Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems

David Dalrymple
Joar Skalse
Yoshua Bengio
Stuart J. Russell
Max Tegmark
Sanjit A. Seshia
Steve Omohundro
Christian Szegedy
Alessandro Abate
Joseph Halpern
Clark Barrett
Ding Zhao
Ben Goldhaber
Nora Ammann

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-45
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-45.html

May 4, 2024

Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems

David Dalrymple [*1] Joar Skalse [*2] Yoshua Bengio [3] Stuart Russell [4] Max Tegmark [5] Sanjit A. Seshia [4] Steve Omohundro [6] Christian Szegedy [7] Alessandro Abate [2] Joe Halpern [8] Clark Barrett [9] Ding Zhao [10] Ben Goldhaber [11] Nora Ammann [12]

Abstract

Ensuring that AI systems reliably and robustly avoid harmful or dangerous behaviours is a crucial challenge, especially for AI systems with a high degree of autonomy and general intelligence. In this paper, we will introduce and define a family of approaches to AI safety, which we will refer to as guaranteed safe (GS) AI. The core feature of these approaches is that they aim to produce AI systems which are equipped with *high-assurance quantitative safety guarantees*. We outline a number of strategies for achieving this goal, describe the main technical challenges, and suggest a number of potential solutions. We also argue for the necessity of this approach to AI safety, and for the inadequacy of the main alternative approaches. Overall, despite a number of difficult technical challenges, GS AI offers a promising path for ensuring robust AI safety through formal methods.

1. Introduction

We introduce and define a family of approaches to AI safety, collectively referred to as *guaranteed safe* (GS) AI. These approaches aim to provide high-assurance quantitative guarantees about the safety of an AI system's behaviour through the use of four core components — a formal safety specification, a world model, a verifier, and deployment infrastructure. We will argue that this strategy is both promising and underexplored, and contrast it with other ongoing efforts in AI safety. We will also outline several ongoing avenues of research within the broader GS research agenda, identify some of their core difficulties, and discuss approaches for overcoming these difficulties.

Critical infrastructure and safety-critical systems require very high safety standards. For example, aircraft, nuclear power plants, and medical devices are subject to exceptionally rigorous safety certification. Moreover, it is plausible that there will soon be AI systems that are at least as safety-critical as these systems (e.g. as AI systems are increasingly deployed in safety-critical contexts with ever greater capabilities and autonomy), which means that they should be required to adhere to standards of safety that are at least as strict. Above some risk or capability thresholds, the burden of demonstrating such safety guarantees should be on the systems' developers. The provided evidence must be adequate to justify high confidence that the AI system is safe. We will argue that approaches based only on experimental tests are insufficient for producing such safety guarantees. Moreover, we will also argue that GS AI presents research avenues that could plausibly produce such safety guarantees in a satisfactory and tractable manner.

In Section 2, we provide the general background and context for this paper, including a brief overview of the AI safety problem and a way to classify AI systems based on their potential for dangerous behaviour. In Section 3, we discuss some avenues for producing stronger quantitative safety guarantees through the use of probabilistic safety assessments. In Section 4, we introduce and define GS AI, together with an extensive discussion of the spectrum of approaches that fall under this agenda, the main challenges to these approaches, and some potential solutions. In Section 5 we provide a further high-level discussion of our proposals and their feasibility, in Section 6 we list some related work, and in Section 7 we conclude the paper.

2. Background

In this section, we provide the background that is required to understand the rest of this paper and its context. This includes an overview of the AI safety problem and how to classify AI systems based on the level of risk they pose.

[*] Equal contribution [1] Advanced Research and Invention Agency [2] Oxford University [3] Mila - Quebec AI Institute / U. Montreal [4] UC Berkeley [5] Massachusetts Institute of Technology [6] Beneficial AI Research [7] x.AI [8] Cornell University [9] Stanford University [10] Carnegie Mellon University [11] FAR AI, Inc. [12] Alignment of Complex Systems. Correspondence to: David 'davidad' Dalrymple <david.dalrymple@aria.org.uk>, Joar Skalse <joar.skalse@cs.ox.ac.uk>.

2.1. The AI Safety Problem

A number of prominent AI experts have raised the concern that AI systems may pose a danger to humans and society. Some experts have argued that sufficiently advanced AI systems may threaten the survival of the human species, or lead to our permanent disempowerment, especially in the case of AI systems that are more intelligent than humans. Such concerns have been raised by Bostrom (2014), Tegmark (2018), Russell (2019; 2024), Pearl (2019), Bengio (2023), Metz (2023), Amodei (2023), and others. These experts have provided many different arguments in support of these concerns, which we will not be able to reproduce here in full. However, here are a few very brief summaries of some of their most central arguments:

1. For an AI system to solve a complex problem in an open-ended domain we must provide it with a formalisation of what it means to solve that problem. However, it appears to be very difficult to create such specifications. This issue has been observed empirically in current AI systems (e.g. Krakovna et al., 2020; Pan et al., 2021; Pang et al., 2023), and studied theoretically (e.g. Zhuang & Hadfield-Menell, 2020; Skalse et al., 2022; 2023; Karwowski et al., 2023; Skalse et al., 2024). This suggests that it is difficult to motivate AI systems to act in accordance with our intentions.

2. In a conflict of interests, greater intelligence is a substantial advantage. For example, in a chess game between a novice and a grandmaster, we should expect the grandmaster to win. More generally, the reason why humans are the most powerful species on the planet is primarily that humans are the most intelligent species. Technological innovation is also the greatest driver of economic growth and military capabilities. Thus, if there are ever AI systems that are substantially more effective at technological innovation than humans, and which are not aligned with human interests, then we should expect human interests to be marginalised.

3. Even if the goals of an AI system are specified correctly, it may still fail to internalise these goals in the intended way. For example, one way to maximise a reward signal that is provided by human feedback may not be to do what the humans wish, but rather to take control of the reward mechanism (Cohen et al., 2022). Similar phenomena have been observed empirically in current AI systems (Shah et al., 2022; Langosco et al., 2023), and studied theoretically (Hubinger et al., 2021).

4. Existing attempts to solve these problems have so far not yielded convincing solutions, despite rather extensive investigations (Ji et al., 2024). This suggests that the problem is fundamentally hard, on a technical level.

For a more complete and in-depth treatment of the arguments for why future AI may pose an existential risk to humanity, see e.g. Bostrom (2014); Russell (2019).

Other AI experts have also pointed to other, more immediate risks from AI systems. For example, generative AI may enable the spread of disinformation, by enabling the creation of convincing deepfakes or by making it cheaper and easier to produce large volumes of content (Brundage et al., 2018). AI may also carry out biased decision making that systematically disadvantages certain groups in unfair ways, even if this is against the wishes of the creators of that AI system (e.g., Kleinberg et al., 2016; Kim & Cho, 2022; Das et al., 2023; Wang et al., 2022). Recommender systems may facilitate invasions of privacy and the spread of extreme content (Stray, 2021; Carroll et al., 2022; Boxell et al., 2020; Settle, 2018; Lelkes et al., 2017). AI may also enable large-scale surveillance (Feldstein, 2019) or the centralisation of economic or political power (Brynjolfsson & Ng, 2023). For an overview of some of these issues, see, e.g., Memarian & Doleck (2023); Hendrycks et al. (2023); Brundage et al. (2018).

These two perspectives on the risks from AI are not mutually exclusive. Moreover, they both point to similar sets of technical challenges. The *AI safety problem* is the problem of ensuring that AI systems reliably and robustly act in ways that are not harmful or dangerous, including (but not limited to) cases where those AI systems are more intelligent than humans. In this paper, we are proposing a family of strategies for solving this problem.

Note that the problem of ensuring that AI is not harmful to humans comprises both a technical problem and a societal problem; solving the technical problems is not sufficient if the solutions are not globally implemented. In this paper, we will primarily focus on the technical aspect of the AI safety problem. For an overview of some of the political and sociological challenges, see e.g. Bostrom (2014); Alaga & Schuett (2023); Koessler & Schuett (2023); Sastry et al. (2024); Schuett et al. (2023).

2.2. AI Safety Levels

The levels of precaution that are appropriate for a given AI system depend on the capabilities of that system. To classify the relevant levels of capability, Anthropic has introduced a framework that they call AI Safety Levels (ASL)[1] as part of their voluntary safety commitments (Anthropic, 2023). This framework classifies AI systems into the following high-level categories:

1. ASL-1 refers to systems which pose no meaningful

[1] This is loosely modelled after the US government's biosafety level (BSL).

catastrophic risk.

2. ASL-2 refers to systems that show early signs of dangerous capabilities, such as the ability to give instructions on how to build bioweapons, but where the information is not yet useful due to insufficient reliability or to not providing information that a search engine could not. Many current language models appear to be ASL-2.

3. ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.

4. ASL-4 and higher (ASL-5+) are not yet defined in as much detail, but generally refer to human level and superhuman levels of intelligence, and thus qualitative escalations in the potential for catastrophic misuse and autonomy.

We refer to this classification scheme throughout the paper. Note that a similar classification scheme also has been defined by the Future of Life Institute (FLI, 2023).

3. Probabilistic Safety Assessment

Current approaches to validating frontier AI models prior to deployment lean on independent testing and red-teaming (as e.g. Ziegler et al., 2022; Perez et al., 2022). Such methods can find concrete examples of unsafe behaviour, which can then be rectified by the developer or lead to the decision to halt training and deployment until mitigating or alternative solutions are put in place. However, these testing regimes do not provide a rigorous, quantifiable safety guarantee: red-teamers could fail to find serious failures, while a model still harbours such failure modes. In the extreme case, a model could have been backdoored in a way that is cryptographically hard to detect without knowing the trigger (Guo et al., 2021). Even in the absence of such malfeasance, weaknesses in AI systems can remain undetected even after extensive testing and real-world usage. For example, ChatGPT was evaluated in great detail over a period of several months (OpenAI et al., 2024), and yet users found ways to circumnavigate its safety precautions within just a single day of it going public (Burgess, 2023). This issue is likely to be even more pertinent for more capable systems tasked with solving more difficult problems. Moreover, it is also important to note that AI systems often will be deployed in *adversarial* settings, where human actors (or other AI systems) actively try to break their safety measures. In such settings, empirical evaluations are likely to be inadequate — there is always a risk that an adversary will be more competent at finding dangerous inputs, unless you have a strong guarantee to the contrary.

We argue that to obtain a high degree of confidence in a system we need a positive safety case providing quantifiable guarantees, using either empirical or theoretical arguments. This is not unique to AI. About 25% of bridges built in the 1870s collapsed within the decade (McCullough, 2001), before a deeper theoretical understanding of civil engineering reduced it to less than 0.4% per decade (Cook, 2014). When probabilistic safety assessments are required from developers, this makes it possible for society to mandate a clear level of safety, in terms of, e.g., the frequency per year of adverse events of certain magnitudes (this is called a societal risk curve). At its simplest, this could consist of extensive testing: if an autonomous vehicle drives a million miles safely without intervention of a test driver, then we can conclude that the failure rate probably is less than one in a million miles when deployed in the same operating domain. However, note that such guarantees may only be valid within the bounds of unreasonable assumptions. For example, if human drivers start driving more aggressively around autonomous vehicles after they are more used to them, then this safety bound might cease to hold.

Empirical bounds may be obtained with weaker assumptions through methods such as adversarial testing, and/or testing in simulations with domain randomization. Even stronger bounds may be obtained through a more mechanistic understanding of the system. Fault trees (Nieuwhof, 1975) are a common safety engineering technique that allows for quantitative analysis, namely they can be interpreted as deduction in probabilistic logic. For example, if two redundant components can be shown to have a failure rate of $\leq n^{-1}$ and the failure rates are independent, the combined failure rate would be $\leq n^{-2}$. This could occur for example in an autonomous driving case when performing object detection on different sensors (e.g. LIDAR, vision, radar), or in a generative model case when using an ensemble of different models to detect malicious inputs. A similar approach might be applicable even in the case of a single monolithic model by leveraging approaches like mechanistic interpretability that seek to understand internal representations of the model (Zhang et al., 2021; Gao & Guan, 2023; Bricken et al., 2023; Michaud et al., 2024).

An alternative approach to obtaining stronger bounds relies on theoretical understanding of the system. For example, a rigorous theory for how deep networks generalise (as built towards by e.g. Kearns & Vazirani, 1994; Watanabe, 2009; 2018; Mingard et al., 2021; 2020) might enable principled extrapolation from empirical testing on a limited validation domain to a broader test domain. Combined, these approaches could enable carefully conducted empirical evaluations to provide substantially stronger safety bounds than exist for contemporary frontier models.

However, any empirical evaluation must ultimately rely on some relatively strong assumptions, such as the distribution of inputs used to validate the models being sufficiently similar to those they are deployed on. This makes it challenging for an empirical approach to rule out instances of deceptive alignment, where a system is acting to subvert the evaluation procedure (Hubinger et al., 2021). It also makes it challenging to give long-horizon safety guarantees, where the distribution of inputs is likely to naturally shift over time.[2] To achieve stronger safety guarantees, which will likely be needed for ASL-3 and beyond, we therefore expect it to be necessary to use a *model-based* approach. We discuss this approach in the following section.

4. Guaranteed Safe AI

In this section, we will introduce and characterise a family of approaches to the AI safety problem, which we refer to as guaranteed safe (GS) AI. We will first provide a definition of GS AI, together with a high-level overview. We will then discuss each of the core components of GS AI, namely a *world model*, a *safety specification*, a *verifier*, and *deployment infrastructure*. For each of the core components, we discuss its role in the overall architecture towards providing high-assurance safety guarantees, highlight key challenges in trying to implement these components, and discuss current approaches to overcoming those challenges.

4.1. Definition of GS AI

The core feature of the GS approach to AI safety is to produce systems consisting of an AI agent and other physical, hardware, and software components which together are equipped with a high-assurance quantitative safety guarantee, taking into account bounded computational resources. This can be contrasted against approaches to AI safety which primarily rely on empirical evaluations, or loose arguments based on qualitative or pre-theoretic intuitions.

A high-assurance quantitative safety guarantee may take the form of a formal proof that the system always will adhere to some safety specification[3] (or distribution over safety specifications) for all inputs, relative to a model (or a distribution over models) of world dynamics. Alternatively, it may be a reliable and sound upper bound on the probability of violating a safety specification. However, especially when we cannot find a proof certificate or formally define the desirable or undesirable behaviour, it may also take the form of an estimate of an upper bound on the probability of harm, with the estimate asymptotically converging with computational resources in order to guarantee a constraint on desirable behaviour.

Definition 4.1. A Guaranteed Safe AI system is one that is equipped with a quantitative safety guarantee that is produced by a (single, set of, or distribution of) **world-model(s)**, a (single, set of, or distribution of) **safety specification(s)**, a **verifier**, and **deployment infrastructure**, satisfying the following additional criteria:

1. The probabilistic specification encodes societal risk criteria, which should ideally be determined by collective deliberation;
2. The verifier provides a quantitative guarantee (in the form of a proof certificate, probabilistic bound, asymptotic guarantee, or other comparable assurance) that the AI system satisfies the specification with respect to the world model;
3. When deployed, multilaterally auditable and redundant verifiers and runtime monitors check for observations that invalidate the world-model, and control high-assurance failover to a verified backup system; and
4. All potential future effects of the AI system that could be relevant to the safety specification should be conservatively over-approximated by the world-model.

There is much to unpack in this definition. Before moving on, we will therefore provide a brief explanatory example of what each of the four core components of Definition 4.1 could look like, together with a motivation for their necessity. Later in this section, we will provide a more in-depth discussion of what each of these components may look like, what challenges they come with, and how those challenges may be overcome. An overview is also provided in Figure 1. But first, let us provide some intuition.

A safety specification corresponds to a property that we wish an AI system to satisfy. For example, we may wish that an AI system never takes any actions that may plausibly cause a human to be harmed. If we have a formal definition of harm, as well as a formal definition of causation, then this safety property could be turned into a well-defined formal specification. Of course, neither of these terms are easy to formalise, but proposals do exist (e.g. Beckers et al., 2022b;a; Halpern, 2016; Pearl, 2009). Other safety specifications may also be desirable. For example, we may wish to require that an AI system is "truthful", or that it can offer "explanations" for its actions, etc. In Section 4.3, we will provide a more detailed overview of some possible methods for obtaining safety specifications, as well as their advantages and challenges. Note that because of the difficulties

[2] Note that the usage of powerful AI is likely to itself create situations that are novel and unprecedented, which makes this a point of particular importance. Stated differently, we should *assume* that "distributional shift" will occur.

[3] In this paper, we use the term "safety" not in the strict sense used in formal methods (Alpern & Schneider, 1987) but in the broader sense used in AI for specifications of significance.

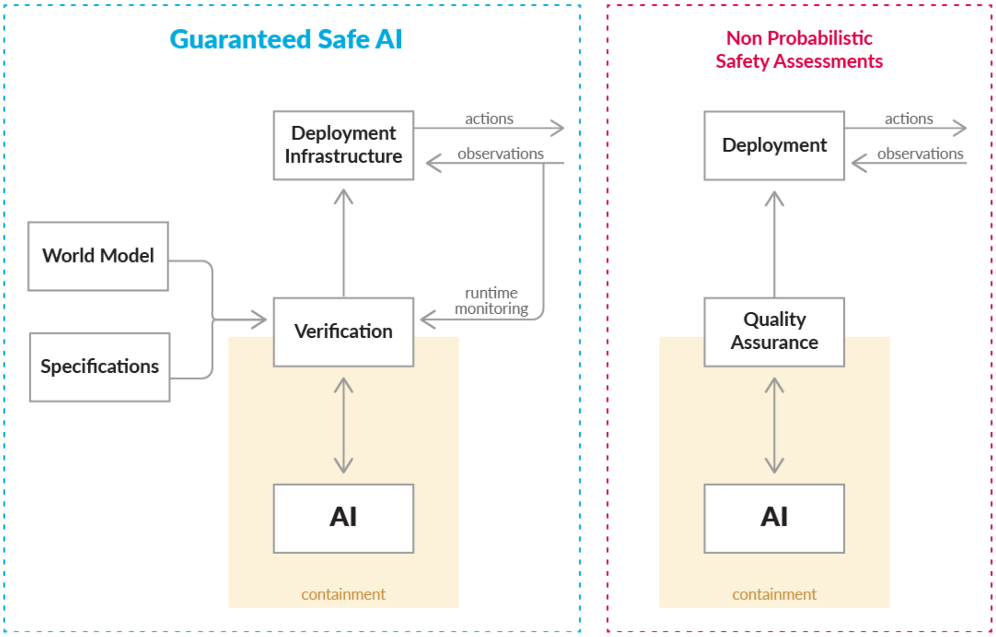

Figure 1. The GS AI approach builds on four components, namely a *world model* that describes the environment of the AI system, a *safety specification* that describes desirable safety properties and is expressed in terms of the world model, a *verifier* that provides a quantitative guarantee of the extent to which an AI system satisfies the safety specification, and *deployment infrastructure* which responds to runtime observations which invalidate world model predictions. Each of these components can in turn be created in a wide variety of ways, which gives rise to a spectrum of approaches within the GS AI agenda.

with formulating adequate safety specifications, it may be desirable to use multiple safety specifications, or to use *distributions* of safety specifications, obtained through some learning process (or otherwise).

Next, many desirable safety specifications necessarily require a *world model* (or distribution over world models) that describes the dynamics of the AI system's environment. For example, suppose we want to ensure that an AI system never takes any actions that leads a human to be harmed, according to some (possibly ambiguous) definition(s) of "harm". In order to do this, we need a model that describes if a given action is likely to lead a human to be harmed (in some context). More generally, without a world model we can only verify specifications defined over input-output relations, but it is often desirable to instead verify specifications over input-*outcome* relations. We may also want to define predicates such as "harm" in terms of not only directly observable quantities (like a human-provided label), but also in terms of unobserved or even counterfactual variables (such as how a group of wise humans would hypothetically judge an outcome). Of course, we must also be able to trust the correctness of this world model, which means that it ideally should be interpretable and understandable. For more discussion of how to create such a world model (or distribution of world models), see Section 4.2. Note that the world model need not be a "complete" model of the world — the level of detail and abstraction that is adequate depends on the safety specification and the AI system's context of use.

Given a safety specification and a world model, we also need a way to produce quantitative assurances for a given AI system. In the most straightforward form, this could take the shape of a formal proof that the AI system (or its output) satisfies the safety specification relative to the world model. This is akin to traditional formal verification (see e.g. Baier & Katoen, 2008; Leino, 2023; Seligman et al., 2023). Of course, such formal verification is often hard to produce, even for relatively simple computer programs. However, further progress in automated reasoning and theorem proving due to integration with data-driven learning (see e.g. Lample et al., 2022; Seshia, 2015; Trinh et al., 2024) could make this substantially easier, and might also scale with further progress in AI more generally. Moreover, if a direct formal proof cannot be obtained, there are weaker alternatives that would still produce a quantitative guarantee. For example, it may take the form of a proof that bounds the probability of failing to satisfy the safety specification, or a proof that the AI system will converge towards satisfying the safety specification (with increasing amounts of data or computational resources, for example). Indeed, many model-based AI algorithms have been designed to satisfy exactly these sorts of guarantees (McMahan et al., 2005; Junges et al., 2016; Mathews & Schmidler, 2022; Hasanbeig et al., 2023). For a more detailed discussion of these issues, see Section 4.4.

Finally, it is desirable to monitor the AI system at runtime, and check for signs that the world model is inaccurate. If such signs are detected, the AI system should not be allowed to continue execution. However, note that for a safety-critical system, it may not be possible (or desirable) to simply shut it down. For example, it would not be safe to disable the controller of a self-driving car while the car is in use. A GS AI system should therefore have a trustworthy backup system, appropriate for its domain of use, that can safely transition the system to a safe state if the main AI controller is disabled. These (and related) issues are discussed in Section 4.5.

Note that the GS AI approach remains agnostic about both how the "core" AI system was trained or produced, and about what containment or boxing methods (Armstrong et al., 2012) are used. If the verifier and the world model together can establish that it satisfies the safety specification, a quantitative safety guarantee is obtained (even if the core AI system remains uninterpretable, etc).

4.2. The World Model

In this section we will discuss the world model, including both a more detailed discussion of what counts as a world model, as well as some possible strategies for constructing such a world model (including their challenges and benefits).

The world model needs to answer queries about what would happen in the world as a result of a given output from the AI. It must also describe the state of the world at a level of granularity that is sufficient for expressing the safety specifications we are interested in.[4] World models also serve to elucidate the AI system designers' assumptions, and we must be mindful that those assumptions may hold only part of the time. Hence, the domain of applicability or epistemic uncertainty regarding different pieces of the world model must be represented and taken into account. For these reasons, the world model, or relevant aspects of it, should be auditable or monitorable at run time.

There are many possible strategies for creating world models. These strategies can roughly be placed on a spectrum, depending on how much safety they would grant if successfully implemented:

- Level 0: You have no world model, and, instead, assumptions about the world are implicit in training data and in aspects of the implementation of the AI system.
- Level 1: You use a trained black-box world simulator as your world model.
- Level 2: You use a machine-learnt generative model of probabilistic causal models, which you can test by checking whether it assigns sufficient credence to specific human-made models (such as e.g. models proposed in the scientific literature).
- Level 3: You use (a distribution over) world model(s) that are generated compositionally, potentially with the help of machine learning, from pieces that have been completely reviewed by humans.
- Level 4: You use (a distribution over) probabilistic causal model(s), potentially generated with the help of machine learning, that are fully audited by human domain experts.
- Level 5: You use world models about real world phenomena that are formally verified as sound abstractions of quantum field theory.
- Level 6: You have no world model, and instead use safety specifications defined over the entire set of all possible worlds.

To get a better intuition for what these levels may look like, let us discuss a number of potential approaches to constructing world models. First of all, in some cases, it may be feasible for engineers to manually create an adequate world model. This has been done in settings where the operating environment is known or controllable, for e.g. controllers in airplanes (Garion et al., 2022; Fremont et al., 2020a) and self-driving cars (Ivanov et al., 2020; Fremont et al., 2020b). Such models are also commonly used in scientific research, including epidemic simulators (Broeck et al., 2011) and particle physics simulators (Baydin et al., 2019b). For AI systems, probabilistic programs (e.g., (Milch et al., 2007; Fremont et al., 2022)) have been shown to be a promising formalism for world modeling. In principle, this manual approach is likely to provide the best understanding of the assumptions underlying the world model, as well as the delineation of its domain of applicability. This approach would produce a world model on Level 4 or 5. However, for AI systems that directly interface with very complex systems (such as e.g. human users, the world economy, or sensitive ecosystems), it may not be possible to create sufficiently accurate world models in a fully manual way.

In such cases, the world model must instead be machine learned (or automatically generated by some other means). One possible approach to creating an interpretable world

[4] Note that a world model with a more abstract state-space may make it easier to express certain safety specifications, or make verification more tractable, but that this may come at the cost of making the predictions less accurate. Also note that the world model need not make predictions about arbitrary properties of the world, and even about properties on which they do make predictions, these predictions need not necessarily be precise.

model with AI is to use large language models (or similar systems) to write probabilistic programs that correspond to the system(s) being modelled (as e.g. Wong et al., 2023; Elmaaroufi et al., 2024). This method has the potential to be scalable, since most of the work is offloaded to AI systems. Moreover, this method may also produce world models that are interpretable by default, if the language models are trained on (or at least prompted with) code written by human programmers, or if the language used to generate pieces of the world model is forced or encouraged to be interpretable (through probabilistic translation to and from natural language). If successful, this approach could produce a world model on Level 2, 3, 4, or possibly 5 in the classification above (depending on the extent to which the resulting model is audited). One of the main challenges with this approach would likely be to ensure that the world model has a high predictive accuracy.

Another approach is to learn a world model from data. However, this approach comes with a number of challenges. In particular, many machine learning methods are prone to being confidently incorrect in novel situations. On a theoretical level, this problem can be solved by Bayesian induction. Unfortunately, Bayesian induction is typically not computationally tractable (Hutter, 2003). However, it may be possible to use deep learning to tractably estimate Bayesian conditional probabilities, as e.g. Deleu et al. (2022); Ke et al. (2022); Deleu et al. (2023); Hollmann et al. (2023); Hu et al. (2023). If we consider an explanatory theory t as a latent variable corresponding to a world model, the plan is to train a large neural network to implicitly estimate and sample from the Bayesian posterior $P(t \mid D)$, where D is the training data. Such a network provides a world model, which can be used to verify the safety of other AI systems (by estimating the probability of harm conditional on running that system, or by estimating the probability of harm conditional on executing each output by that system, for example). Moreover, as we make those networks larger and train them for longer, we are in principle guaranteed that they will converge toward the Bayesian optimal answers. This means that we can continue training at run-time, or at least estimate the error made by the neural network through a sampling process. We can also decrease the risks associated with an insufficiently trained neural network by encouraging the AI-generated hypotheses to be somewhat human-readable, for example by regularising the AI to generate hypotheses that can be converted to natural language and back with as little error as possible. Note that this would allow human inspection of the generated theories, even if the neural net activations are not themselves interpretable. If successful, this approach would produce a world model on Level 2 (or 5, if the model passes additional checks). See Bengio (2024) for a more complete discussion of this approach. Other approaches to scalable Bayesian inference are also explored in works such as Gothoskar et al. (2023); Saad et al. (2023); Baydin et al. (2019a).

A very ambitious approach would be to use world models that are formally verified as being sound abstractions of the basic laws of physics. Note that while physics is not a completed field, there is good reason to believe that our current best theories are completely accurate in certain domains (see e.g. Carroll, 2021). We can therefore be confident that a system truly satisfies a specification if that specification is verified for that system relative to our best theories of physics, and that specification also includes the requirement that the system is not moved beyond the domains where our theories are known to be accurate. If successful, this would produce a world model on Level 5. One of the main challenges for this approach will be the immense computational complexity of producing and verifying such models. For more details, see Tegmark & Omohundro (2023).

Note that in many cases, it will likely be necessary for the world model to model the behaviour of humans. Moreover, humans are very complex, and it seems dubious to presuppose that it is possible to create a model of human behaviour that is both interpretable and highly accurate (especially noting that such a model itself would constitute an AGI system). This may introduce a fundamental trade-off between interpretability and predictive accuracy. However, note that interpretability can be maintained if human behaviour is modelled using *nondeterminism* (where "nondeterminism" here should be understood in the same sense as "nondeterministic automata" and "nondeterministic Turing machines", rather than as a synonym for "probabilistic"). The same applies to other highly complex systems. Also note that the levels in the classification above can be mixed within a single GS AI system. For example, a system can use stronger models of its engineering systems and weaker models of its interactions with the social domain. You may wish for the world model to be as rigorous as possible, which may vary depending on different domains (e.g., you may want to model your sensors with Level 5, but you will not be able to use that rigour for e.g. social phenomena).

A potential argument against the GS research agenda is that the world may be so complex that it is infeasible even in principle to create a sufficiently accurate world model (because of chaotic dynamics, etc). We have several responses to this point. First of all, a world model can (and should) of course include model uncertainty, and this uncertainty can be taken into account when the safety specifications are verified.[5] In this way, the strength of the resulting formal

[5] As a very simple example, suppose that the safety specification is given relative to a finite time horizon of n steps, and that we have reason to believe that the world model is wrong with probability at most ϵ per step over the first n steps. Then if a policy can be proven to satisfy this specification relative to the world model, we should believe that it will satisfy the specification with probability

guarantees will be appropriately sensitive to the reliability of the world model. Moreover, many of the most concerning loss-of-control scenarios with advanced AGI systems involve cases where the AGI is assumed to be able to generate and execute complicated plans with high reliability. In other words, it is reasonable to assume that if AI systems can be powerful enough to pose a serious danger to humanity, then it is possible to create sufficiently accurate world models. Finally, we would like to note that this argument essentially is fully general. *Any* strategy for creating safe AI systems must rely on some beliefs or assumptions about the real world, and these assumptions could be wrong. Stated differently, if it is impossible to create world models that are sufficiently accurate to ensure that a given AI system adheres to some safety specification, then it is presumably in general impossible to ensure that this safety specification is satisfied by that system.

4.3. The Safety Specification

In this section we will discuss the safety specification(s), the difficulties with creating such specifications, and some potential strategies for overcoming these difficulties.

Note that a safety specification in general is different from a reward function (though they include bounded reward functions as a special case). In particular, a safety specification may include properties defined by probabilistic temporal logics, causal counterfactual queries, or even hyperproperties, which reward functions cannot typically express (Seshia et al., 2018; Subramani et al., 2024). However, safety specifications cannot include unbounded evaluations of the world-state. For example, safety specifications could be expressed in terms of PCTL (Hansson & Jonsson, 1994), or in terms of conjunction of linear inequalities of reachability probabilities, for example. Also note that a specification logic can invoke neural components as predicates.

There are many possible strategies for creating safety specifications. These strategies can roughly be placed on a spectrum, depending on how much safety it would grant if successfully implemented. One way to do this is as follows:

- Level 0: No safety specification is used.
- Level 1: The safety of the system is evaluated by a pool of human judges.
- Level 2: The system uses a safety specification that is expressed in natural language but interpreted by a black-box AI system.
- Level 3: The system uses hand-written safety specifications for limited safety properties that are relatively tractable to express in a formal language.
- Level 4: The system uses a specification that is written in (probabilistic) logic at the top level, but which makes use of (uninterpreted) neural components to represent learned bindings of certain human concepts to real physical states.
- Level 5: The system uses compositional specifications that are made up of parts that are all human audited, but synthesised by AI.
- Level 6: The system uses hand-written safety specifications for comprehensive safety properties that require substantial effort to express formally.
- Level 7: The safety specification completely encodes all things that humans might want, in all contexts.

It is often very difficult to create useful formal safety specifications in many of the domains in which AI systems operate. Suppose, for example, that we want to ensure that a chatbot never gives advice that is “harmful”. How should this specification be formalised? A robust formalisation of this specification would require a very detailed world model, since whether or not a given piece of advice will turn out to be harmful may depend on diverse facts about the real world in complicated ways. Alternatively, we could instead require the AI to never give advice that it “believes” to be harmful. However, verifying this specification requires a reliable way of extracting “beliefs” from an AI system, which may be just as difficult. Moreover, “harm” is a vague predicate, in the sense that there are edge-cases where it is controversial whether a given person is harmed or not. Similar issues occur if we want to ensure that an AI system never lies, or that it always follows instructions from humans, etc.

Moreover, if a specification is formalised in the wrong way, then it can often be satisfied by some perverse and unintended behaviours. The field of moral philosophy has produced several formal frameworks that are meant to capture good conduct in humans. However, all of these frameworks recommend counterintuitive and seemingly perverse actions in at least some situations, and none of them are endorsed by a majority of all moral philosophers (Bourget & Chalmers, 2023). We should therefore expect this problem to be hard.

In many cases, it is possible to find *proxies* for complex predicates (such as “harm”) which are easier to define and measure. However, while such a proxy may robustly correlate with our intuitive judgements of harm in normal situations, they may still reliably come apart if those proxies are used as an optimisation target. This phenomenon is known as *Goodhart’s law*, which is an informal principle sometimes stated as “when a measure becomes a target, it ceases to be a good measure”. Goodhart’s law was first introduced by Goodhart (1975), and has since been studied more formally in works such as Manheim & Garrabrant (2019); Hennessy

at least $(1-\epsilon)^n$ in the real world.

& Goodhart (2023); Zhuang & Hadfield-Menell (2020); Skalse et al. (2022); Karwowski et al. (2023). This means that a safety specification may need to be highly accurate in order to remain robustly reliable, especially when applied to a powerful AI system.

One strategy is to attempt to *learn* safety specifications from data. This is explored by the field of *reward learning*, where the specification is assumed to have the form of a reward function. If we assume that a human's preferences can be captured by a reward function, and if we can learn a representation of this reward function from some data source, then we may be able to prove that a given AI system will respect the preferences which are embodied by that reward function. However, reward learning also faces serious difficulties. In particular, most data sources are insufficient for identifying the underlying reward function uniquely, even in the limit of infinite data, and this irreducible ambiguity may be problematic (Ng & Russell, 2000; Dvijotham & Todorov, 2010; Cao et al., 2021; Kim et al., 2021; Skalse et al., 2023; Schlaginhaufen & Kamgarpour, 2023). Moreover, many reward learning algorithms are highly sensitive to the modelling assumptions they make about their data source (Armstrong & Mindermann, 2018; Freedman et al., 2020; Viano et al., 2021; Skalse & Abate, 2023; 2024). Finally, the learnt reward model is itself typically not interpretable, which is a serious issue (see e.g. Michaud et al., 2020; Jenner & Gleave, 2022). This means that much work is required before reward learning can be a reliable source of specifications. For an overview, see Casper et al. (2023). Also note that there are approaches to learning safety specifications which do not fall under the umbrella of reward learning, see e.g. Bengio (2024).

Another approach is to attempt to create "conservative" safety predicates which aim to be *sufficient* (but not *necessary*) for safety. For example, we may require that the AI has no incentive to influence any part of the external world, or to find out any information about the external world (e.g. Armstrong et al., 2012; Armstrong & O'Rorke, 2018; Armstrong & O'Rourke, 2018; Everitt et al., 2021; van Merwijk et al., 2022). Alternatively, we may attempt to create safety predicates which make an AI system more safe, even if they do not *ensure* safe behaviour (as e.g. Soares et al., 2015; Orseau & Armstrong, 2016; Hadfield-Menell et al., 2017, etc). These approaches come with their own challenges, see e.g. Bostrom (2014) and the above cited works. Moreover, while certainly challenging, it may still be feasible to simply specify the safety predicates manually. This is (partially) attempted in works such as e.g. Beckers et al. (2022a;b).

It is also important to note that the task of formalising specifications can be easier with a *system-level* approach. As advocated by Seshia et al. (2022), even when individual AI components perform tasks that are hard to formalise, it can still be the case that the relevant notion of safety can be formalised at the full system level. For example, one can formalise system-level safety for autonomous vehicles even when it is not possible to formalise correctness for object detection and classification components in the autonomy stack (Dreossi et al., 2019). In some cases, this can make it feasible to define specifications directly.

Another approach to creating strong safety specifications is described by Tegmark & Omohundro (2023). The authors propose that the actions of potentially dangerous systems should be mediated through "provable gatekeepers" which enforce safety rules. The core component of these gatekeepers is a device they call a "provable contract". Each provable contract is responsible for a piece of critical infrastructure, and is equipped with a number of formal safety constraints. It takes as input a control program for its infrastructure, together with a proof that the control program obeys its safety constraints. It then only runs the control program if the proof is valid. Unlike traditional design-time application of formal methods, this approach may involve the generation and checking of mathematical proofs as part of the operation of the system. This enables a vastly richer means of control and trust between parties. In general, untrusted AIs may still be used to solve problems, create safe software and hardware designs, and safe contracts and interactions. In this way powerful but untrusted AIs can be used to reliably create a safe and trusted infrastructure.

Another strategy for creating safety specifications is provided by *Cooperative Inverse Reinforcement Learning* (CIRL), as described by Hadfield-Menell et al. (2024). CIRL formulates the interaction between an AI and a human as a two-player Markov game (with one player being the AI, and one player being the human). Both players have the same reward function, but only the human knows what the reward function is. This means that the two players must cooperate to obtain a high reward, and that the AI system must listen to feedback from the human. Optimal solutions to the CIRL game produce behaviours such as active teaching, active learning, and communicative actions. The idea is that this problem formulation will encourage the AI to be *corrigible* (in the sense of Soares et al., 2015), instead of following some goal dogmatically. An AI could be verified to adhere to various safety specifications within the CIRL game. These specifications could take several forms, but a natural choice would be to require that the AI is *provably beneficial* to the human (or some variation thereof). Note that such specifications may require the world model to also make modelling assumptions about the behaviour of the human (and in particular about how the behaviour of the human relates to its preferences). Also note that many of the challenges to reward learning also apply to CIRL.

While these challenges are serious, it is important to note that most approaches to AI safety require formal safety specifications (or at least formalisations of what an AI system should be optimised for). These difficulties are thus not unique to the GS research agenda. It is also important to note that further development in AI capabilities will tend to make it easier to create good safety specifications. For example, AI systems could be used to suggest new specifications, to critique proposed specifications, or to generate examples of cases where two candidate specifications differ. Progress in AI could thus also accelerate the creation of good safety specifications.

4.4. The Verifier

In this section we will discuss the verifier, the difficulties with creating such a verifier, and some potential strategies for overcoming these difficulties.

The verifier may produce different kinds of formal guarantees, depending on what is feasible in a given context. We can thus place different kinds of verifiers on a spectrum, based on the strength of their corresponding guarantees:

- Level 0: No quantitative guarantee is produced.
- Level 1: A heuristic assurance is given by ad-hoc empirical testing of the AI system.
- Level 2: A standardised set of tests is used, which thus provides safety assurances that are auditable and comparable across systems.
- Level 3: You use a property-based test which includes domain randomisation and current state-of-the-art automated evaluations. In other words, you have some template for what the test looks like, but randomly populate the template to get more coverage.
- Level 4: You use black-box fuzzing, wherein an automated tool gives test vectors as input to the system and, depending on the response, generates different test vectors to fool the system. In other words, you use a form of automated red-teaming.
- Level 5: You use white-box fuzzing, which is akin to Level 4, except that your tool not only looks at the input-output behaviour of the AI system, but also considers its internal states and tries to make certain internal structures flip to get more coverage.
- Level 6: You use probabilistic inference with asymptotic convergence. This is akin to Level 5, but with the additional guarantee that if the evaluation system would run forever, then it would eventually literally cover every possible input to the system.
- Level 7: You combine asymptotic coverage with white-box fuzzing. This might include adversarial gradient optimisation, whereby you first cover areas where you most expect safety concerns to spring up, and where the system in the limit of infinite time would cover every possible system input.
- Level 8: Akin to Level 7, but with the additional requirement that you have some non-asymptotic convergence bounds, thanks to some formula for how much of the total state space your system covers at a given time. You can thus run it for a finite amount of time, and know how much is left that you have not covered
- Level 9: You have a sound bound on the probability of failure, meaning that the true probability of something happening is less than or equal to the value established by your verifier. This includes the case where the verifier is able to establish that the probability of failure is 0 (relative to the world model).
- Level 10: Akin to level 9, but with the additional requirement that the proof is concise enough that humans can read, understand, and check it.

To obtain strong guarantees, we would need a verifier on Level 8, 9, or 10. This, of course, raises the question of whether such guarantees could feasibly be obtained for AI systems. Even for relatively simple properties of software, obtaining such guarantees currently has a very high burden of highly specialised cognitive labour. However, it may be possible to train AI systems without ASL-4 capabilities to automate much of this labour at a near-expert level of sophistication without raising significant safety concerns. For example, consider an AI agent that only interacts with a formal theorem-prover and is only allowed to grow a library of formal facts. In this case the agent will never be exposed to any actions that would meaningfully influence the real world, and it would never learn to communicate with humans besides formalising and reasoning about existing statements or developing related theories. Note that the scope of that training would be very limited, so the AI is only exposed to natural language mathematics and computer science. It will therefore not have any significant knowledge of the outside world, besides these restricted domains. This restricts the AI's potential for power-seeking behaviour.

Here we outline a potential system that could acquire human-expert-level reasoning capabilities by letting it process all of the mathematical and computer science literature automatically by formalising it (Szegedy, 2020). We first index all of the existing informal and formal literature with an ASL-3 system and create a neural retrieval system. Then we use similarly capable language models to extract statements and definitions from the text and formalise it. Such models already exist, but they need further tuning on existing formalisation to reach sufficient quality for the initial phase of a bootstrap loop. The informal to formal translation data for this phase can be generated using cycle-consistency training (Lample et al., 2017), along with the small amount of human-verified formalization of informal statements, just like the proof-generation part can be trained on a relatively large corpus of existing formal theorems. The next phase is a reinforcement learning loop that trains the translation model and the neural prover in lockstep; In this step, both models can be assumed to be retrieval-augmented large language models, while the translation takes in informal (natural language) specifications it outputs formal specification. Then the prover attempts to prove them, potentially utilising the natural language proof- sketches. Whenever a statement is proven to be correct (that is, verified by the formal prover), and it is useful for proving other statements like its informal counterpart, it can be considered to be a correct transcription and used for training the translation model. In addition the prover can be trained on the successful proof traces. While there is ongoing research work on the overall feedback loop, there are multiple existing proof-of-concept solutions for various components of this reinforcement learning system (Szegedy, 2020), which makes it likely that a full solution can be implemented in the coming years.

A more extensive discussion on the challenges for extending formal methods to handle the unique characteristics of AI systems is provided by Seshia et al. (2022), who presents fifteen principles for providing provable guarantees of safety for AI systems. We summarise some of the key ideas pertaining to the verifier here, and point the reader to (Seshia et al., 2022) for details. First of all, one important piece is devising suitable *abstractions of AI components* such as deep neural networks. Seshia et al. (2022) advocates for designing abstractions of AI components that are easier to formally analyse and which can be generated algorithmically. In this context, abstractions that are modular and more interpretable have also been found to be easier to analyse by formal verification tools. Another key part is compositional reasoning; the idea is to construct a proof of safety of the overall AI system by decomposing the top-level proof obligation into sub-obligations on individual components, and then apply formal verification to the sub-obligations. One challenge here is that not all AI components have formal specifications – in other words, we need to do compositional verification without compositional specification! Seshia et al. (2022) tackle this by developing techniques to automatically infer a decomposition of the system into components along with *interface contracts* between them. Such a decomposition could be spatial and/or temporal, and has been shown to aid in scalable analysis (Dreossi et al., 2019; Yalcinkaya et al., 2023). Finally, it is crucial to integrate formal methods into the design process for AI components, for example, the training of deep learning components, and also to connect design-time formal methods with run-time assurance. Ideas from verified inductive synthesis of programs can be useful in training AI components with provable guarantees (e.g. Dreossi et al., 2018; Abate et al., 2023). Similarly, environment assumptions generated during world modelling at design time can be synthesised into efficient runtime monitors to supervise the operation of the AI at runtime (e.g. Torfah et al., 2022). Such runtime monitors form gatekeepers ensuring that AIs are only run in environments where their safety can be guaranteed (and otherwise transfer control to the backup system).

4.5. The Deployment Infrastructure

In this section we will discuss the deployment infrastructure, and a few different ways in which such a system may be constructed.

The deployment infrastructure is the hardware and software context into which the verified outputs of the AI system are deployed. The "verified AI outputs" may refer to individual actions or entire AI systems (that were themselves verified as safe). The purpose of the deployment infrastructure is to grant additional safety by enabling a rapid automatic response to observations that invalidate the predictions of the world model (resulting from unknown unknowns or unexpected uncertainty). Furthermore, the deployment systems can shorten the time horizon of outputs that need to be verified, thereby reducing the complexity of verification and increasing its tractability. The deployment infrastructure may contain a backup system which can be used to transition the overall system into a safe state if an observation is detected that invalidates the world model. As with the other components of GS AI, there are different ways to design the backup system, and these designs may be placed on a spectrum along the (continuous) time horizon at which the verified AI output is deployed:

- Level 0: The AI system is deployed to operate autonomously with an infinite time horizon.
- Level 1: The AI system is deployed for a finite time horizon, and if at the end of that time horizon it is not redeployed, it shuts down or transitions to a safe mode.
- Level 2: The deployment is limited to a single action at the time. If no new action is deployed, the system shuts down or transitions to a safe mode.
- Level 3: The verified AI output is itself a small AI system which autonomously responds to observations and takes action (output system with finite time horizon using some tamper proof clock).

It is worth noting that in cases where a system is safety critical, it may not be possible to simply turn it off entirely. In these cases, the system instead requires a safe mode backup which it can safely transition into. Such backup systems may be created in different ways, depending on the GS AI system's context of use. Also note that GS AI systems themselves may be useful for creating bug-free software and tamper-proof hardware that could be used to produce safe backup systems.

5. Discussion

We have argued that safety-critical AI systems need to be equipped with quantitative safety guarantees, and that empirical evaluations alone are insufficient for producing such guarantees with an adequate level of assurance. We have also introduced GS AI as a potential avenue for tractably obtaining such safety guarantees, discussed the main challenges with this research agenda, and some ways in which those challenges may be overcome. In this section, we will provide a further discussion of some potential benefits to the GS AI research agenda.

An important benefit to GS AI is that it makes democratic oversight easier, because concrete safety specifications can be audited and discussed by outside observers and regulators. AI technology will have an immense (and potentially unprecedented) impact on most, if not all, areas of life. This makes it crucial to enable as many people as possible to have a say in how this technology is deployed and used. However, such democratic input and oversight is not possible if important AI systems are specified, trained, and evaluated using procedures that are opaque to (or kept secret from) the wider public. In other safety-critical industries, probabilistic safety assessment means that the developer must specify all the assumptions or premises needed to deduce that their system meets the required societal risk thresholds. These assumptions can then be challenged by the regulator (and society at large) if they are not deemed socially acceptable. GS AI enables the same kind of oversight for AI systems.

Another important benefit of the GS research agenda is that it may produce AI safety solutions whose costs are amortised over time. Any potential safety measure (in AI or elsewhere) faces the issue that if said safety measure is costly to implement, then there is an incentive to disregard it. This reduces the value of approaches to AI safety that impose a high "safety tax". For example, a comprehensive suite of rigorous empirical evaluations may be expensive and time-consuming to carry out, and this cost would (presumably) have to be paid again for each new AI system that is created. This would in turn create an incentive to save resources by cutting corners. By contrast, some approaches to GS AI may allow for most of these costs to be amortised, which may substantially reduce the incentive to disregard the corresponding safety measures. Once satisfactory safety specifications have been identified and scalable methods for formal verification have been developed, new AI systems could likely be verified against these safety predicates at a much lower marginal cost. Such solutions would also be more scalable in view of the impressive speed of AI progress that otherwise threatens to outpace progress in safety.

It is important to note that GS AI may not be the only method for producing AI systems with verifiable quantitative safety guarantees. For example, another potential approach is to extract interpretable policies from black-box algorithms via automated mechanistic interpretability and directly proving safety guarantees about these policies. This approach differs from GS AI in that it does not make use of a world model that is separate from the policy; instead, it requires that the policy can itself be made interpretable. This strategy may be easier if it is intractably difficult to create a sufficiently good world model or adequate methods for doing formal verification relative to that world model. However, it may also be more difficult, especially because the policy may be more complex than the world model. For example, the rules of chess are less complex than a policy which is good at chess, and it is much easier to specify the axioms of Euclidean geometry than it is to specify a computer program that is good at proving theorems about Euclidean geometry. In a similar way, it may be much easier to create an interpretable world model than to create a performant interpretable policy. However, it is ultimately an empirical question whether it is easier to create interpretable world models or interpretable policies in a given domain of operation.

We also want to emphasise that there is a spectrum of approaches for safety assessments, ranging from easy ones that provide weaker safety assurances (such as evaluations and red teaming) to more expensive procedures that provide stronger safety guarantees (such as the approaches within the GS AI spectrum). Given the uncertainty about when each of the ASL safety levels will be crossed, we need an "anytime" portfolio approach of R & D efforts spanning this spectrum. This will allow us to maximise the expected effectiveness of the feasible safety techniques at each stage. This would involve investing in cheaper techniques, such as empirical evaluations, but also more ambitious approaches, such as those presented by the GS AI framework. It is also important to note that quick partial successes in GS AI are both plausible and useful. For example, provably compliant cybersecurity, geofencing, remote kill-switches, verified sensors and actuators, etc, may all be feasible even if the most ambitious proposals within the GS AI agenda are not, and would still provide notable benefits in their own right.

6. Relevant Existing Work

In this section, we briefly provide an overview on some existing work that GS AI approaches build on, or which are otherwise relevant to or related to the GS AI agenda.

First of all, the field of *computational learning theory* (CLT) is concerned with the mathematical analysis of learning algorithms and learning problems. These investigations have produced various formal guarantees for large classes of learning algorithms, primarily in the form of generalisation guarantees and regret bounds. These bounds typically show that a given learning algorithm under some given circumstances is guaranteed to attain a given level of performance with at least some given probability, and may also show how these probabilities scale in terms of the amount of training data, etc (see e.g. Kearns & Vazirani, 1994). Like the GS approach to AI safety, CLT is also concerned with deriving formal guarantees for AI systems. However, CLT is typically concerned with guarantees concerning narrow performance metrics, whereas we are concerned with specifications that would provide strong safety assurances for advanced systems operating in open-ended environments.

There is also a large literature on formal verification of neural networks and AI-enabled systems. This literature has proposed a range of algorithms which can be used to ensure that a given neural network (potentially belonging to some restricted class) satisfies a given specification (typically specified over its interfaces), or that a system containing one or more neural components satisfies a safety specification. It would not be possible for us to provide a comprehensive overview of these algorithms here, but for an introduction, see e.g. Seshia et al. (2022); Albarghouthi (2021).

Another area of computer science that is relevant to GS AI is the literature on *correct-by-construction* program synthesis. This is an approach to software development that aims to produce programs that are guaranteed to be correct with respect to their specifications from the outset, rather than relying on traditional testing and debugging methods to identify and fix errors after the fact. The core idea behind correct-by-construction synthesis is to use formal methods to systematically construct programs that satisfy a given set of formal specifications or correctness properties. This is typically achieved by encoding the specifications as logical formulas or constraints, and then using automated solvers or synthesis algorithms to derive programs that provably satisfy those constraints. For an overview, see e.g. Gulwani et al. (2017); Jha & Seshia (2017); Edwards et al. (2023).

More broadly, GS AI is also related to the field of safety engineering, which is an area of engineering science that focuses on identifying, evaluating, and mitigating potential hazards and risks associated with various systems and processes. For an overview of the techniques used in this field, see e.g. Ericson (2015); Leveson (2012); Dhillon (2003).

7. Conclusion

In this paper, we have introduced and defined the concept of guaranteed safe (GS) AI. GS AI aims to ensure the safety of AI systems by equipping them with formal, verifiable safety guarantees. We have argued that the GS approach is necessary, given the limitations of other methods such as empirical testing and interpretability.

We have also acknowledged that GS AI faces serious technical challenges. Creating accurate and interpretable world models, formulating precise safety specifications, and performing formal verification at scale are all difficult problems. However, we have suggested potential strategies and research directions for making progress on these problems.

Overall, we believe the GS agenda is crucial for ensuring robust and reliable AI safety in advanced AI systems. While empiricism and transparency are useful tools, they do not provide the strong safety assurances that formal verification can. And although formal verification is challenging, the GS research program offers a promising path toward making it feasible at scale.

Much work remains to fully develop the GS approach. But given its importance for avoiding AI risks, we argue that the GS agenda deserves substantially more attention and resources than it currently receives. With a concerted research effort on the core technical problems, significant progress could be made. We hope this paper provides a useful starting point and motivation for a wider pursuit of the GS program.

8. Acknowledgements

We would like to thank Anca Dragan, Adam Gleave, Zac Hatfield-Dodds, and Joshua Tenenbaum for their valuable contributions and feedback on the idea and paper.

References

Abate, A., Edwards, A., Giacobbe, M., Punchihewa, H., and Roy, D. Quantitative verification with neural networks. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPICS.CONCUR.2023.22. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.CONCUR.2023.22.

Alaga, J. and Schuett, J. Coordinated pausing: An evaluation-based coordination scheme for frontier ai developers, 2023.

Albarghouthi, A. Introduction to neural network verification, 2021.

Alpern, B. and Schneider, F. B. Recognizing safety and liveness. *Distributed Comput.*, 2(3):117–126, 1987.

Amodei, D. Written testimony of Dario Amodei before the U.S. Senate Committee on the Judiciary, Subcommitee on Privacy, Technology, and the Law, 2023. URL https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf.

Anthropic. Anthropic's responsible scaling policy, 2023. URL https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf.

Armstrong, S. and Mindermann, S. Occam's razor is insufficient to infer the preferences of irrational agents. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31, pp. 5603–5614, Montréal, Canada, 2018. Curran Associates, Inc., Red Hook, NY, USA.

Armstrong, S. and O'Rorke, X. Good and safe uses of ai oracles, 2018.

Armstrong, S. and O'Rourke, X. 'indifference' methods for managing agent rewards, 2018.

Armstrong, S., Sandberg, A., and Bostrom, N. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines*, 22:299–324, 2012.

Baier, C. and Katoen, J.-P. *Principles of Model Checking*. The MIT Press, 2008. ISBN 026202649X.

Baydin, A. G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., Munk, A., Naderiparizi, S., Gram-Hansen, B., Louppe, G., Ma, M., Zhao, X., Torr, P., Lee, V., Cranmer, K., Prabhat, and Wood, F. Etalumis: bringing probabilistic programming to scientific simulators at scale. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '19, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450362290. doi: 10.1145/3295500.3356180. URL https://doi.org/10.1145/3295500.3356180.

Baydin, A. G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., Munk, A., Naderiparizi, S., Gram-Hansen, B., Louppe, G., Ma, M., Zhao, X., Torr, P., Lee, V., Cranmer, K., Prabhat, and Wood, F. Etalumis: bringing probabilistic programming to scientific simulators at scale. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '19, New York, NY, USA, 2019b. Association for Computing Machinery. ISBN 9781450362290. doi: 10.1145/3295500.3356180. URL https://doi.org/10.1145/3295500.3356180.

Beckers, S., Chockler, H., and Halpern, J. A causal analysis of harm. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 2365–2376. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/100c1f131893d3b4b34bb8db49bef79f-Paper-Conference.pdf.

Beckers, S., Chockler, H., and Halpern, J. Y. Quantifying harm, 2022b.

Bengio, Y. Written testimony of Yoshua Bengio before the U.S. Senate Committee on the Judiciary, Subcommitee on Privacy, Technology, and the Law, 2023. URL https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_bengio.pdf.

Bengio, Y. Towards a cautious scientist AI with convergent safety bounds, 2024. URL https://yoshuabengio.org/2024/02/26/towards-a-cautious-scientist-ai-with-convergent-safety-bounds/.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 9780199678112. URL https://books.google.com/books?id=7_H8AwAAQBAJ.

Bourget, D. and Chalmers, D. J. Philosophers on philosophy: The 2020 philpapers survey. *Philosophers' Imprint*, 23 (1), 2023. doi: 10.3998/phimp.2109.

Boxell, L., Gentzkow, M., and Shapiro, J. M. Cross-country trends in affective polarization. Working Paper 26669, National Bureau of Economic Research, January 2020. URL http://www.nber.org/papers/w26669.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, pp. 2, 2023.

Broeck, W. V. d., Gioannini, C., Gonçalves, B., Quaggiotto, M., Colizza, V., and Vespignani, A. The gleamviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infectious Diseases*, 11(1):37, Feb 2011. ISSN 1471-2334. doi: 10.1186/1471-2334-11-37. URL https://doi.org/10.1186/1471-2334-11-37.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hEigeartaigh, S. O., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., and Amodei, D. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018.

Brynjolfsson, E. and Ng, A. Big ai can centralize decision-making and power, and that's a problem. *Missing links in ai governance*, 65, 2023.

Burgess, M. The hacking of chatgpt is just getting started, Apr 2023. URL `wired.com/story/chatgpt-jailbreak-generative-ai-hacking/`. [Online; posted 12-April-2023].

Cao, H., Cohen, S. N., and Szpruch, L. Identifiability in inverse reinforcement learning. *arXiv preprint*, arXiv:2106.03498 [cs.LG], 2021.

Carroll, M., Dragan, A., Russell, S., and Hadfield-Menell, D. Estimating and penalizing induced preference shifts in recommender systems, 2022.

Carroll, S. M. The quantum field theory on which the everyday world supervenes, 2021.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

Cohen, M., Hutter, M., and Osborne, M. Advanced artificial agents intervene in the provision of reward. *AI magazine*, 43(3):282–293, 2022.

Cook, W. Bridge failure rates, consequences, and predictive trends. 2014. URL `https://api.semanticscholar.org/CorpusID:107360532`.

Das, S., Stanton, R., and Wallace, N. Algorithmic fairness. *Annual Review of Financial Economics*, 15 (Volume 15, 2023):565–593, 2023. ISSN 1941-1375. doi: https://doi.org/10.1146/annurev-financial-110921-125930. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-financial-110921-125930`.

Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., and Bengio, Y. Bayesian structure learning with generative flow networks, 2022.

Deleu, T., Nishikawa-Toomey, M., Subramanian, J., Malkin, N., Charlin, L., and Bengio, Y. Joint bayesian inference of graphical structure and parameters with a single generative flow network, 2023.

Dhillon, B. *Engineering Safety: Fundamentals, Techniques, And Applications*. Series On Industrial And Systems Engineering. World Scientific Publishing Company, 2003. ISBN 9789813102361. URL `https://books.google.co.uk/books?id=P_E7DQAAQBAJ`.

Dreossi, T., Ghosh, S., Yue, X., Keutzer, K., Sangiovanni-Vincentelli, A., and Seshia, S. A. Counterexample-guided data augmentation. In *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Dreossi, T., Donzé, A., and Seshia, S. A. Compositional falsification of cyber-physical systems with machine learning components. *Journal of Automated Reasoning*, 63(4): 1031–1053, 2019.

Dvijotham, K. and Todorov, E. Inverse optimal control with linearly-solvable MDPs. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 335–342, Haifa, Israel, June 2010. Omnipress, Madison, Wisconsin, USA.

Edwards, A., Peruffo, A., and Abate, A. A general verification framework for dynamical and control models via certificate synthesis, 2023.

Elmaaroufi, K., Shankar, D., Cismaru, A., Vazquez-Chanlatte, M., Sangiovanni-Vincentelli, A., Zaharia, M., and Seshia, S. A. Generating probabilistic scenario programs from natural language. *ArXiV e-Prints*, 2024.

Ericson, C. *Hazard Analysis Techniques for System Safety*. Wiley, 2015. ISBN 9781118940389. URL `https://books.google.co.uk/books?id=cTikBgAAQBAJ`.

Everitt, T., Carey, R., Langlois, E., Ortega, P. A., and Legg, S. Agent incentives: A causal perspective, 2021.

Feldstein, S. *The global expansion of AI surveillance*, volume 17. Carnegie Endowment for International Peace Washington, DC, 2019.

FLI. AI governance scorecard and safety standards policy, 2023. URL `https://futureoflife.org/wp-content/uploads/2023/11/FLI_Governance_Scorecard_and_Framework.pdf`.

Freedman, R., Shah, R., and Dragan, A. Choice set misspecification in reward inference. In *IJCAI-PRICAI-20 Workshop on Artificial Intelligence Safety*, 2020. doi: 10.48550/ARXIV.2101.07691. URL `https://arxiv.org/abs/2101.07691`.

Fremont, D. J., Chiu, J., Margineantu, D. D., Osipychev, D., and Seshia, S. A. Formal analysis and redesign of a neural network-based aircraft taxiing system with VerifAI. In *32nd International Conference on Computer Aided Verification (CAV)*, July 2020a.

Fremont, D. J., Kim, E., Pant, Y. V., Seshia, S. A., Acharya, A., Bruso, X., Wells, P., Lemke, S., Lu, Q., and Mehta, S. Formal scenario-based testing of autonomous vehicles: From simulation to the real world. In *23rd IEEE International Conference on Intelligent Transportation Systems (ITSC)*, September 2020b.

Fremont, D. J., Kim, E., Dreossi, T., Ghosh, S., Yue, X., Sangiovanni-Vincentelli, A. L., and Seshia, S. A. Scenic: A language for scenario specification and data generation. *Machine Learning Journal*, 2022.

Gao, L. and Guan, L. Interpretability of machine learning: Recent advances and future prospects, 2023.

Garion, C., Hattenberger, G., Pollien, B., Roux, P., and Thirioux, X. Formal Verification for Autopilot - Preliminary state of the art. Technical report, ISAE-SUPAERO ; ONERA – The French Aerospace Lab ; ENAC, March 2022. URL https://hal.science/hal-03255656.

Goodhart, C. Problems of monetary management: the UK experience in papers in monetary economics. *Monetary Economics*, 1, 1975.

Gothoskar, N., Ghavami, M., Li, E., Curtis, A., Noseworthy, M., Chung, K., Patton, B., Freeman, W. T., Tenenbaum, J. B., Klukas, M., and Mansinghka, V. K. Bayes3d: fast learning and inference in structured generative models of 3d objects and scenes, 2023.

Gulwani, S., Polozov, O., and Singh, R. Program synthesis. *Foundations and Trends® in Programming Languages*, 4(1-2):1–119, 2017. ISSN 2325-1107. doi: 10.1561/2500000010. URL http://dx.doi.org/10.1561/2500000010.

Guo, W., Tondi, B., and Barni, M. An overview of backdoor attacks against deep neural networks and possible defences, 2021.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. The off-switch game, 2017.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. Cooperative inverse reinforcement learning, 2024.

Halpern, J. Y. *Actual Causality*. MIT Press, Cambridge, MA, 2016. ISBN 978-0-262-03502-6. doi: 10.7551/mitpress/9780262035026.001.0001.

Hansson, H. and Jonsson, B. A logic for reasoning about time and reliability. *Form. Asp. Comput.*, 6(5): 512–535, sep 1994. ISSN 0934-5043. doi: 10.1007/BF01211866. URL https://doi.org/10.1007/BF01211866.

Hasanbeig, H., Kroening, D., and Abate, A. Certified reinforcement learning with logic guidance. *Artificial Intelligence*, 322(C):103949, 2023. doi: 10.1016/j.artint.2023.103949.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks, 2023.

Hennessy, C. A. and Goodhart, C. A. E. Goodhart's Law and Machine Learning: A Structural Perspective. *International Economic Review*, pp. iere.12633, March 2023. ISSN 0020-6598, 1468-2354. doi: 10.1111/iere.12633. URL https://onlinelibrary.wiley.com/doi/10.1111/iere.12633.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second, 2023.

Hu, E. J., Malkin, N., Jain, M., Everett, K., Graikos, A., and Bengio, Y. Gflownet-em for learning compositional latent variable models, 2023.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems, 2021.

Hutter, M. A gentle introduction to the universal algorithmic agent aixi, 2003.

Ivanov, R., Carpenter, T. J., Weimer, J., Alur, R., Pappas, G. J., and Lee, I. Case study: verifying the safety of an autonomous racing car with a neural network controller. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, HSCC '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370189. doi: 10.1145/3365365.3382216. URL https://doi.org/10.1145/3365365.3382216.

Jenner, E. and Gleave, A. Preprocessing reward functions for interpretability, 2022.

Jha, S. and Seshia, S. A. A Theory of Formal Synthesis via Inductive Learning. *Acta Informatica*, 54(7):693–726, 2017.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., and Gao, W. Ai alignment: A comprehensive survey, 2024.

Junges, S., Jansen, N., Dehnert, C., Topcu, U., and Katoen, J.-P. Safety-constrained reinforcement learning for MDPs. In Chechik, M. and Raskin, J.-F. (eds.), *Tools and Algorithms for the Construction and Analysis of Systems*, pp. 130–146, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. ISBN 978-3-662-49674-9.

Karwowski, J., Hayman, O., Bai, X., Kiendlhofer, K., Griffin, C., and Skalse, J. Goodhart's law in reinforcement learning, 2023.

Ke, N. R., Chiappa, S., Wang, J., Goyal, A., Bornschein, J., Rey, M., Weber, T., Botvinic, M., Mozer, M., and Rezende, D. J. Learning to induce causal structure, 2022.

Kearns, M. J. and Vazirani, U. *An Introduction to Computational Learning Theory*. The MIT Press, 08 1994. ISBN 9780262276863. doi: 10.7551/mitpress/3897.001.0001. URL https://doi.org/10.7551/mitpress/3897.001.0001.

Kim, J.-Y. and Cho, S.-B. An information theoretic approach to reducing algorithmic bias for machine learning. *Neurocomputing*, 500:26–38, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.09.081. URL https://www.sciencedirect.com/science/article/pii/S0925231222005987.

Kim, K., Garg, S., Shiragur, K., and Ermon, S. Reward identification in inverse reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5496–5505, Virtual, July 2021. PMLR.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores, 2016.

Koessler, L. and Schuett, J. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries, 2023.

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., and Legg, S. Specification gaming: the flip side of AI ingenuity, 2020. URL deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Lample, G., Lachaux, M.-A., Lavril, T., Martinet, X., Hayat, A., Ebner, G., Rodriguez, A., and Lacroix, T. Hypertree proof search for neural theorem proving, 2022.

Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., and Krueger, D. Goal misgeneralization in deep reinforcement learning, 2023.

Leino, K. R. M. *Program Proofs*. MIT Press, 2023.

Lelkes, Y., Sood, G., and Iyengar, S. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1): 5–20, 2017. ISSN 00925853, 15405907. URL http://www.jstor.org/stable/26379489.

Leveson, N. *Engineering a Safer World: Systems Thinking Applied to Safety*. 01 2012. ISBN 9780262298247. doi: 10.7551/mitpress/8179.001.0001.

Manheim, D. and Garrabrant, S. Categorizing Variants of Goodhart's Law, February 2019. URL http://arxiv.org/abs/1803.04585. arXiv:1803.04585 [cs, q-fin, stat].

Mathews, J. and Schmidler, S. C. Finite sample complexity of sequential monte carlo estimators on multimodal target distributions, 2022.

McCullough, D. *The Great Bridge: The Epic Story of the Building of the Brooklyn Bridge*. Simon & Schuster, 2001. ISBN 9780743217378. URL https://books.google.co.uk/books?id=bOM93rb22YEC.

McMahan, H. B., Likhachev, M., and Gordon, G. J. Bounded real-time dynamic programming: Rtdp with monotone upper bounds and performance guarantees. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pp. 569–576, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102423. URL https://doi.org/10.1145/1102351.1102423.

Memarian, B. and Doleck, T. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2023.100152. URL https://www.sciencedirect.com/science/article/pii/S2666920X23000310.

Metz, C. 'the godfather of A.I.' leaves Google and warns of danger ahead, May 2023. URL https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html. [Online; posted 4-May-2023].

Michaud, E. J., Gleave, A., and Russell, S. Understanding learned reward functions, 2020.

Michaud, E. J., Liao, I., Lad, V., Liu, Z., Mudide, A., Loughridge, C., Guo, Z. C., Kheirkhah, T. R., Vukelić, M., and Tegmark, M. Opening the ai black box: program synthesis via mechanistic interpretability. *arXiv preprint arXiv:2402.05110*, 2024.

Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., and Kolobov, A. Blog: Probabilistic models with unknown objects. *Statistical Relational Learning*, pp. 373, 2007.

Mingard, C., Skalse, J., Valle-Pérez, G., Martínez-Rubio, D., Mikulik, V., and Louis, A. A. Neural networks are a priori biased towards boolean functions with low entropy, 2020.

Mingard, C., Valle-Pérez, G., Skalse, J., and Louis, A. A. Is sgd a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021. URL http://jmlr.org/papers/v22/20-676.html.

Ng, A. Y. and Russell, S. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, volume 1, pp. 663–670, Stanford, California, USA, 2000. Morgan Kaufmann Publishers Inc.

Nieuwhof, G. An introduction to fault tree analysis with emphasis on failure rate evaluation. *Microelectronics Reliability*, 14(2):105–119, 1975. ISSN 0026-2714. doi: 10.1016/0026-2714(75)90024-4. URL https://www.sciencedirect.com/science/article/pii/0026271475900244.

OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

Orseau, L. and Armstrong, S. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*, 2016. URL https://api.semanticscholar.org/CorpusID:2912679.

Pan, A., Bhatia, K., and Steinhardt, J. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*, October 2021.

Pang, R. Y., Padmakumar, V., Sellam, T., Parikh, A. P., and He, H. Reward Gaming in Conditional Text Generation, February 2023. URL http://arxiv.org/abs/2211.08714. arXiv:2211.08714 [cs].

Pearl, J. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Pearl, J. Judea pearl: Causal reasoning, counterfactuals, and the path to agi, 2019. URL https://youtu.be/pEBI0vF45ic?si=lE_gCwZAFfDU7GG2&t=3869.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models, 2022.

Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019. ISBN 9780525558620. URL https://books.google.co.uk/books?id=M1eFDwAAQBAJ.

Russell, S. Written testimony of Stuart Russell before the U.S. Senate Committee on the Judiciary, Subcommitee on Privacy, Technology, and the Law, 2024. URL https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf.

Saad, F., Patton, B., Hoffman, M. D., A. Saurous, R., and Mansinghka, V. Sequential Monte Carlo learning for time series structure discovery. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29473–29489. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/saad23a.html.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing power and the governance of artificial intelligence, 2024.

Schlaginhaufen, A. and Kamgarpour, M. Identifiability and generalizability in constrained inverse reinforcement learning, 2023.

Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., and Garfinkel, B. Towards best practices in agi safety and governance: A survey of expert opinion, 2023.

Seligman, E., Schubert, T., and Kumar, M. A. K. *Formal verification: an essential toolkit for modern VLSI design*. Elsevier, 2023.

Seshia, S. A. Combining induction, deduction, and structure for verification and synthesis. *Proceedings of the IEEE*, 103(11):2036–2051, 2015.

Seshia, S. A., Desai, A., Dreossi, T., Fremont, D., Ghosh, S., Kim, E., Shivakumar, S., Vazquez-Chanlatte, M., and Yue, X. Formal specification for deep neural networks. In *Proceedings of the International Symposium on Automated Technology for Verification and Analysis (ATVA)*, pp. 20–34, October 2018.

Seshia, S. A., Sadigh, D., and Sastry, S. S. Toward verified artificial intelligence. *Communications of the ACM*, 65 (7):46–55, 2022.

Settle, J. E. *Frenemies: How Social Media Polarizes America*. Cambridge University Press, 2018.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022.

Skalse, J. and Abate, A. Misspecification in inverse reinforcement learning, 2023.

Skalse, J. and Abate, A. Quantifying the sensitivity of inverse reinforcement learning to misspecification, 2024.

Skalse, J., Farnik, L., Motwani, S. R., Jenner, E., Gleave, A., and Abate, A. Starc: A general framework for quantifying differences between reward functions, 2024.

Skalse, J. M. V., Howe, N. H. R., Krasheninnikov, D., and Krueger, D. Defining and Characterizing Reward Gaming. In *Advances in Neural Information Processing Systems*, May 2022.

Skalse, J. M. V., Farrugia-Roberts, M., Russell, S., Abate, A., and Gleave, A. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pp. 32033–32058. PMLR, 2023.

Soares, N., Fallenstein, B., Armstrong, S., and Yudkowsky, E. Corrigibility. In Walsh, T. (ed.), *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015*, volume WS-15-02 of *AAAI Technical Report*. AAAI Press, 2015. URL http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

Stray, J. Designing recommender systems to depolarize, 2021.

Subramani, R., Williams, M., Heitmann, M., Holm, H., Griffin, C., and Skalse, J. On the expressivity of objective-specification formalisms in reinforcement learning, 2024.

Szegedy, C. A promising path towards autoformalization and general artificial intelligence. In *Intelligent Computer Mathematics: 13th International Conference, CICM 2020, Bertinoro, Italy, July 26–31, 2020, Proceedings 13*, pp. 3–20. Springer, 2020.

Tegmark, M. *Life 3.0: Being human in the age of artificial intelligence*. Vintage, 2018.

Tegmark, M. and Omohundro, S. Provably safe systems: the only path to controllable agi, 2023.

Torfah, H., Xie, C., Junges, S., Vazquez-Chanlatte, M., and Seshia, S. A. Learning monitorable operational design domains for assured autonomy. In *Proceedings of the International Symposium on Automated Technology for Verification and Analysis (ATVA)*, October 2022.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, Jan 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06747-5. URL https://doi.org/10.1038/s41586-023-06747-5.

van Merwijk, C., Carey, R., and Everitt, T. A complete criterion for value of information in soluble influence diagrams, 2022.

Viano, L., Huang, Y.-T., Kamalaruban, P., Weller, A., and Cevher, V. Robust inverse reinforcement learning under transition dynamics mismatch. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=t8HduwpoQQv.

Wang, X., Zhang, Y., and Zhu, R. A brief review on algorithmic fairness. *Management System Engineering*, 1(1): 7, Nov 2022. ISSN 2731-5843. doi: 10.1007/s44176-022-00006-z. URL https://doi.org/10.1007/s44176-022-00006-z.

Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2009.

Watanabe, S. *Mathematical Theory of Bayesian Statistics*. Chapman and Hall/CRC, 2018.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., and Tenenbaum, J. B. From word models to world models: Translating from natural language to the probabilistic language of thought, 2023.

Yalcinkaya, B., Torfah, H., Fremont, D. J., and Seshia, S. A. Compositional simulation-based analysis of AI-based autonomous systems for markovian specifications. In *Runtime Verification - 23rd International Conference (RV)*, volume 14245 of *Lecture Notes in Computer Science*, pp. 191–212. Springer, 2023.

Zhang, Y., Tino, P., Leonardis, A., and Tang, K. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5): 726–742, October 2021. ISSN 2471-285X. doi: 10.1109/tetci.2021.3100641. URL http://dx.doi.org/10.1109/TETCI.2021.3100641.

Zhuang, S. and Hadfield-Menell, D. Consequences of misaligned AI. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pp. 15763–15773, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

Ziegler, D. M., Nix, S., Chan, L., Bauman, T., Schmidt-Nielsen, P., Lin, T., Scherlis, A., Nabeshima, N., Weinstein-Raun, B., de Haas, D., Shlegeris, B., and Thomas, N. Adversarial training for high-stakes reliability, 2022.