas.repec.org/p/jhu/papers/423.html.

[35] Christian Geist and Ulle Endriss. 2011. Automated search for impossibility theorems in social choice theory: Ranking sets of objects. *J. Artif. Intell. Res.* 40, 1 (Jan. 2011), 143–174.

[36] C. Giraud-Carrier. 2005. Toward a justification of meta-learning: Is the no free lunch theorem a showstopper? Retrieved from https://www.semanticscholar.org/paper/Toward-a-Justification-of-Meta-learning-3A-Is-the-No-Giraud-Carrier/fee1abe79f179f465d2725be63e97a50034bc511.

[37] Charles A. E. Goodhart. 1984. Problems of monetary management: The UK experience. In *Monetary Theory and Practice: The UK Experience*, Charles A. E. Goodhart (Ed.). Macmillan Education UK, London, 91–121. DOI: https://doi.org/10.1007/978-1-349-17295-5_4

[38] Nelson Goodman. 1946. A query on confirmation. *J. Philos.* 43, 14 (1946), 383–385. DOI: https://doi.org/10.2307/2020332

[39] Andreia P. Guerreiro, Carlos M. Fonseca, and Luís Paquete. 2021. The hypervolume indicator: Computational problems and algorithms. *Comput. Surv.* 54, 6 (July 2021), 119:1–119:42. DOI: https://doi.org/10.1145/3453474

[40] Kurt Gödel. 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme I. *Monatshefte für Mathematik und Physik* 38, 1 (Dec. 1931), 173–198. DOI: https://doi.org/10.1007/BF01700692

[41] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'17)*. 220–227. DOI: https://doi.org/10.24963/ijcai.2017/32

[42] Robin Hanson. 2016. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press, Oxford, New York.

[43] Jose Hernandez-Orallo. 1998. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS'98)*. ICSC Press, 146–163.

[44] Jaakko Hintikka. 1970. Information, deduction, and the a priori. *Noûs* 4, 2 (1970), 135–152. DOI: https://doi.org/10.2307/2214318

[45] William Howe and Roman V. Yampolskiy. 2021. Impossibility of unambiguous communication as a source of failure in AI systems. In *Proceedings of the Workshop on Artificial Intelligence Safety 2021 co-located with the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21)*, Virtual, August, 2021. CEUR Workshop Proceedings 2916. Retrieved from https://ceur-ws.org/Vol-2916/paper_14.pdf.

[46] Hsueh-Ling Huynh and Balazs Szentes. 1999. Believing the Unbelievable: The Dilemma of Self-Belief. Retrieved from https://personal.lse.ac.uk/szentes/docs/bub3.pdf.

[47] Aapo Hyvärinen and Petteri Pajunen. 1999. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.* 12, 3 (Apr. 1999), 429–439. DOI: https://doi.org/10.1016/S0893-6080(98)00140-3

[48] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv:1805.00899 [cs, stat]* (Oct. 2018).

[49] Sarah Isufi, Kristijan Poje, Igor Vukobratovic, and Mario Brcic. 2022. Prismal view of ethics. *Philosophies* 7, 6 (2022), 134. DOI: https://doi.org/10.3390/philosophies7060134

[50] Mislav Juric, Agneza Sandic, and Mario Brcic. 2020. AI safety: State of the field through quantitative lens. In *Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. MIPRO Association, 1254–1259. DOI: https://doi.org/10.23919/MIPRO48935.2020.9245153

[51] Richard M. Karp. 1972. Reducibility among combinatorial problems. In *Complexity of Computer Computations*. Springer, 85–103.

[52] Jerzy Klamka. 1972. Uncontrollability and unobservability of multivariable systems. *IEEE Trans. Automat. Contr* 17, 5 (Oct. 1972), 725–726. DOI: https://doi.org/10.1109/TAC.1972.1100128

[53] Jerzy Klamka. 2013. Controllability of dynamical systems. A survey. *Bull. Polish Acad. Sci.-Technic. Sci.* 61, 2 (2013), 335–342. Retrieved from http://journals.pan.pl/dlibra/publication/edition/83697 DOI - 10.2478/bpasts-2013-0031.

[54] Jon Kleinberg. 2002. An impossibility theorem for clustering. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS'02)*. MIT Press, Cambridge, MA, 463–470.

[55] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (Nov. 2016).

[56] Roger Koppl and J. Barkley Rosser Jr. 2002. All that I have to say has already crossed your mind. *Metroeconomica* 53, 4 (2002), 339–360. DOI: https://doi.org/10.1111/1467-999X.00147

[57] Tor Lattimore and Marcus Hutter. 2013. No free lunch versus Occam's razor in supervised learning. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence: Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30–December 2, 2011*, David L. Dowe (Ed.). Springer, Berlin, 223–235. DOI: https://doi.org/10.1007/978-3-642-44958-1_17

[58] Davor Lauc. 2018. How gruesome are the no-free-lunch theorems for machine learning? *Croat. J. Philos.* XVIII, 54 (2018), 479–485. Retrieved from https://www.ceeol.com/search/article-detail?id=734617.

[59] F. William Lawvere. 1969. Diagonal arguments and cartesian closed categories. In *Category Theory, Homology Theory and Their Applications II (Lecture Notes in Mathematics)*, Barry Mitchell, Jan-Erik Roos, Friedrich Ulmer, Hans-Berndt Brinkmann, Stephen U. Chase, Paul Dedecker, R. R. Douglas, P. J. Hilton, F. Sigrist, Charles Ehresmann, K. W. Gruenberg, Max A. Knus, F. William Lawvere, and Saunders Mac Lane (Eds.). Springer, Berlin, 134–145. DOI: https://doi.org/10.1007/BFb0080769

[60] Yujia Li, David Choi, and Junyoung Chung. 2022. Competitive programming with AlphaCode. Retrieved from https://storage.googleapis.com/deepmind-media/AlphaCode/competition_level_code_generation_with_alphacode.pdf.

[61] Henry W. Lin, Max Tegmark, and David Rolnick. 2017. Why does deep and cheap learning work so well? *J. Statist. Phys.* 168, 6 (Sept. 2017), 1223–1247. DOI: https://doi.org/10.1007/s10955-017-1836-5

[62] Seth Lloyd. 2000. Ultimate physical limits to computation. *Nature* 406, 6799 (Aug. 2000), 1047–1054. DOI: https://doi.org/10.1038/35023282

[63] Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*. Retrieved from http://proceedings.mlr.press/v97/locatello19a.html.

[64] M. H. Löb. 1955. Solution of a problem of Leon Henkin. *J. Symbol. Logic* 20, 2 (1955), 115–118. DOI: https://doi.org/10.2307/2266895

[65] John McDowell. 1979. Virtue and reason. *Monist* 62, 3 (July 1979), 331–350. DOI: https://doi.org/10.5840/monist197962319

[66] Sean McKeever and Michael Ridge. 2005. The many moral particularisms. *Canad. J. Philos.* 35, 1 (2005), 83–106. Retrieved from https://www.jstor.org/stable/40232238.

[67] El Mahdi El Mhamdi, Rachid Guerraoui, Hadrien Hendrikx, and Alexandre Maurer. 2017. Dynamic safe interruptibility for decentralized multi-agent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, 129–139.

[68] Richard E. Nisbett and Timothy DeCamp Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* 84, 3 (1977), 231–259. DOI: https://doi.org/10.1037/0033-295X.84.3.231

[69] Caspar Oesterheld. 2017. Multiverse-wide Cooperation via Correlated Decision Making. Retrieved from https://longtermrisk.org/multiverse-wide-cooperation-via-correlated-decision-making/.

[70] Laurent Orseau and Stuart Armstrong. 2016. Safely interruptible agents. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI'16)*. AUAI Press, Arlington, Virginia, 557–566.

[71] Derek Parfit. 1986. *Reasons and Persons*. Oxford University Press, Oxford, UK. DOI: https://doi.org/10.1093/019824908X.001.0001

[72] David Parmenter. 2019. *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs (4th edition ed.)*. Wiley, Hoboken, NJ.

[73] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.

[74] Stanislas Polu, Jesse Michael Han, Kunhao Zheng, Mantas Baksys, Igor Babuschkin, and Ilya Sutskever. 2022. Formal mathematics statement curriculum learning. *arXiv:2202.01344 [cs]* (Feb. 2022).

[75] Carson J. Reynolds. 2005. On the computational complexity of action evaluations. In *Proceedings of the 6th International Conference of Computer Ethics: Philosophical Enquiry*. Retrieved from https://www.media.mit.edu/publications/on-the-computational-complexity-of-action-evaluations/.

[76] Lev Reyzin. 2019. Unprovability comes to machine learning. *Nature* 565, 7738 (Jan. 2019), 166–167. DOI: https://doi.org/10.1038/d41586-019-00012-4

[77] Henry Gordon Rice. 1953. Classes of recursively enumerable sets and their decision problems. *Trans. Amer. Math. Soc.* 74, 2 (1953), 358–366. DOI: https://doi.org/10.2307/1990888

[78] Kailash Karthik Saravanakumar. 2021. The impossibility theorem of machine fairness—A causal perspective. *arXiv:2007.06024 [cs, stat]* (Jan. 2021).

[79] Gerhard Schurz. 2008. The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philos. Sci.* 75, 3 (July 2008), 278–305. DOI: https://doi.org/10.1086/592550

[80] Gerhard Schurz. 2017. No free lunch theorem, inductive skepticism, and the optimality of meta-induction. *Philos. Sci.* 84, 5 (Dec. 2017), 825–839. DOI: https://doi.org/10.1086/693929

[81] Sonia Sehra, David Flores, and George D. Montañez. 2021. Undecidability of underfitting in learning algorithms. In *2nd International Conference on Computing and Data Science (CDS'21)*. Stanford, CA, USA, 591–594. DOI: https://doi.org/10.1109/CDS52072.2021.00107

[82] Kunal Sharma, M. Cerezo, Zoë Holmes, Lukasz Cincio, Andrew Sornborger, and Patrick J. Coles. 2022. Reformulation of the no-free-lunch theorem for entangled datasets. *Phys. Rev. Lett.* 128, 7 (Feb. 2022), 070501. DOI: https://doi.org/10.1103/PhysRevLett.128.070501

[83] Kenneth O. Stanley and Joel Lehman. 2015. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, Cham, Switzerland.

[84] Marilyn Strathern. 1997. "Improving Ratings": Audit in the British university system. *Eur. Rev.* 5, 3 (July 1997), 305–321. DOI: https://doi.org/10.1002/(SICI)1234-981X(199707)5:3⟨305::AID-EURO184⟩3.0.CO;2-4

[85] Pingzhong Tang and Fangzhen Lin. 2009. Computer-aided proofs of Arrow's and other impossibility theorems. *Artif. Intell.* 173, 11 (July 2009), 1041–1053. DOI: https://doi.org/10.1016/j.artint.2009.02.005

[86] Alfred Tarski. 1936. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, Alfred Tarski (Ed.). Oxford University Press, 152–278.

[87] Jessica Taylor. 2016. Quantilizers: A safer alternative to maximizers for limited optimization. In *Proceedings of the AAAI Workshop: AI, Ethics, and Society*.

[88] Hugo Touchette and Seth Lloyd. 2004. Information-theoretic approach to the study of control systems. *Phys. A: Statist. Mech. Applic.* 331, 1 (Jan. 2004), 140–172. DOI: https://doi.org/10.1016/j.physa.2003.09.007

[89] Peter Shiu-Hwa Tsu. 2017. Can virtue be codified?: An inquiry on the basis of four conceptions of virtue. In *Virtue's Reasons*. Routledge.

[90] Alan Turing. 1937. On computable numbers, with an application to the entscheidungsproblem. *Proc. London Math. Societ.* s2-42, 1 (1937), 230–265. DOI: https://doi.org/10.1112/plms/s2-42.1.230

[91] Leslie Gabriel Valiant. 1984. A theory of the learnable. *Commun. ACM* 27, 11 (Nov. 1984), 1134–1142. DOI: https://doi.org/10.1145/1968.1972

[92] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. 2018. Human-aligned artificial intelligence is a multiobjective problem. *Ethics Inf. Technol.* 20, 1 (Mar. 2018), 27–40. DOI: https://doi.org/10.1007/s10676-017-9440-6

[93] Jan van Leeuwen and Jiri Wiedermann. 2021. *Impossibility Results for the Online Verification of Ethical and Legal Behaviour of Robots*. Technical Report UU-PCS-2021-02. Utrecht University, Utrecht. Retrieved from http://www.cs.uu.nl/groups/AD/UU-PCS-2021-02.pdf.

[94] Vernor Vinge. 1993. Technological singularity. Retrieved from https://frc.ri.cmu.edu/hpm/book98/com.ch1/vinge.singularity.html.

[95] Tyler Volkoff, Zoë Holmes, and Andrew Sornborger. 2021. Universal compiling and (no-)free-lunch theorems for continuous-variable quantum learning. *PRX Quant.* 2, 4 (Nov. 2021), 040327. DOI: https://doi.org/10.1103/PRXQuantum.2.040327

[96] Pei Wang. 1995. On the working definition of intelligence. Technical Report 94, *Center for Research on Concepts and Cognition*, Indiana University, Bloomington, IN. Retrieved from https://www.researchgate.net/publication/2339604_On_the_Working_Definition_of_Intelligence.

[97] David H. Wolpert. 1996. The existence of a priori distinctions between learning algorithms. *Neural Computat.* 8, 7 (Oct. 1996), 1391–1420. DOI: https://doi.org/10.1162/neco.1996.8.7.1391

[98] David H. Wolpert. 2001. Computational capabilities of physical systems. *Phys. Rev. E* 65, 1 (Dec. 2001), 016128. DOI: https://doi.org/10.1103/PhysRevE.65.016128

[99] David H. Wolpert. 2008. Physical limits of inference. *Phys. D: Nonlin. Phenom.* 237, 9 (July 2008), 1257–1281. DOI: https://doi.org/10.1016/j.physd.2008.03.040

[100] David H. Wolpert. 2018. Constraints on physical reality arising from a formalization of knowledge. *arXiv:1711.03499 [physics]* (June 2018).

[101] David H. Wolpert. 2020. What is important about the no free lunch theorems? *arXiv:2007.10928 [cs, stat]* (July 2020).

[102] David H. Wolpert and William Macready. 1997. No free lunch theorems for optimization. *IEEE Trans. Evolut. Computat.* 1, 1 (Apr. 1997), 67–82. DOI: https://doi.org/10.1109/4235.585893

[103] David H. Wolpert and William Macready. 2005. Coevolutionary free lunches. *IEEE Trans. Evolut. Computat.* 9, 6 (Dec. 2005), 721–735. DOI: https://doi.org/10.1109/TEVC.2005.856205

[104] Tobias Wängberg, Mikael Böörs, Elliot Catt, Tom Everitt, and Marcus Hutter. 2017. A game-theoretic analysis of the off-switch game. In *Artificial General Intelligence (Lecture Notes in Computer Science)*, Tom Everitt, Ben Goertzel, and Alexey Potapov (Eds.). Springer International Publishing, Cham, 167–177. DOI: https://doi.org/10.1007/978-3-319-63703-7_16

[105] Roman V. Yampolskiy. 2017. What are the ultimate limits to computational techniques: Verifier theory and unverifiability. *Phys. Script.* 92, 9 (July 2017), 093001. DOI: https://doi.org/10.1088/1402-4896/aa7ca8

[106] Roman V. Yampolskiy. 2019. Personal universes: A solution to the multi-agent value alignment problem. *arXiv:1901.01851 [cs]* (Jan. 2019).

[107] Roman V. Yampolskiy. 2020. Unexplainability and incomprehensibility of AI. *J. Artif. Intell. Conscious.* 07, 02 (Sept. 2020), 277–291. DOI: https://doi.org/10.1142/S2705078520500150

[108] Roman V. Yampolskiy. 2020. Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent. *J. Artif. Intell. Conscious.* 07, 01 (Mar. 2020), 109–118. DOI: https://doi.org/10.1142/S2705078520500034

[109] Roman V. Yampolskiy. 2022. On the controllability of artificial intelligence: An analysis of limitations. *Journal of Cyber Security and Mobility* 11, 3 (2022), 321–404. DOI: https://doi.org/10.13052/jcsm2245-1439.1132

[110] Noson S. Yanofsky. 2003. A universal approach to self-referential paradoxes, incompleteness and fixed points. *Bull. Symbol. Logic* 9, 3 (Sept. 2003), 362–386. DOI: https://doi.org/10.2178/bsl/1058448677

[111] Simon Zhuang and Dylan Hadfield-Menell. 2021. Consequences of misaligned AI. *arXiv:2102.03896 [cs]* (Feb. 2021).

Received 28 February 2022; revised 29 April 2023; accepted 31 May 2023