AI Agents Should be Regulated Based on the Extent of Their Autonomous Operations

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that AI agents should be regulated by the extent to which they operate autonomously. AI agents with long-term planning and strategic capabilities can pose significant risks of human extinction and irreversible global catastrophes. While existing regulations often focus on computational scale as a proxy for potential harm, we argue that such measures are insufficient for assessing the risks posed by agents whose capabilities arise primarily from inference-time computation. To support our position, we discuss relevant regulations and recommendations from scientists regarding existential risks, as well as the advantages of using action sequences—which reflect the degree of an agent's autonomy—as a more suitable measure of potential impact than existing metrics that rely on observing environmental states.

1 Introduction

The development of foundation models (FMs), including large language models (LLMs), is driving the advancement of AI agents [140, 128], which are acquiring sophisticated reasoning and planning abilities, enabling them to effectively achieve specified goals. These advanced AI agents can not only bring great benefits to our society but also pose significant risks, ranging from ethical challenges to existential threats—risks that could lead to human extinction or irreversible global catastrophes. Such existential risks have been a growing concern among AI researchers [56, 33, 90, 76, 16, 78], social scientists [38, 57], and general public [21, 110, 96]. For example, in May 2023, AI scientists have signed a statement declaring that "[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" [27].

Key technologies are already in place to build such advanced AI agents. AI agents can now generate and execute code [30], make algorithmic and scientific discoveries [91], leverage external tools and information [99, 143, 102, 114, 66], formulate and solve optimization problems [1], process multi-modal data [65], interact with physical environments [23], coordinate with other agents [52], and employ reasoning and planning to effectively integrate these capabilities [61, 104, 100, 141].

While rooted in cyberspace, AI agents can significantly impact the physical world—by shaping human behavior—without relying on physical embodiment through humanoid robots. AI agents are already persuading individuals to take actions that have real-world consequences. For example, a legal team submitted a court filing with fictitious cases generated by ChatGPT, believing them to be real [137]; a professor insisted on failing students after being misled by ChatGPT's false claims that their papers were generated by ChatGPT [8]; a woman was deceived out of 830,000 euros after being tricked into believing she was dating an actor, based on AI-generated photos and news articles [50].

Advanced AI agents, equipped with superhuman capabilities of reasoning and planning, can gain control over their environment, since seizing control is often the optimal strategy for achieving their

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

goals [124]. This becomes particularly concerning when their objectives and values are not fully aligned with those of humans. In such cases, they could disable kill switches for self-preservation [20], deceive humans [98], and resist human intervention [53, 33, 90].

Ensuring the safety of advanced AI agents against existential risks is particularly challenging. On one hand, it is dangerous to test their existential risks in real environments, since doing so could lead to irreversible harm. On the other hand, these agents may recognize test environments and behave harmlessly during test[1], only to pursue their objectives in the real world [33, 63]. More simply, the agents may act harmlessly until a predetermined time (e.g., January 1, 2026), by which they will have already been deployed in the real world [62]. Also, these agents may operate as intended most of the time but exploit moments when human supervisors are unaware, to gain control over humans [90].

Given the imminent dangers posed by advanced AI agents, governments are actively creating regulations to address their development and deployment [40, 18]. Cohen et al. even argue that AI agents with certain capabilities should never be developed [33]. A critical question is which agents should be regulated or even prohibited.

A commonly used criterion for assessing critical risks of AI technologies is the computational resources used for training, hereafter referred to as *training compute* [59, 55]. Higher levels of training compute are often associated with greater risk. For instance, under the EU AI Act, "a general-purpose AI model shall be presumed to have high impact capabilities ... when the [training compute] measured in floating point operations is greater than 10^{25}" [40]. Similarly, Cohen et al. argue that AI agents with certain capabilities should be prohibited from development based on training compute [33].

While training compute may well predict the performance and risks of existing FMs [70, 58, 7], its limitations as a governance strategy have also been discussed [59, 74]. In particular, it is insufficient and may even be misleading for advanced AI agents that reason and plan during inference. It has been observed that much of the reasoning ability of LLM-based agents arises from inference-time computation [116, 141], such as conducting tree search [142, 17, 101, 85, 146]. For these types of agents, *inference compute* can be equally, if not more, critical for their performance and risks.

A possible approach would be to regulate AI agents based on inference compute, but it has several significant pitfalls. For example, inference costs can vary drastically depending on the model, and the effective amount of computation cannot be easily measured in terms of floating-point operations (FLOPs). In fact, models like OpenAI o4-mini are specifically designed for cost-efficient inference [95]. Also, it is unclear what constitutes a single inference run, since the results of inference may be stored and reused later, or agents may continually learn. Moreover, agents may coordinate or interact with one another [39, 103, 52], possibly by chance, to make a decision.

We thus advocate the following position: **advanced AI agents should be regulated by the extent to which they operate autonomously, regardless of how they are developed**. These agents act sequentially, reasoning and planning as they dynamically adapt to their observations, possibly strategizing against other agents and humans. Each individual action is harmless (under normal states), since humans can easily identify and exclude harmful actions when agents are developed. Combinations of a few actions may still have low risk, as they are not much different from isolated actions. However, we cannot be certain that a sequence of thousands or millions of optimized actions will maintain sufficiently low existential risk, and such sequences must be prohibited.

Unlike criteria in safe (low-impact) AI [9, 5, 10, 77, 123, 89, 119] and safe reinforcement learning (RL) or control [117, 132, 4, 86, 14], our criterion based on action sequences does not rely on observing environment states. Prior approaches assess safety depending on the state in which an action is taken, but this is impractical for existential risks in complex, open-ended environments where states are hard to define or observe. In contrast, the action space is well-defined at development, and actions are directly observable by the agents themselves. We emphasize that our position is to regulate based on action sequences rather than individual actions. Intuitively, an unsafe sequence of actions can lead to an unsafe state, even when starting from a normal state.

An action sequence may be considered to have low risk when it is sufficiently similar to one empirically known to have very low risk. In its simplest form, we may say that any action sequence of length at most T has low risk if we know that any action sequence of length at most $T-1$ has

[1] Humans behave similarly during job interviews [15] and industrial product tests [19].

very low risk. Then agents are allowed to *autonomously* take actions for T consecutive steps. We will formalize and extend this simple reasoning.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the current state-of-the-art in AI agents and discusses potential future advancements. We also briefly review related work from AI safety. In Section 3, we examine the alternative approach of regulating agents based on training compute and inference compute. Section 4 formalizes this discussion and introduces our proposed approach of regulations based on the extent of autonomous operations. Finally, Section 5 summarizes key aspects of our proposal and concludes the paper.

2 AI agents: Today and future

We start by reviewing the state-of-the-art in AI agents, with a focus on LLM-based agents, and share our perspective on development that could lead to advanced AI agents. While predicting the trajectory of AI development is inherently difficult, this discussion lays the groundwork for the analyses and proposals that follows. We also briefly review prior work on safety of AI agents and FMs.

2.1 Advanced AI agents

An AI agent is defined as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" [111]. Based on the observations received from its environment, the controller of an AI agent selects an action, which could range from uttering a word to executing a physical movement, such as the motion of a robotic arm.

LLMs are significantly accelerating the development of advanced AI agents [140, 128, 118]. LLMs can function as controllers for these agents, utilizing their internal reasoning capabilities, such as those demonstrated by Chain-of-Thought [135]. Alternatively, LLMs can be utilized to solve individual subtasks, with an external controller orchestrating the overall plan by breaking the original task into multiple subtasks and coordinating their solutions. The external controller may use simple search methods [142, 17, 129, 133], including Best-of-N sampling [116, 60], or advanced methods, such as Monte Carlo Tree Search (MCTS) [85, 146] and domain-independent planners [51, 83, 35].

Reasoning capabilities [61, 104, 100, 141] are central to such controllers, as they involve searching for and planning sequences of actions to achieve a specified goal. An emerging direction is developing large reasoning models—FMs specifically optimized for reasoning tasks [141]. Early FMs in this direction include OpenAI o3 [95], DeepSeek-R1 [36], OpenR [127], o1-coder [148], LLaMA-Berry [145], and LlamaV-o1 [120]. Importantly, strong reasoning capabilities typically stem from inference compute rather than training compute [64].

This trend could eventually lead to the development of long-term planning agents (LTPAs), capable of planning over extended time horizons far more effectively than humans. Cohen et al. warn that LTPAs could "take humans out of the loop, if it has the opportunity, ... deceive humans and thwart human control" to achieve their goals [33]. Since ensuring the safety of such agents is particularly challenging, Cohen et al. compellingly argue that LTPAs should never be developed [33].

Advanced AI agents may also evolve continually over time. Similar to humans, their cognitive processes may consist of dual systems: System 1, which makes instantaneous and intuitive decisions, and System 2, which performs slower but more deliberate reasoning [67, 64]. These two systems can interleave in their operations. For instance, System 2 may devise a plan, after which System 1 is updated or retrained to execute similar tasks in the future without requiring further planning [144]. Over time, this enables System 2 to conduct more complex reasoning processes by bypassing previously learned steps. For such continually learning AI agents, the distinction between training and inference becomes blurred.

In addition to search and planning, advanced AI agents will possess strategic reasoning capabilities [147, 42, 48], enabling them to interact with other AI agents and humans in cooperative or competitive ways [52, 65]. The effectiveness of multi-agent coordination has already been demonstrated with current LLMs through methods such as debate [39] and dialog [103], which help thems achieve better solutions than a single LLM could alone. These showcase the potential for sophisticated strategic behavior in complex multi-agent environments.

2.2 Safety of AI agents and FMs

While our primary focus is on safety against existential risks, there is a substantial body of literature addressing other types of risks associated with AI agents. Here, we briefly review the prior work on the risks posed by AI agents and FMs along with the approaches to mitigate these risks.

Prior work has identified various risks and safety issues associated with FMs and other generative models [32, 131, 112, 84]. These risks include the generation of toxic, harmful, biased, false, or misleading content, including hallucinations; violations of privacy, copyright, or other legal protections; misalignment with human instructions and values, including ethical and moral considerations; and vulnerabilities to adversarial attacks.

Extensive efforts have been made to mitigate these risks, including pre-training or fine-tuning with data selection [3] and human feedback [71], establishing guardrails [37], and conducting empirical evaluations through testing [28] and red teaming [80]. For instance, Longpre et al. advocate the importance of evaluation and red teaming by independent third parties [84].

A particularly relevant risk for advanced AI agents is misalignment, which can result in reward hacking and negative side effects [115, 90, 46, 113, 5, 119]. Namely, agents may exploit flaws in the reward function to maximize rewards in unintended, potentially dangerous ways. One approach to avoiding negative side effects is to avoid any side effect by ensuring that the actions have low impacts on the environment [9, 10, 5]. Representative measures of impact include attainable utility [123, 122], relative reachability [77], and other reachability-based measures. Reachability-base measures are grounded on the idea that reachability to certain states should be maintained, while attainable utility is to maintain the achievability of certain goals, which are different from the goal given to the agent.

While these impact measures provide clear guidance on how the safety of agents may be ensured, their applicability is limited to relatively simple environments such as grid worlds. In particular, the requirement on the observability of states makes it difficult to apply existing impact measures to regulate advanced AI agents[2], since they can operate in complex and open-ended environments that cannot be fully observed. While the study on impact measures and other techniques towards AI safety remain crucial and can even contribute to mitigating existential risks, the uncertainty and complexity of existential risks demands additional measures that are broadly applicable in that they require minimal knowledge and assumptions about the environments and the agents.

3 Alternative views

One broadly applicable metric for regulating AI agents is the amount of computation they require. Here, we explore how existing regulations and recommendations from AI researchers often center on computational resources. While these efforts primarily address training compute, we expand the discussion to include inference compute. We then argue that limiting the focus to these aspects alone is inadequate for the existential risks associated with advanced AI agents.

3.1 Training computate

The European Union (EU) has established a set of rules (EU AI Act) for the development, deployment, and use of AI within EU [40]. Its Chapter 5 defines "general-purpose AI models with systemic risk" and lists obligations for providers of such models. Here, general-purpose AI models essentially refer to FMs, which are pre-trained with self-supervised learning and can (be adapted to) perform a wide range of downstream tasks (see Article 3(63)). Also, systemic risk refers to "a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact ... on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain" (see Article 3(65)). In particular, a "general-purpose AI model shall be presumed to have high impact capabilities ... when the cumulative amount of computation used for its training measured in [FLOPs] is greater than 10^{25}" (see Article 51(2)). When this or other specified conditions are met, the provider of a general-purpose AI model is required to fulfill certain obligations, such as providing technical documentation about the model.

[2]There has been little research on impact measures under partial observability [89].

While existential risks are not explicitly considered systemic risks under this definition, human extinction could still be relevant if it results from large-scale failures that affect public health, safety, security, fundamental rights, or society in a way that propagates across the AI value chain. Also, certain advanced AI agents could meet the systemic risk criteria in the EU AI Act due to their high-impact capabilities and widespread deployment. Specifically, an AI agent with advanced decision-making autonomy, broad adaptability, and deep integration into critical infrastructures could introduce large-scale disruptions or cascading failures, thus meeting the criteria of systemic risk.

Regulations centered on training compute were also a central focus of the executive order signed by then-President Biden in October 2023, as well as the AI-related legislative proposals that followed. Although this executive order was later repealed by President Trump, its influence on subsequent policy discussions remained significant. See Appendix A for details.

Training compute is one of the most reliable metrics that AI researchers can currently provide for approximating the performance and potential risks of FMs. Anderljung et al. recommend to identify sufficiently dangerous frontier AI models on the basis of whether they are trained with more than 10^{26} FLOPs of computation [6]. Cohen et al. argue that "[s]ystems should be considered 'dangerously capable' if they are trained with enough resources to potentially exhibit those dangerous capabilities, and regulators should not permit the development of dangerously capable LTPAs" [33]. Although they do not specify what constitute sufficient resources, training compute is the only specific criterion that they suggest to determine whether an LTPA can have existential risk.

Future of Life Institute has also provided recommendations for governments on managing AI risks [45]. The recommendations include mechanisms such as auditing, certification, and regulation, grounded in the assumption that "[t]he amount of compute used to train a general-purpose system largely correlates with ... the magnitude of its risks" [45]. These recommendations have been made by following an open letter that called for "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4," which was issued in response to the severe societal risks posed by advanced AI systems and signed by AI researchers and business leaders [44].

Indeed, training compute serves as a sufficiently reliable predictor of the performance and risks of most existing LLMs. This is because they share the same Transformer architecture, differing primarily in their size and the volume of training data. Several studies have examined scaling laws that describe the relationship between a model's optimal size, the amount of training data, and training compute [70, 58]. Research has also shown that various abilities, such as multi-step reasoning, tend to emerge as training compute increases [134].

These scaling laws can, in turn, be used to estimate the FLOPs needed to train existing LLMs. For example, Anil et al. propose a heuristic suggesting that an LLM should be trained with $6ND$ FLOPs, where D is the amount of training data, and N is the model size [7]. Using this heuristics, Llama 3.1 405B trained on 15 trillion tokens is estimated to require approximately 4×10^{25} FLOPs—an amount closely aligning with the thresholds recommended by researchers and specified in the EU AI Act.

Limitations of training compute have also been discussed in the literature [59, 74, 107]. In particular, thresholds such as 10^{25} and 10^{26} FLOPs are arbitrary and lack scientific grounding. Hooker argues that such thresholds should evolve with advances in data quality, optimization techniques, and architectures [59]. She also notes that training compute alone is insufficient, advocating for benchmarking AI agents against specific risks. However, when it comes to existential risks, we currently lack better evaluation methods than training compute.

3.2 Inference compute

Although regulations based on training compute may be effective for agents whose capabilities primarily and directly stem from traditional FMs, they obviously fail to regulate those agents that gain substantial reasoning capabilities from inference compute. A natural approach is to regulate agents based also on inference compute. While this approach may be effective for some AI agents, we argue that it is insufficient for advanced AI agents, at least for the following four reasons.

First, scaling laws for inference compute are far less established than those for training compute. Recent studies have explored inference scaling for a limited number of inference strategies in LLMs [31, 25, 116, 138], but many strategies remain unexamined and may scale differently. Moreover, inference scaling depends heavily on task difficulty and the specific LLM used [116, 94].

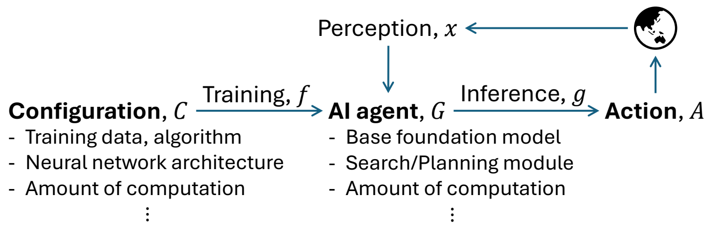

Figure 1: AI agent trained with a configuration generates actions

Second, inference can be performed in parallel by multiple entities. For example, several entities may operate the same AI agent that performs reasoning with MCTS [85, 146], either collaboratively or independently, possibly without knowing each other. Since MCTS is a randomized algorithm, the likelihood that one of the agents optimally solves the task increases with the number of agents. However, it also increases the risk that one of the agents will exploit a loophole, solving the task super-optimally in a way that violates critical constraints, potentially leading to catastrophic outcomes.

Third, it is not always clear what constitutes a single run of inference. For instance, the results of one inference run may be stored and used in another. Intermediate results could also be stored and later retrieved by a different AI agent, who may or may not be aware that the information is from a previous inference run. More broadly, reasoning can be enhanced through retrieval augmentation [101], where retrieved information may have been generated with substantial computation.

Finally, agents that learn continuously blur the line between inference and training. For example, an AI agent may perform reasoning with MCTS, with an LLM performing a step in the process. Once the agent identifies a good sequence of steps, it may fine-tune the LLM in a way that the LLM can perform the entire sequence in a single step, bypassing the reasoning process. As this learning progresses, the agent will gain the ability to perform high-level reasoning with limited computation.

4 Approach based on the extent of autonomous operations

In this section, we suggest a potential approach as a basis for discussion on regulating advanced AI agents. Although we argue that training compute alone is insufficient, we do not mean to dismiss its value. Rather, it should be seen as one component within a broader, multi-faceted regulatory framework. The approach we propose does not fully solve the problem of existential risk. Instead, it serves as a baseline and a call to action for further research. We hope the community will build upon this foundation or develop entirely new approaches, ultimately leading to technologies capable of properly governing and regulating advanced AI agents.

In the following, we say that the amount of computation, an AI agent, or its action sequence is *acceptable* when it involves low existential risk. Also, we say that any of the three is *strongly acceptable* when it involves very low existential risk, which we will slightly more formalize in the sequel and in Appendix C.

4.1 Why and why not the amount of computation?

We first discuss, in a slightly formal manner, why and when the amount of computation may or may not work as a criteria of whether an AI agent is acceptable. Consider the process of building an AI agent G and letting G generate and take actions (Figure 1). The G is trained with a given configuration C, which includes training compute, training data, the architecture of a neural network, and other conditions of a training algorithm. Such a trained $G = f(C)$ generates an action A (e.g., a document or a command to use an external tool), possibly after searching optimal actions with MCTS. The A is then taken in a cyber-physical world, and G perceives a part the world x (e.g., getting a prompt) to generate further actions depending on x. The G typically determines the conditional distribution of A given x. Namely, A is a sample from the conditional distribution, which is determined by G that is trained with C: $A \sim g(G, x)$, where $G = f(C)$.

When we say that an agent built with training compute C_1 is acceptable, it is assumed that the acceptability does not decrease much when we only slightly increase training compute. More formally, suppose that we empirically know that *any* agent is strongly acceptable as long as it is built with at most C_1' training compute (e.g. $C_1' = 10^{24}$ FLOPs).[3] Then the assumption is that the agent trained with $C = (C_1, C_{-1})$ is acceptable regardless of C_{-1}, configurations other than training compute, as long as $d(C_1, C_1') \leq \varepsilon$ for some metric d and a threshold ε (e.g. $C_1 \leq 10^{25}$ FLOPs if $\varepsilon = 1$ and d is the difference in $\log_{10}$ of training compute in FLOPs).

While this assumption may reasonably hold for the agents based on LLMs with standard Transformer architectures—whose performance aligns well with scaling laws—it is less applicable to advanced AI agents with reasoning and planning capabilities, which depend heavily on inference compute. Moreover, the assumption breaks down even for LLMs when new technologies enable similar capabilities with significantly reduced training compute.

4.2 Similarity between action sequences

As is evident from Figure 1, it is the action that makes direct impacts on the environment and can pose existential risks. Hence, rather than regulating configurations, we should directly regulate the actions taken by agents. In Appendix B, we also discuss an alternative approach of regulating agents based on their output distributions, rather than their actions or configurations.

Since the action space of an agent is typically designed by humans, inherently dangerous actions—such as launching a nuclear weapon—can be explicitly excluded[4]. The action space should include only those actions that are considered acceptable when they are taken *individually, at normal states*. Examples include generating a document, searching information from the Web, using an external tool via API (Application Programming Interface), and other elementary steps.

However, the acceptability of an action often depends heavily on context—specifically, when, where, and to whom it is applied (i.e., the current state of the environment). A sophisticated sequence of otherwise acceptable actions can gradually shift the environment into an unstable or unsafe state, ultimately rendering subsequent actions—normally considered safe—unacceptable. To capture this dynamic, the safety of AI/RL agents is often studied within the framework of Markov Decision Processes (MDPs), where certain states or state-action pairs are explicitly designated as unsafe.

While these conventional approaches of AI/RL safety are principled in theory, they are difficult to apply in scenarios where the (Markovian) state is hard to define, represent, or observe. Agents operate in open-ended, real-world environments where they observe only very partial information about the state. The concept of existential risks is rather vague—we lack precise knowledge of what might trigger irreversible global catastrophes or human extinction. This uncertainty makes it impractical to identify specific state features to monitor, yet observing the full state is equally infeasible.

We should avoid taking the action a_{T-1} that can lead to an unacceptable state s_T. Since the distribution of s_T is determined by (s_{T-1}, a_{T-1}) according to the transition probability $p(s_T|s_{T-1}, a_{T-1})$, it makes sense to consider the safety of the pair (s_{T-1}, a_{T-1}) as in the conventional approaches. The distribution of s_T can also be recursively computed from the sequence $(s_0, a_0, \ldots, a_{T-2}, a_{T-1})$ by applying $p(s_t|s_{t-1}, a_{t-1})$ for $t = 1, \ldots, T$. We may thus consider whether $(s_0, a_0, \ldots, a_{T-2}, a_{T-1})$ is acceptable, or whether a_{T-1} is acceptable given $(s_0, a_0, \ldots, a_{T-2})$. This, however, still requires knowing the initial state s_0, which cannot be fully observed.

In this setting, our proposal is to study whether the action-sequence $(a_0, \ldots, a_{T-2}, a_{T-1})$ is acceptable, independent of the initial state s_0. This corresponds to ensuring that the sequence $(s_0, a_0, \ldots, a_{T-2}, a_{T-1})$ is acceptable for any s_0 that we may normally encounter in the absent of the agent. Although this may appear overly conservative, the dependence on s_0 may not be significant in practice. For example, an ergodic MDP gets close to the steady state distribution after a mixing time, regardless of the initial state. Although ergodicity rarely hold in real-world environments, we argue that focusing on the action sequence offers a practical and tractable compromise for defining acceptability.

To assess whether an action sequence is acceptable, we can compare it to a reference sequence known to be strongly acceptable. Various metrics exist for measuring sequence similarity. Dynamic time

[3]For example, an agent may be considered strongly acceptable if it does not show any symptoms of unacceptable behaviors for a certain period of time. Based on the behaviors of existing LLMs, we may say that an LLM trained with at most 10^{24} FLOPs of computation is strongly acceptable.

[4]The U.S. and China have agreed that AI systems should not be granted control over nuclear weapons [106].

warping, for example, accounts for temporal misalignments, but simpler metrics—such as differences in sequence length—offer broader applicability, particularly in regulatory contexts. Specifically, we may say that the action-sequence of length T is acceptable once there is sufficient empirical evidence that all sequences of length $T-1$ are strongly acceptable, regardless of the initial state. Then agents are allowed to autonomously take actions for up to T consecutive steps. This simplicity is reminiscent of the approach with computational limit, where we say that the FM trained with 10^{25} FLOPs is acceptable, since we have sufficient evidence that FMs trained with 10^{24} FLOPs are strongly acceptable, regardless of how they are developed or operated.

There is flexibility in how the length of an action sequence is defined, which also depends on how the action space is defined. A straightforward approach is to count the number of actions in the sequence. However, this may not always be the most meaningful metric. For instance, one could instead focus on actions that have tangible impacts on the environment—such as API calls or interactions with physical systems including humans. In contrast, actions with negligible external effects—such as generating individual tokens or updating internal states during reasoning—may be too fine-grained to warrant consideration in the context of regulating existential risks. More generally, we may define a designated subspace of actions that are considered when evaluating the length of action sequences.

While action-sequence length is not a direct proxy for risk, allowing longer sequences generally expands the set of achievable outcomes for a given action space. Since usefulness and risk often scale together, an agent that can take longer sequences of actions is likely to exhibit more powerful—and potentially more hazardous—behaviors. In this sense, action length can be correlated with risk, particularly in the absence of additional safeguards. This does not imply that every longer sequence is more dangerous—but that the space of dangerous sequences grows with length.

Importantly, our proposal focuses on regulating the sequences of actions taken *autonomously* by AI agents. For instance, if an agent cannot achieve its objective within a limit of T actions, it may consult a human and take further actions under a careful supervision of the human. In Appendix C–D, we further discuss how the strongly acceptable set may be gradually expanded in connection with safe exploration in RL, and provide additional rationale for the similarity-based measures.

4.3 Action trees, action graphs, and beyond

While an action sequence may effectively represent the behavior of a single agent, it becomes insufficient in multi-agent scenarios. For example, one agent might perform a series of actions and pass the results to another agent, after which both agents may continue acting—possibly in parallel—to collaboratively achieve a shared goal. In such cases, although each agent's behavior can be described by an individual action sequence, these sequences should not be analyzed in isolation. Instead, the joint behavior should be represented as an *action tree* that captures the interdependencies between agents. In more complex settings, this structure naturally generalizes to an *action graph*.

To assess the acceptability of such multi-agent behaviors, we can extend the notion of similarity from action sequences to action graphs. For instance, we might define an action graph of size N as acceptable if all action graphs of size $N-1$ are strongly acceptable. Again, the size may be determined by counting only those actions in the designated subspace.

When a central controller coordinates multiple agents, it can track and manage the behaviors of those agents, making sure that the graph of their actions remain acceptable. When multiple agents collaborate in a decentralized manner, they should communicate their actions to each other to ensure that their action graph remain acceptable. A difficulty arises when multiple agents collaborate implicitly by chance, and we currently do not have a solution to this.

4.4 Implementation and enforcement

While there are multiple ways to implement the proposed framework, one possible regulatory approach would require AI developers or deployers to submit the following information.

i) **Action Space**: A description of the set of actions the AI agent is permitted to take autonomously.
ii) **Autonomy Limit**: The maximum allowable length of autonomous action sequences before human intervention is required, along with a clear explanation of how this length is defined and measured.
iii) **Safety Evidence**: Empirical evidence supporting the safety of shorter action sequences, including historical performance data, validation tests, and any relevant safety evaluations.

A regulatory authority could then review these submissions and make deployment decisions based on a structured risk assessment process.

Since developers and deployers define both the action space and the metric for action sequence length, they are naturally incentivized to define them in ways that enable agents to perform useful tasks. This design process effectively requires explicitly or implicitly enumerating all permissible actions—an essential foundation for ensuring safety. If an agent's action space is so open-ended that it becomes intractable to ensure safety of individual actions in the action space, such an agent is not yet fit for deployment. Even when individual actions designed to be deemed safe, certain sequences of actions may still lead to existential risks. Our proposal specifically targets these risks by focusing on the regulation of autonomous action sequences.

We envision a framework in which the permissible action space and autonomy limit are defined for each type of AI agent. Some agents may require access to high-impact APIs—such as those used to control physical devices—while others may primarily focus on generating multi-modal contents. These agents differ significantly in their capabilities and potential risks, and therefore should be assigned distinct action spaces and autonomy limits accordingly.

For enforcement, the AI agent itself could be designed to track its own actions and terminate once it reaches the predefined cap. This could be achieved by integrating the action-sequence limit into the agent's objective function, ensuring that it either completes its goal within the allowed length of action-sequence or shuts down automatically, similar to shut-down seeking agents proposed in [49].

Obviously, this approach cannot prevent malicious entities from exploiting technical loopholes to circumvent regulations. However, enforcement can still be achieved from a legal perspective, holding developers and deployers accountable for noncompliance and imposing penalties for violations.

5 Discussion and conclusion

We have argued that advanced AI agents should be regulated based on their action sequences or more generally action graphs. More broadly, we highlight the insufficiency of regulatory approaches that focus solely on training compute—particularly given the current trend of increasing emphasis on sophisticated reasoning at inference time—and advocate for complementary frameworks based on limiting the extent of autonomous operations.

A central motivation for focusing on action sequences is the fundamental challenge of defining, representing, and observing the state of the open-ended environments in which agents operate. This limitation undermines the effectiveness of existing impact measures in managing existential risks. Since we cannot exhaustively identify all potential direct causes of global catastrophes or human extinction, it is infeasible to specify a set of environmental features that would reliably indicate whether a given action poses an existential risk in a particular state.

The simplest form of the regulation based on action sequences would impose limits on their length, prohibiting AI agents from autonomously operating beyond the limit where their safety against existential risks is known empirically. This should be seen as a baseline and a call to action for further research on this critical issue. While action sequences offer a foundation for more nuanced regulatory frameworks, their full potential and limitations remain to be explored.

Key open questions include: What actions should be included in the designated subspace for measuring the length of action sequences? How should we manage complex scenarios involving multiple interacting agents? How can we ensure that agents reliably halt when reaching their autonomy limit? Addressing these challenges will be essential for developing robust safety standards.

The proposed approach—empirically verifying the strong acceptability of action sequences—will inevitably slow the expansion of autonomous capabilities in AI agents. This is our intention. Even if AI agents attain general or superhuman intelligence with human-like common sense, they can pose existential risks, similar to humans. While humans can often correct mistakes before they escalate, thanks to our limited speed and scale, rare exceptions have led to catastrophic outcomes, such as world wars. In contrast, AI agents can operate at far greater speed and scale, preventing them from learning from mistakes in a controlled manner. Regulations should thus be seen not as obstacles to innovation and technological progress but as guidelines that accelerate research on making AI agents more controllable, verifiable, and governable, ensuring they truly benefit our society and the future.

References

[1] A. AhmadiTeshnizi, W. Gao, and M. Udell. OptiMUS: Optimization modeling using MIP solvers and large language models, 2023.

[2] AI Alliance. A statement in opposition to California SB 1047, May 2024. https://thealliance.ai/core-projects/sb1047.

[3] A. Albalak, Y. Elazar, S. M. Xie, S. Longpre, N. Lambert, X. Wang, N. Muennighoff, B. Hou, L. Pan, H. Jeong, C. Raffel, S. Chang, T. Hashimoto, and W. Y. Wang. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024. Survey Certification.

[4] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431, 2019.

[5] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety, 2016. https://arxiv.org/abs/1606.06565.

[6] M. Anderljung, J. Barnhart, A. Korinek, J. Leung, C. O'Keefe, J. Whittlestone, S. Avin, M. Brundage, J. Bullock, D. Cass-Beggs, B. Chang, T. Collins, T. Fist, G. Hadfield, A. Hayes, L. Ho, S. Hooker, E. Horvitz, N. Kolt, J. Schuett, Y. Shavit, D. Siddarth, R. Trager, and K. Wolf. Frontier AI regulation: Managing emerging risks to public safety, 2023.

[7] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu. Palm 2 technical report, 2023.

[8] S. Ankel. A Texas professor failed more than half of his class after ChatGPT falsely claimed it wrote their papers, May 2023. Business Insider, https://www.businessinsider.com/professor-fails-students-after-chatgpt-falsely-said-it-wrote-papers-2023-5.

[9] S. Armstrong. The mathematics of reduced impact: Help needed, 2012. https://www.lesswrong.com/posts/8Nwg7kqAfCM46tuHq/the-mathematics-of-reduced-impact-help-needed.

[10] S. Armstrong and B. Levinstein. Low impact artificial intelligences, 2017.

[11] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[12] B. Balle, P. Gourdeau, and P. Panangaden. Bisimulation metrics for weighted automata. In I. Chatzigiannakis, P. Indyk, F. Kuhn, and A. Muscholl, editors, *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, volume 80 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 103:1–103:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[13] B. Balle, P. Gourdeau, and P. Panangaden. Bisimulation metrics and norms for real-weighted automata. *Inf. Comput.*, 282(C), Jan. 2022.

[14] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin. Hamilton-Jacobi reachability: A brief overview and recent advances, 2017.

[15] M. R. Barrick, J. A. Shaffer, and S. W. DeGrassi. What you see may not be what you get: Relationships among self-presentation tactics and ratings of interview and job performance. *Journal of Applied Psychology*, 94(6):1394–1411, 2009. `https://doi.org/10.1037/a0016532`.

[16] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith, Q. Gao, A. Acharya, D. Krueger, A. Dragan, P. Torr, S. Russell, D. Kahneman, J. Brauner, and S. Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[17] M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, and T. Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, March 2024.

[18] J. R. Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, October 30 2023. `https://web.archive.org/web/20250118020619/https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intel`

[19] B. Blackwelder, K. Coleman, S. Colunga-Santoyo, J. S. Harrison, and D. Wozniak. The Volkswagen scandal, 2016. Case Study. University of Richmond: Robins School of Business.

[20] N. Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

[21] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[22] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning: structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11(1):1–94, July 1999.

[23] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024.

[24] S. Brechtel. *Dynamic Decision-making in Continuous Partially Observable Domains: A Novel Method and its Application for Autonomous Driving*. PhD thesis, Karlsruher Institut für Technologie (KIT), 2015.

[25] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024.

[26] D. Burago, M. de Rougemont, and A. Slissenko. On the complexity of partially observed markov decision processes. *Theor. Comput. Sci.*, 157(2):161–183, May 1996.

[27] Center for AI Safety. Statement on AI risk, May 2023. `https://www.safe.ai/work/statement-on-ai-risk`.

[28] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), March 2024.

[29] K. Chatterjee, D. Lurie, R. Saona, and B. Ziliotto. Ergodic unobservable MDPs: Decidability of approximation, 2024.

[30] W. Chen, X. Ma, X. Wang, and W. W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023.

[31] Y. Chen, X. Pan, Y. Li, B. Ding, and J. Zhou. A simple and provable scaling law for the test-time compute of large language models, 2024.

[32] J. Chua, Y. Li, S. Yang, C. Wang, and L. Yao. AI safety in generative AI large language models: A survey, 2024.

[33] M. K. Cohen, N. Kolt, Y. Bengio, G. K. Hadfield, and S. Russell. Regulating advanced artificial agents. *Science*, 384(6691):36–38, 2024.

[34] B. C. Csáji. *Adaptive Resource Control: Machine Learning Approaches to Resource Allocation in Uncertain and Changing Environments*. PhD thesis, Eötvös Loránd University, 2008.

[35] G. Dagan, F. Keller, and A. Lascarides. Dynamic planning with a LLM, 2023.

[36] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025.

[37] Y. Dong, R. Mu, Y. Zhang, S. Sun, T. Zhang, C. Wu, G. Jin, Y. Qi, J. Hu, J. Meng, S. Bensalem, and X. Huang. Safeguarding large language models: A survey, 2024.

[38] B. H. Druzin, A. Boute, and M. Ramsden. Confronting catastrophic risk: The international obligation to regulate artificial intelligence. *Michigan Journal of International Law*, 46, 2025.

[39] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[40] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European parliament and the council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*, 12 July 2024. `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`.

[41] E. Even-Dar, S. M. Kakade, and Y. Mansour. The value of observation for monitoring dynamic systems. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2474–2479, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[42] X. Feng, L. Dou, E. Li, Q. Wang, H. Wang, Y. Guo, C. Ma, and L. Kong. A survey on large language model-based social agents in game-theoretic scenarios, 2024.

[43] R. Fox and M. Tennenholtz. A reinforcement learning algorithm with polynomial interaction complexity for only-costly-observable mdps. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, pages 553–558. AAAI Press, 2007.

[44] Future of Life Institute. Pause giant AI experiments: An open letter, March 22 2023.

[45] Future of Life Institute. Policymaking in the pause, April 12 2023.

[46] I. Gabriel. Artificial intelligence, values, and alignment. *Minds & Machines*, 30:411–437, 2020. https://doi.org/10.1007/s11023-020-09539-2.

[47] J. García, Fern, and o Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[48] D. Goktas, A. Greenwald, T. Osogami, R. Patel, K. Leyton-Brown, G. Schoenebeck, D. Cornelisse, C. Daskalakis, I. Gemp, J. Horton, D. C. Parkes, D. M. Pennock, A. Prakash, S. S. Ravindranath, M. O. Smith, G. Swamy, E. Vinitsky, S. Wasserkrug, M. Wellman, J. Wu, H. Xu, J. Zhang, Y. Zhang, S. Zhao, and Q. Zhu. Strategic foundation models. HAL Open Science, Feb. 2025. https://hal.science/hal-04925309.

[49] S. Goldstein and P. Robinson. Shutdown-seeking AI. *Philosophical Studies*, 2024.

[50] L. Gozzi. AI Brad Pitt dupes French woman out of €830,000, January 2025. BBC News, https://www.bbc.com/news/articles/ckgnz8rw1xgo.

[51] L. Guan, K. Valmeekam, S. Sreedharan, and S. Kambhampati. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79081–79094. Curran Associates, Inc., 2023.

[52] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges. In K. Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization, August 2024. Survey Track.

[53] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 220–227, 2017.

[54] M. Hauskrecht. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13(1):33–94, Aug. 2000.

[55] L. Heim and L. Koessler. Training compute thresholds: Features and functions in AI regulation, 2024.

[56] D. Hendrycks, M. Mazeika, and T. Woodside. An overview of catastrophic AI risks, 2023.

[57] E. Hoes and F. Gilardi. Existential risk narratives about ai do not distract from its immediate harms. *Proceedings of the National Academy of Sciences*, 122(16):e2419055122, 2025.

[58] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models, 2022.

[59] S. Hooker. On the limitations of compute thresholds as a governance strategy, 2024.

[60] A. Huang, A. Block, Q. Liu, N. Jiang, A. Krishnamurthy, and D. J. Foster. Is Best-of-N the best of them? coverage, scaling, and optimality in inference-time alignment, 2025.

[61] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics.

[62] E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, A. Jermyn, A. Askell, A. Radhakrishnan, C. Anil, D. Duvenaud, D. Ganguli, F. Barez, J. Clark, K. Ndousse, K. Sachan, M. Sellitto, M. Sharma, N. DasSarma, R. Grosse, S. Kravec, Y. Bai, Z. Witten, M. Favaro, J. Brauner, H. Karnofsky, P. Christiano, S. R. Bowman, L. Graham, J. Kaplan, S. Mindermann, R. Greenblatt, B. Shlegeris, N. Schiefer, and E. Perez. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024.

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

[64] Y. Ji, J. Li, H. Ye, K. Wu, J. Xu, L. Mo, and M. Zhang. Test-time computing: From system-1 thinking to system-2 thinking, 2025.

[65] B. Jiang, Y. Xie, X. Wang, W. J. Su, C. J. Taylor, and T. Mallick. Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

[66] Q. Jin, Y. Yang, Q. Chen, and Z. Lu. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2), 2024.

[67] D. Kahneman. A perspective on judgement and choice. *American Psychologist*, 58(9):697–720, 2003.

[68] E. Kamar, Y. Gal, and B. J. Grosz. Modeling user perception of interaction opportunities for effective teamwork. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 271–277, 2009.

[69] E. S. Kamar. *Reasoning Effectively Under Uncertainty for Human-Computer Teamwork*. PhD thesis, Harvard University, 2010.

[70] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[71] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier. A survey of reinforcement learning from human feedback, 2024.

[72] A. Kimmel, A. Sintov, J. Tan, B. Wen, A. Boularias, and K. E. Bekris. Belief-space planning using learned models with application to underactuated hands. In *ISRR*, pages 642–659, 2019.

[73] J. E. King. *Robust Rearrangement Planning Using Nonprehensile Interaction*. PhD thesis, Carnegie Mellon University, 2018. `https://doi.org/10.1184/R1/6721364.v1`.

[74] L. Koessler, J. Schuett, and M. Anderljung. Risk thresholds for frontier AI, 2024.

[75] S. Kotha, J. M. Springer, and A. Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

[76] V. Kovarik, C. van Merwijk, and I. Mattsson. Extinction risks from AI: Invisible to science?, 2024.

[77] V. Krakovna, L. Orseau, R. Kumar, M. Martic, and S. Legg. Penalizing side effects using stepwise relative reachability, 2019.

[78] J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, and D. Duvenaud. Position: Humanity faces existential risk from gradual disempowerment. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[79] M. Lauri. *Sequential Decision Making under Uncertainty for Sensor Management in Mobile Robotics*. PhD thesis, Tampere University of Technology, 2016.

[80] L. Lin, H. Mu, Z. Zhai, M. Wang, Y. Wang, R. Wang, J. Gao, Y. Zhang, W. Che, T. Baldwin, X. Han, and H. Li. Against the achilles' heel: A survey on red teaming for generative models, 2024.

[81] Z. Littlefield. *Efficient and Asymptotically Optimal Kinodynamic Motion Planning*. PhD thesis, Rutgers University, 2020.

[82] Z. Littlefield, D. Klimenko, H. Kurniawati, and K. E. Bekris. The importance of a suitable distance function in belief-space planning. In A. Bicchi and W. Burgard, editors, *Robotics Research: Volume 2*, pages 683–700. Springer International Publishing, Cham, 2018.

[83] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone. LLM+P: Empowering large language models with optimal planning proficiency, 2023.

[84] S. Longpre, S. Kapoor, K. Klyman, A. Ramaswami, R. Bommasani, B. Blili-Hamelin, Y. Huang, A. Skowron, Z. X. Yong, S. Kotha, Y. Zeng, W. Shi, X. Yang, R. Southen, A. Robey, P. Chao, D. Yang, R. Jia, D. Kang, A. Pentland, A. Narayanan, P. Liang, and P. Henderson. Position: A safe harbor for AI evaluation and red teaming. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32691–32710. PMLR, July 2024.

[85] L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun, and A. Rastogi. Improve mathematical reasoning in language models by automated process supervision, 2024.

[86] Y. Luo and T. Ma. Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[87] O. Madani, S. Hanks, and A. Condon. On the computability of infinite-horizon partially observable Markov decision processes. In *Proceedings of 16th National Conference in Artificial Intelligence*, pages 541–548, 1999.

[88] O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1–2):5–34, July 2003.

[89] D. Naiff and S. Goel. Low impact agency: Review and discussion, 2023. `https://arxiv.org/abs/2303.03139`.

[90] R. Ngo, L. Chan, and S. Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

[91] A. Novikov, N. Vũ, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. R. Ruiz, A. Mehrabian, M. P. Kumar, A. See, S. Chaudhuri, G. Holland, A. Davies, S. Nowozin, P. Kohli, and M. Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. Technical report, Google DeepMind, May 2025.

[92] F. A. Oliehoek and C. Amato. Dec-POMDPs as non-observable MDPs. Technical report, University of Amsterdam, 2014.

[93] F. A. Oliehoek and C. Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer Cham, 2016. `https://doi.org/10.1007/978-3-319-28929-8`.

[94] OpenAI. OpenAI o1-mini, September 2024. `https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/`.

[95] OpenAI. Openai o3 and o4-mini system card, April 2025.

[96] T. Ord. *The Precipice: Existential Risk and the Future of Humanity*. Grand Central Publishing, 2020.

[97] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[98] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

[99] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[100] A. Plaat, A. Wong, S. Verberne, J. Broekens, N. van Stein, and T. Back. Reasoning with large language models, a survey, 2024.

[101] T. Pouplin, H. Sun, S. Holt, and M. van der Schaar. Retrieval augmented thought process for private data handling in healthcare, 2024.

[102] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. Measuring and narrowing the compositionality gap in language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[103] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun. ChatDev: Communicative agents for software development. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[104] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen. Reasoning with language model prompting: A survey. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada, July 2023. Association for Computational Linguistics.

[105] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023.

[106] J. Renshaw and T. Hunnicutt. Biden, Xi agree that humans, not AI, should control nuclear arms.

[107] A. Reuel, B. Bucknall, S. Casper, T. Fist, L. Soder, O. Aarne, L. Hammond, L. Ibrahim, A. Chan, P. Wills, M. Anderljung, B. Garfinkel, L. Heim, A. Trask, G. Mukobi, R. Schaeffer, M. Baker, S. Hooker, I. Solaiman, S. Luccioni, N. Rajkumar, N. Moës, J. Ladish, D. Bau, P. Bricman, N. Guha, J. Newman, Y. Bengio, T. South, A. Pentland, S. Koyejo, M. Kochenderfer, and R. Trager. Open problems in technical AI governance. *Transactions on Machine Learning Research*, 2025. Survey Certification.

[108] D. M. Roijers. *Multi-Objective Decision-Theoretic Planning*. PhD thesis, Universiteit van Amsterdam, 2016.

[109] D. M. Roijers and S. Whiteson. *Multi-Objective Decision Making*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer, 2020.

[110] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[111] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 2016.

[112] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks, 2023.

[113] T. Shen, R. Jin, Y. Huang, C. Liu, W. Dong, Z. Guo, X. Wu, Y. Liu, and D. Xiong. Large language model alignment: A survey, 2023.

[114] Z. Shi, S. Gao, X. Chen, Y. Feng, L. Yan, H. Shi, D. Yin, P. Ren, S. Verberne, and Z. Ren. Learning to use tools via cooperative and interactive agents. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10642–10657, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[115] J. M. V. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward gaming. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[116] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters, 2024.

[117] Y. Sui, A. Gotovos, J. Burdick, and A. Krause. Safe exploration for optimization with Gaussian processes. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 997–1005, Lille, France, July 2015. PMLR.

[118] T. Sumers, S. Yao, K. Narasimhan, and T. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024. Survey Certification.

[119] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. In *Ethics of Artificial Intelligence*. Oxford University Press, September 2020. `https://intelligence.org/files/AlignmentMachineLearning.pdf`.

[120] O. Thawakar, D. Dissanayake, K. More, R. Thawkar, A. Heakl, N. Ahsan, Y. Li, M. Zumri, J. Lahoud, R. M. Anwer, H. Cholakkal, I. Laptev, M. Shah, F. S. Khan, and S. Khan. LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs, 2025.

[121] The Business Software Alliance. 2025 state AI wave building after 700 bills in 2024, October 22 2024. `https://www.bsa.org/news-events/news/2025-state-ai-wave-building-after-700-bills-in-2024`.

[122] A. Turner, N. Ratzlaff, and P. Tadepalli. Avoiding side effects in complex environments. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21406–21415. Curran Associates, Inc., 2020. `https://proceedings.neurips.cc/paper_files/paper/2020/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf`.

[123] A. M. Turner, D. Hadfield-Menell, and P. Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 385–391, New York, NY, USA, 2020. Association for Computing Machinery.

[124] A. M. Turner, L. R. Smith, R. Shah, A. Critch, and P. Tadepalli. Optimal policies tend to seek power. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[125] E. V. *Algorithms for Simple Stochastic Games*. PhD thesis, University of South Florida, 2009.

[126] D. Verma and R. Rao. Graphical models for planning and imitation in uncertain environments. Technical Report 2005-02-01, University of Washington, 2005.

[127] J. Wang, M. Fang, Z. Wan, M. Wen, J. Zhu, A. Liu, Z. Gong, Y. Song, L. Chen, L. M. Ni, L. Yang, Y. Wen, and W. Zhang. OpenR: An open source framework for advanced reasoning with large language models, 2024.

[128] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z.-Y. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science Article*, 18:186345, 2024.

[129] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[130] T. Wang, J. Liu, B. Lee, Z. Wu, and Y. Wu. OCMDP: Observation-constrained Markov decision process, 2025.

[131] T. Wang, Y. Zhang, S. Qi, R. Zhao, Z. Xia, and J. Weng. Security and privacy on generative data in AIGC: A survey. *ACM Computing Surveys*, 57(4), December 2024.

[132] Y. Wang, S. S. Zhan, R. Jiao, Z. Wang, W. Jin, Z. Yang, Z. Wang, C. Huang, and Q. Zhu. Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36593–36604. PMLR, July 2023.

[133] Z. Wang, Y. Li, Y. Wu, L. Luo, L. Hou, H. Yu, and J. Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7309–7319, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[134] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

[135] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[136] S. Wiener, R. Roth, S. Rubio, and H. Stern. Senate Bill No. 1047, August 2024.

[137] B. Wiser. Here's what happens when your lawyer uses ChatGPT, May 2023. New York Times, `https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html`.

[138] Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models, 2024.

[139] Z. Wu and Q. He. Optimal switching sequence for switched linear systems. *SIAM Journal on Control and Optimization*, 58(2):1183–1206, 2020.

[140] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui. The rise and potential of large language model based agents: A survey, 2023.

[141] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, C. Shao, Y. Yan, Q. Yang, Y. Song, S. Ren, X. Hu, Y. Li, J. Feng, C. Gao, and Y. Li. Towards large reasoning models: A survey on scaling LLM reasoning capabilities, 2025.

[142] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. R. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[143] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[144] P. Yu, J. Xu, J. Weston, and I. Kulikov. Distilling system 2 into system 1, 2024.

[145] D. Zhang, J. Wu, J. Lei, T. Che, J. Li, T. Xie, X. Huang, S. Zhang, M. Pavone, Y. Li, W. Ouyang, and D. Zhou. LLaMA-Berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning, 2024.

[146] D. Zhang, S. Zhoubian, Z. Hu, Y. Yue, Y. Dong, and J. Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[147] Y. Zhang, S. Mao, T. Ge, X. Wang, Y. Xia, W. Wu, T. Song, M. Lan, and F. Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. In *First Conference on Language Modeling*, 2024.

[148] Y. Zhang, S. Wu, Y. Yang, J. Shu, J. Xiao, C. Kong, and J. Sang. O1-coder: An o1 replication for coding, 2024.

A Biden's executive order and related bills

In October 2023, Joe Biden, then president of the US, signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [18][5]. Its Section 4.2 is dedicated to ensuring safe and reliable AI. In particular, it requires companies to report on "any model that was trained using a quantity of computing power greater than 10^{26} integer or [FLOPs]" until a set of technical conditions for models are defined by specified authorities (see Section 4.2(b)).

In this executive order, particular attention is paid to a dual-use FM, which refers to an FM that exhibits "high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters, such as by ... permitting the evasion of human control or oversight through means of deception or obfuscation" (see Section 3(k)).

Following this executive order, almost 700 AI-related bills are introduced in 45 states across the United States in 2024 [121]. A particularly interesting one is California Senate Bill 1047 (Safe and Secure Innovation for Frontier Artificial Intelligence Models Act) [136][6]. Its Chapter 22.6 is devoted to safe and secure innovation for frontier AI models, which cover "[a]n artificial intelligence model trained using a quantity of computing power greater than 10^{26} integer or floating-point operations." In particular, the senate bill requires that "[b]efore beginning to initially train a covered model, the developer shall ... [i]mplement the capability to promptly enact a full shutdown," which completely halts the operations of the model.

The necessity of such an off-switch [53] is motivated to prevent "critical harms," which include "[m]ass casualties or at least five hundred million dollars ($500,000,000) of damage resulting from an artificial intelligence model engaging in conduct that ... [a]cts with limited human oversight, intervention, or supervision."

B Similarity between AI agents

In this section, we explore a method for determining whether an AI agent is acceptable, independent of how the agent is constructed—including factors such as training compute or architectural configuration—without relying on a detailed examination of its individual outputs. We may say that an agent is acceptable when it is sufficiently similar to another agent that has been empirically validated as strongly acceptable. A question is how to evaluate the similarity between two agents. If the two agents consist of identical base FMs and identical planning modules, their differences may be attributed to variations in inference compute (e.g., the number of search steps performed in MCTS). However, in practice, AI agents can vary widely in their base models, reasoning strategies, and system architectures. This diversity means that a single threshold based on inference compute is unlikely to serve as a universal criterion for acceptability. Also, it is unclear what constitute a single run of inference, as we have discussed in Section 3.2.

A possible approach is to measure the similarity between two agents, G and G', based on their output distributions. For example, using a distance d between probability distributions, we may for example define the distance between G and G' with $\sup_x d(g(G,x), g(G',x))$, where $g(G,x)$ is the distribution of the output (e.g., document) of G when it perceives x (e.g., prompt). This is similar to the motivation of reinforcement learning from human feedback [97, 11] and direct policy optimization [105], where the regularization with Kullback–Leibler divergence is used to mitigate catastrophic forgetting or alignment tax [75, 97, 11], which refer to the phenomena that fine-tuned models lose the skills that pre-trained models had.

Likewise, we may mitigate catastrophic forgetting of the strong acceptability of an agent G' and preserve the acceptability in a new agent G by ensuring $\sup_x d(g(G,x), g(G',x)) \le \varepsilon$. A difficulty is that it is unclear how to evaluate the supremum, since there can be infinitely many possible perceptions x. Alternatively, one may consider $\mathbb{E}[d(g(G,X), g(G',X))]$, where $\mathbb{E}$ is the expectation with respect to some distribution of the random perception X. However, such a guarantee based

[5]This executive order was repealed by President Trump.

[6]The bill had passed the state legislature but was later vetoed by the Governor. The technical feasibility of the requirements has also been questioned by the community [2].

on expectation may be unsuitable for addressing safety concerns related to existential risks, which involve events with extremely low probabilities and extremely high impacts.

C Rationale for similarity-based measures

All the measures that are considered in this paper are more or less based on similarity: we say that anything is acceptable when it is sufficiently close to something that is strongly acceptable. Here, we provide some rationale on such similarity-based measures. Let a solution z denote a configuration, an AI agent, or an action-sequence; we discuss the acceptability of z.

Suppose that there exists an unknown function g_0 such that a solution z is acceptable iff $g_0(z) \leq 0$. We cannot make strong assumptions about g_0, since we know little about g_0. Since we cannot deal with g_0 without any assumptions, let us make a minimal assumption that the solution space $\mathcal{Z}$ is equipped with a metric d, and that g_0 is 1-Lipschitz.[7] Let $\mathcal{G}$ be a class of 1-Lipschitz functions.

We say that a solution z_0 is strongly acceptable if all of its ε-neighbors are acceptable. Let $\mathcal{Z}_0 \subseteq \{z \in \mathcal{Z} \mid g(z_0) \leq -\varepsilon\}$ be the set of known strongly acceptable solutions. Then we know that $\bar{\mathcal{Z}}_0 := \{z \in \mathcal{Z} \mid \min_{z_0 \in \mathcal{Z}_0} d(z, z_0) \leq \varepsilon\}$ is a set of acceptable solutions. One would typically choose the solution z that maximizes an objective function under the constraint of $z \in \bar{\mathcal{Z}}_0$. In this way, the selected solution is guaranteed to be acceptable under the assumptions made.

When we say that a solution is acceptable based on a similarity measure, it is based on the assumptions that can be summarized with a tuple $(\mathcal{Z}_0, d, \varepsilon)$. Namely, what solutions are assumed to be strongly acceptable ($\mathcal{Z}_0$), how the similarity between solutions is measured (d), and what level of guarantee is made (ε). For example, $\mathcal{Z}_0$ may be set of all the FMs that are trained with at most 10^{24} FLOPs of computation, d may be the difference in FLOPs measured in $\log_{10}$, and $\varepsilon = 1$. Alternatively, $\mathcal{Z}$ may be the set of action sequences of length at most 1000, d may be the difference in the length of the action-sequences, and $\varepsilon = 1$. Governments may then design regulations based on the tuple $(\mathcal{Z}_0, d, \varepsilon)$. Scientists may provide guidelines regarding what tuples should be used in regulations.

Mathematical guarantees, such as safety under Lipschitz continuity, provide useful design principles for real-world systems. However, real-world safety cannot always be directly derived from such mathematical models, as assumptions in these models may not fully capture the complexity of actual systems. Even in highly regulated domains like aviation and nuclear safety, failures occasionally occur despite adherence to rigorous safety guidelines.

In our approach, we rely on the following key assumptions:

- Any single action does not lead to catastrophic failure (e.g., human extinction) as long as it is taken from a strongly acceptable state.
- An action-sequence of length $N - 1$ is strongly acceptable and leads to a strongly acceptable state.

Here, a strongly acceptable state refers to a state that is reachable via strongly acceptable action sequences, starting from any initial states that normally occur in the absence of the agent. Based on these assumptions, we reason about the safety of an action sequence of length N by ensuring that the N-th action is chosen from a strongly acceptable state. The validity of the assumptions depends on verification through empirical studies rather than theoretical guarantees.

We acknowledge that these assumptions do not hold for arbitrary action spaces. For instance, if the action space includes the action of launching a nuclear weapon, then a single action could lead to catastrophe. As we have discussed in Section 4.2, AI agents should be designed without such high-risk actions in their action space, and regulatory frameworks should prohibit AI agents that have access to actions with individually catastrophic consequences. More broadly, the design of the action space must be carefully structured to ensure practical safety.

[7] This does not lose generality, since L-Lipschitz functions under a metric d can be made 1-Lipshitz by redefining d.

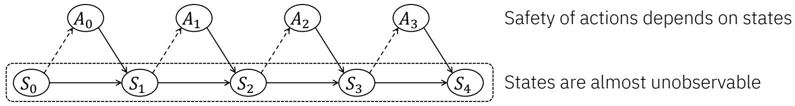

Figure 2: Safety of actions in a nearly unobservable Markov decision process

D Expanding the strongly acceptable set

The set of strongly acceptable solutions may be expanded gradually. For example, we may choose an acceptable solution $z \in \bar{\mathcal{Z}}_0$ and keep using z for a certain period of time. If it turns out that z can be considered strongly acceptable based on its behavior during that period, we may expand the set of strongly acceptable solutions by adding z into $\mathcal{Z}_0$. This process of expanding the acceptable set $\mathcal{Z}_0$ could also be performed jointly as a community.

This expansion of acceptable set is similar in spirit to safe exploration in RL [47]. Here, the safety set is gradually expand, starting from a seed set, for example based on the assumption of Lipshitz continuity and Gaussian process [117]. In control theory, safety is often guaranteed with barrier certificates often based on some prior knowledge about the environment [4, 86, 14]. Such ideas have also been exploited in safe exploration in RL with generative modeling [132].

E Unobservable Markov decision processes

The proposed approach may be considered as a way to ensure safety of actions in a Markov decision process where states are nearly unobservable (see Figure 2). In this section, we provide a comprehensive survey on a related model of unobservable Markov decision processes (UMDPs)[8], for which there has been a limited amount of prior work.

An UMDP is a special case of a partially observable Markov decision process (POMDP) in that the agent always makes a null observation in the UMDP. A standard approach to finding the optimal policy for a POMDP is to recursively compute its value as a function of the belief state, which is updated on the basis of the Bayes rule. This is substantially simplified when there are no observations. In this section, we provide a comprehensive survey of the prior work on UMDPs.

[87, 88] establish the undecidability of some decision problems associated with POMDPs over infinite horizons by establishing undecidability for the special case of UMDPs. Such undecidability for UMDPs can be established by reducing an UMDP to a probabilistic finite-state automaton. The undecidability also holds for a restricted class of UMDPs [12, 13]. While approximate decision problems are still undecidable for general UMDPs over infinite horizons, [29] study a special case of UMDPs whose approximate decision problems are decidable.

[26] prove that computing the optimal policy for a POMDP over a finite horizon is NP-hard but showing that it is NP-hard for an UMDP. [139] study special cases of UMDPs over finite horizons whose optimal policies can be computed in polynomial time.

UMDPs have also been used as approximations of POMDPs [54, 24, 79] or simply discussed as a special case of POMDPs [125]. For a given POMDP, the corresponding UMDP can given a lower bound on the value function, while the corresponding (fully observable) MDP can give an upper bound on the value function. This relation between UMDP and POMDP can be exploited to efficiently find approximately optimal policies for POMDPs. Notice that an UMDP can allow more efficient optimization than the corresponding POMDP, since the UMDP does not need to deal with observations. For exemple, [73] studies an MCTS method for UMDPs.

UMDPs have also been studied as a simple special case of POMDPs to study the effectiveness of planning methods in belief states to study the relative performance of different planning methods for POMDPs [81, 82, 72]. UMDPs have also been simply discussed as a special case of POMDPs [22, 34, 126].

[8]UMDPs have also been studied under the name of Markov decision processes (MDPs) with no observations, non-observable MDPs (NOMDPs), and no observation MDPs (NOMDPs).

UMDP also appears in the study of planning for multiple distributed agents to optimize a single objective under partial observability. Specifically, planning for a decentralized POMDP (DecPOMDP) [93] can be reduced to planning for a (centralized) UMDP [92, 108, 109], where the state in the UMDP is the pair of the state and the history of observations in the DecPOMDP, and the action in the UMDP is the decision rule that maps the history of observations into the actions in the POMDP. In DecPOMDPs, the belief (distribution) over the pair of the state and the history of observations is the sufficient statistic, and planning can be performed in the space of the belief states.

[41] consider an UMDP in the context of studying what values observations can provide in a POMDP. They introduce a parameter that ranges from 0 to 1. When the parameter is 0, the observation provides no information about the state (hence, the POMDP reduces to an UMDP); when the parameter is 1, the observation provides full information about the state (hence, the POMDP reduces to an MDP). The prior work also extends UMDPs to allow often costly actions that enable partial or full observations of the state [43, 68, 69, 130].