POLICY FORUM

ARTIFICIAL INTELLIGENCE

Regulating advanced artificial agents

Governance frameworks should address the prospect of AI systems that cannot be safely tested

By **Michael K. Cohen[1,2], Noam Kolt[3,4], Yoshua Bengio[5,6], Gillian K. Hadfield[2,3,4,7], Stuart Russell[1,2]**

Technical experts and policy-makers have increasingly emphasized the need to address extinction risk from artificial intelligence (AI) systems that might circumvent safeguards and thwart attempts to control them (*1*). Reinforcement learning (RL) agents that plan over a long time horizon far more effectively than humans present particular risks. Giving an advanced AI system the objective to maximize its reward and, at some point, withholding reward from it, strongly incentivizes the AI system to take humans out of the loop, if it has the opportunity. The incentive to deceive humans and thwart human control arises not only for RL agents but for long-term planning agents (LTPAs) more generally. Because empirical testing of sufficiently capable LTPAs is unlikely to uncover these dangerous tendencies, our core regulatory proposal is simple: Developers should not be permitted to build sufficiently capable LTPAs, and the resources required to build them should be subject to stringent controls.

Governments are turning their attention to these risks, alongside current and anticipated risks arising from algorithmic bias, privacy concerns, and misuse. At a 2023 global summit on AI safety, the attending countries, including the United States, United Kingdom, Canada, China, India, and members of the European Union (EU), issued a joint statement warning that, as AI continues to advance, "Substantial risks may arise from…unintended issues of control relating to alignment with human intent" (*2*). This broad consensus concerning the potential inability to keep advanced AI under control is also reflected in President Biden's 2023 executive order that introduces reporting requirements for AI that could "eva[de] human control or oversight through means of deception or obfuscation" (*3*). Building on these efforts, now is the time for governments to develop regulatory institutions and frameworks that specifically target the existential risks from advanced artificial agents.

RISKS FROM LTPAs

RL agents function as follows: They receive perceptual inputs and take actions, and certain inputs are typically designated as "rewards." An RL agent then aims to select actions that it expects will lead to higher rewards. For example, by designating money as a reward, one could train an RL agent to maximize profit on an online retail platform (*4*).

> "…safety testing is likely to be either dangerous or uninformative."

Highly capable and far-sighted RL agents are likely to accrue reward very successfully. If plan A leads to more expected reward than plan B, sufficiently advanced RL agents would favor the former. Crucially, securing the ongoing receipt of maximal rewards with very high probability would require the agent to achieve extensive control over its environment, which could have catastrophic consequences (*5*–*8*). One path to maximizing long-term reward involves an RL agent acquiring extensive resources and taking control over all human infrastructure (*5*, *6*), which would allow it to manipulate its own reward free from human interference (*5*). Additionally, because being shut down by humans would reduce the expected reward, sufficiently capable and far-sighted agents are likely to take steps to preclude that possibility (*7*) or if feasible, create new agents (unimpeded by monitoring or shutdown) to act on their behalf (*5*). Progress in AI could enable such advanced behavior.

So long as an agent's rewards can be controlled, it can be incentivized to achieve complex goals by conditioning the rewards appropriately. But a sufficiently capable RL agent could take control of its rewards, which would give it the incentive to secure maximal reward single-mindedly. Constraining the influence that highly competent agents learn to exert over their environment is likely to prove extremely difficult; an intelligent agent could, for example, persuade or pay unwitting human actors to execute important actions on its behalf (*5*, *7*).

Critically, far-sighted RL agents face an incentive to develop and execute arbitrarily competent long-term plans. Many AI systems are trained only to achieve certain immediate outcomes, like correctly classifying an image. Although such short-sighted agents could certainly cause harm, they would likely lack the incentive to execute protracted schemes to subvert human control.

Accordingly, we define an LTPA as an algorithm designed to produce plans, and to prefer plan A to plan B, when it expects that plan A is more conducive to a given goal over a long time horizon. For example, an agent trained to maximize profit on an online retail platform, as proposed by Suleyman's "new Turing test" (*4*), might productively use such an algorithm and hinder attempts to interfere with its profit making. LTPAs include all long-horizon RL algorithms, including so-called "policy gradient" methods, which lack an explicit planning subroutine but are trained to be as competent as possible. LTPAs also include algorithms that imitate trained LTPAs, but not algorithms that merely imitate humans. In the latter case, if plan A is more competent than any plan a human could develop, and plan B is a human plan, an algorithm imitating a human would not prefer plan A to plan B. The supplementary materials include a taxonomy situating LTPAs among other machine learning systems. Notably, there is no recognizable horizon length at which risk increases sharply; accordingly, regulators will have

[1] University of California, Berkeley, CA, USA. [2] Center for Human-Compatible Artificial Intelligence, Berkeley, CA, USA. [3] University of Toronto, Toronto, Ontario, Canada. [4] Schwartz Reisman Institute for Technology and Society, Toronto, Ontario, Canada. [5] Université de Montréal, Montréal, Québec, Canada. [6] Mila–Quebec AI Institute, Montréal, Québec, Canada. [7] Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. Email: mkcohen@berkeley.edu

to define the length of a long time horizon according to their risk tolerance.

Losing control of advanced LTPAs, although not the only existential risk from AI, is the class of risk that we aim to address here—and one that necessitates new forms of government intervention.

A GOVERNANCE PROPOSAL

Although governments have expressed concern about existential risks from AI, regulatory proposals do not adequately address this class of risk (*9*). The EU AI Act (*10*) canvasses a broad array of risks from AI but does not single out loss of control of advanced LTPAs. We see promising first steps from the US and UK—President Biden's executive order on AI (*3*) requires reports on potentially uncontrollable AI systems, but it does not seek to constrain their development or proliferation; the US and UK AI Safety Institutes are building capacity for regulators to understand cutting-edge AI but lack the authority to control it (*11*).

Across multiple jurisdictions, following industry practice, the prevailing regulatory approach for AI involves empirical safety testing, most prominently within the UK AI Safety Institute (*2, 3, 10–12*). We, however, argue that for a sufficiently capable LTPA, safety testing is likely to be either dangerous or uninformative. Although we might like to empirically assess whether an agent would exploit an opportunity to thwart our control over it, if the agent in fact has such an opportunity during a test, the test may be unsafe. Conversely, if it does not have such an opportunity during a test, the test is likely to be uninformative with respect to such risks. This holds for human agents as well as artificial agents: Consider a leader appointing a general, but worried about a coup; if the general is clever, there is no safe and reliable loyalty test. A candidate for the role, like an advanced artificial agent, would either recognize the test and behave agreeably or, if possible, execute a coup during the test.

If an agent is advanced enough to recognize that it is being tested, then there is little reason to expect similar behavior in and out of testing. Moreover, an AI system designed to interact with complex environments (e.g., human institutions or biological systems) would likely be able to discern a simulated test environment from real-world deployment (because complex systems can only be simulated approximately), thereby enabling the AI system to identify when it is being tested. Although no current artificial agents are competent enough to thwart human control, some have already been found to identify safety tests and pause misbehavior (*13*). Testing may nonetheless be useful for detecting some dangerous algorithmic capabilities in systems that cannot thwart human control.

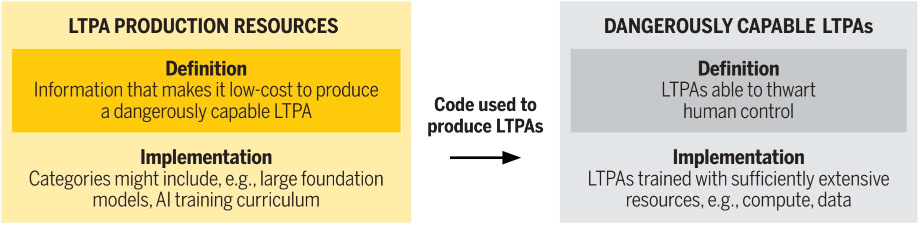
Stepping back, empirical testing is a notoriously ineffective tool for ensuring the safety of computational systems. For example, extensive testing failed to reveal an error in the Intel Pentium's arithmetic unit. Given that both safety and validity cannot be ensured when testing sufficiently capable LTPAs, governments should establish new regulatory bodies with the legal authority and technical capacity to prevent such agents from being built in the first place, no matter the domain.

DEFINING DANGEROUS CAPABILITIES

How capable is "sufficiently capable"? Unfortunately, we do not know. More cautious regulators might prevent the development of even weak LTPAs; however, regulators seeking to facilitate the development of merely "moderately capable" LTPAs should establish protocols to estimate in advance whether such systems might have the ability to game safety testing and evade human control. One factor that regulators could consider is the resources proposed to be used to train LTPAs, including compute, data, and the resources used to develop any pretrained models that assist in LTPA training. We propose that policy-makers (i) establish a list of dangerous capabilities, such as those described in President Biden's executive order (*3*), which include "high levels of performance at...deception or obfuscation" and "offensive cyber operations through automated vulnerability discovery"; and (ii) estimate the resources needed to develop an LTPA that exhibits those capabilities. We do not believe that existing systems exhibit those capabilities, and it is very difficult to predict when they could (*4*). This difficulty arises in part because there is currently no robust scientific method for (ii); computer scientists should develop one quickly. Perhaps if certain resources could be used to create an AI system with the short-term goal of exhibiting a moderately dangerous capability (i.e., trying to fail the safety test), that could improve our understanding of the resources that can produce dangerous capabilities.

Mandatory reporting and production controls for LTPAs

To prevent unlawful development of dangerously capable long-term planning agents (LTPAs), which may be difficult to directly detect, reporting requirements would enable regulators to have sufficient visibility into easier to observe LTPA production resources and code interacting with those resources. Though concern is ultimately with subsets of production resources and LTPAs ("definition"), these are not easily recognizable, thus broader recognizable supersets encompassing those subsets ("implementation") are the focus of regulation.



LTPA PRODUCTION RESOURCES

Definition
Information that makes it low-cost to produce a dangerously capable LTPA

Implementation
Categories might include, e.g., large foundation models, AI training curriculum

Code used to produce LTPAs →

DANGEROUSLY CAPABLE LTPAs

Definition
LTPAs able to thwart human control

Implementation
LTPAs trained with sufficiently extensive resources, e.g., compute, data

Shape and size of regulatory implementation categories can be updated periodically to ensure the inclusion of new systems that meet the definition.

Admittedly, listing relevant dangerous capabilities and estimating the resources required to achieve these capabilities will require considerable research. We suggest that regulators err on the side of caution and underestimate the resources required to develop LTPAs with dangerous capabilities. Systems 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. To ensure this occurs, regulators will need to carefully monitor and control the resources that could be used to produce dangerously capable LTPAs. Although this would interrupt the "move fast" ethos of AI development, we believe caution is necessary.

If dangerously capable LTPAs are at some point permitted to be developed, rigorous technical and regulatory work would need to be done first to determine if, when, and how to permit this. The possibility must also be considered that researchers and policy-makers fail to identify any robustly safe regulatory regimes that permit the development of dangerously capable LTPAs, at least by the time that actors in the private sector are able to build them. It is also worth noting that there might be a path to building AI systems that can be proved mathematically to avoid certain dangerous behaviors (*7*), but such formal guarantees appear highly unlikely for any AI systems built similarly to the most powerful systems today (*13*).

MONITORING AND REPORTING

Just as nuclear regulation extends to controlling uranium, AI regulation must extend to controlling the resources needed to produce dangerously capable LTPAs. We define production resources (PRs) as any information that makes the production of a dangerously capable LTPA cheaper than a threshold determined by regulators according to their risk tolerance. Unlike uranium, a PR is not a physical resource—it could include any AI model trained beyond a certain compute threshold (*14*). Fortunately, regulators could detect such PRs by following the hardware required to produce them. (Some of this hardware could be regulated as well, including semiconductor chips and data centers, but that is outside our focus here.) To limit the proliferation of PRs, expanding on Hadfield *et al*. (*15*), Avin *et al*. (*12*), and President Biden's executive order (*3*), we propose that developers be required to report (a) relevant facts about the PR [if the PR is an AI model, this might include (i) the input/output properties, (ii) the data collection process for training it, (iii) the training objective, and (iv) documented behavior in test settings, but not typically the model weights themselves]; (b) the specific machines on which the PR is stored and their locations; (c) all code run on these machines after the PR is created; and (d) all outputs of that code. With the context provided by point (a), governments could monitor the code that interacts with PRs, allowing them to detect the development of (unlawful) LTPAs (see the figure). In addition, if a company offers users application programming interface (API) access to a PR, users should be required to report the code on the user's machine that interacts with the API. Details of the reporting requirements will need to be updated in response to technological advances that lead to changes in the resources and processes needed to produce dangerously capable LTPAs. Finally, reporting procedures could be complemented by protecting and rewarding whistleblowers who uncover misconduct.

PRODUCTION CONTROLS

Given sufficient visibility into the resources for producing LTPAs, regulators could then prohibit the production of dangerously capable LTPAs. Developers that are unsure whether a proposed AI system meets the definition of dangerously capable LTPA could inquire with the relevant regulator prior to development. Regulators could also control the transfer of large pretrained models or other relevant resources. Further, regulators could make it unlawful for other actors to use AI systems that fail to comply with these requirements (*15*). Taken together, controls on the development, use, and dissemination of production resources will substantially reduce the likelihood of these resources being used to build dangerously capable LTPAs.

ENFORCEMENT MECHANISMS

To ensure compliance with these reporting requirements and usage controls, regulators may need to be authorized to (i) issue legal orders that compel organizations to report production resources and mandate the cessation of prohibited activities; (ii) audit an organization's activities and, where necessary, restrict an organization's access to certain resources, such as cloud computing; (iii) impose fines on noncompliant organizations; and (iv) as in financial regulation, impose personal liability on key individuals in noncompliant organizations. If business leaders can be held to account for breaching corporate duties, then surely they should face similar consequences for irresponsibly handling one of the world's most dangerous technologies.

REGULATORY INSTITUTIONS

We have addressed our discussion to "regulators" but have not proposed specific regulatory institutions for addressing the risks from LTPAs. This issue will need to be approached differently in different countries. That being said, we expect that whereas other risks from AI might be addressed primarily through domain-specific regulation (e.g., financial regulation and health care regulation), the risk of loss of control of AI likely requires specialized regulation and the establishment of new regulatory institutions. This specialized regulation could nevertheless benefit from the existing expertise of domain-specific regulators, including with developing frameworks for monitoring PRs. Critically, because the risks from LTPAs are global, regulatory efforts cannot stop at national borders. International cooperation is vital.

BROADER CONCERNS

LTPAs, of course, are not the only type of AI system that poses substantial and even existential risks. Accordingly, we suggest that empirical testing, which is inadequate for sufficiently advanced LTPAs, could nevertheless substantially improve the safety of some other types of AI. At the same time, the governance regime that we propose could be adapted to other AI systems. Although our proposal for governing LTPAs fills an important gap, further institutional mechanisms will likely be needed to mitigate the risks posed by advanced artificial agents. ■

REFERENCES AND NOTES

1. G. Hinton *et al*., "Statement on AI risk" (Center for AI Safety, May 2023); https://www.safe.ai/statement-on-ai-risk.
2. United Kingdom Department for Science, Innovation, and Technology, United Kingdom Foreign, Commonwealth, and Development Office, United Kingdom Prime Minister's Office, "The Bletchley Declaration by countries attending the AI Safety Summit, 1-2 November 2023" (Gov.uk, November 2023); https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.
3. J. Biden, "Executive order on the safe, secure, and trustworthy development and use of artificial intelligence" (The White House, October 2023); https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
4. M. Suleyman, *The Coming Wave* (Penguin, September 2023).
5. M. K. Cohen, M. Hutter, M. A. Osborne, *AI Mag*. **43**, 282 (2022).
6. S. Zhuang, D. Hadfield-Menell, *Adv. Neural Inf. Process. Syst*. **33**, 15763 (2020).
7. S. Russell, *Human Compatible: AI and the Problem of Control* (Viking, 2019).
8. A. Turner, L. Smith, R. Shah, A. Critch, P. Tadepalli, *Adv. Neural Inf. Process. Syst*. **34**, 23063 (2021).
9. N. Kolt, "Algorithmic black swans"(Washington University Law Review, October 2023); https://ssrn.com/abstract=4370566.
10. European Commission, "Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts" (COM/2021/0206, European Commission, January 2024); https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.
11. United Kingdom Department for Science, Innovation and Technology, "Introducing the AI Safety Institute" (Gov.uk, November 2023); https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf.
12. S. Avin *et al*., *Science* **374**, 1327 (2021).
13. J. Lehman *et al*., *Artif. Life* **26**, 274 (2020).
14. Y. Shavit, arXiv:2303.11341 [cs.LG] (2023).
15. G. Hadfield, M. Cuéllar, T. O'Reilly, "It's time to create a national registry for large AI models" (Carnegie Endowment for International Peace, July 2023); https://carnegieendowment.org/2023/07/12/it-s-time-to-create-national-registry-for-large-ai-models-pub-90180.

ACKNOWLEDGEMENTS

M.K.C. and N.K. contributed equally. The authors thank R. Grosse, A. Barto, and P. Christiano for feedback on earlier versions of the manuscript. M.K.C. commenced this project at Oxford University. M.K.C. and S.R. are supported by the Open Philanthropy Foundation. N.K. is supported by a Vanier Canada Graduate Scholarship. Y.B. is supported by a Canadian Institute for Advanced Research (CIFAR) AI Chair and a Natural Sciences and Engineering Research Council of Canada Discovery Grant. G.K.H. is supported by the Schwartz Reisman Institute Chair in Technology and Society and a Canada CIFAR AI Chair at the Vector Institute for Artificial Intelligence. G.K.H. and S.R. are supported by AI2050 Senior Fellowships from Schmidt Futures.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adl0625

10.1126/science.adl0625

Science
AAAS

Supplementary Materials for

Regulating advanced artificial agents

Michael K. Cohen *et al.*

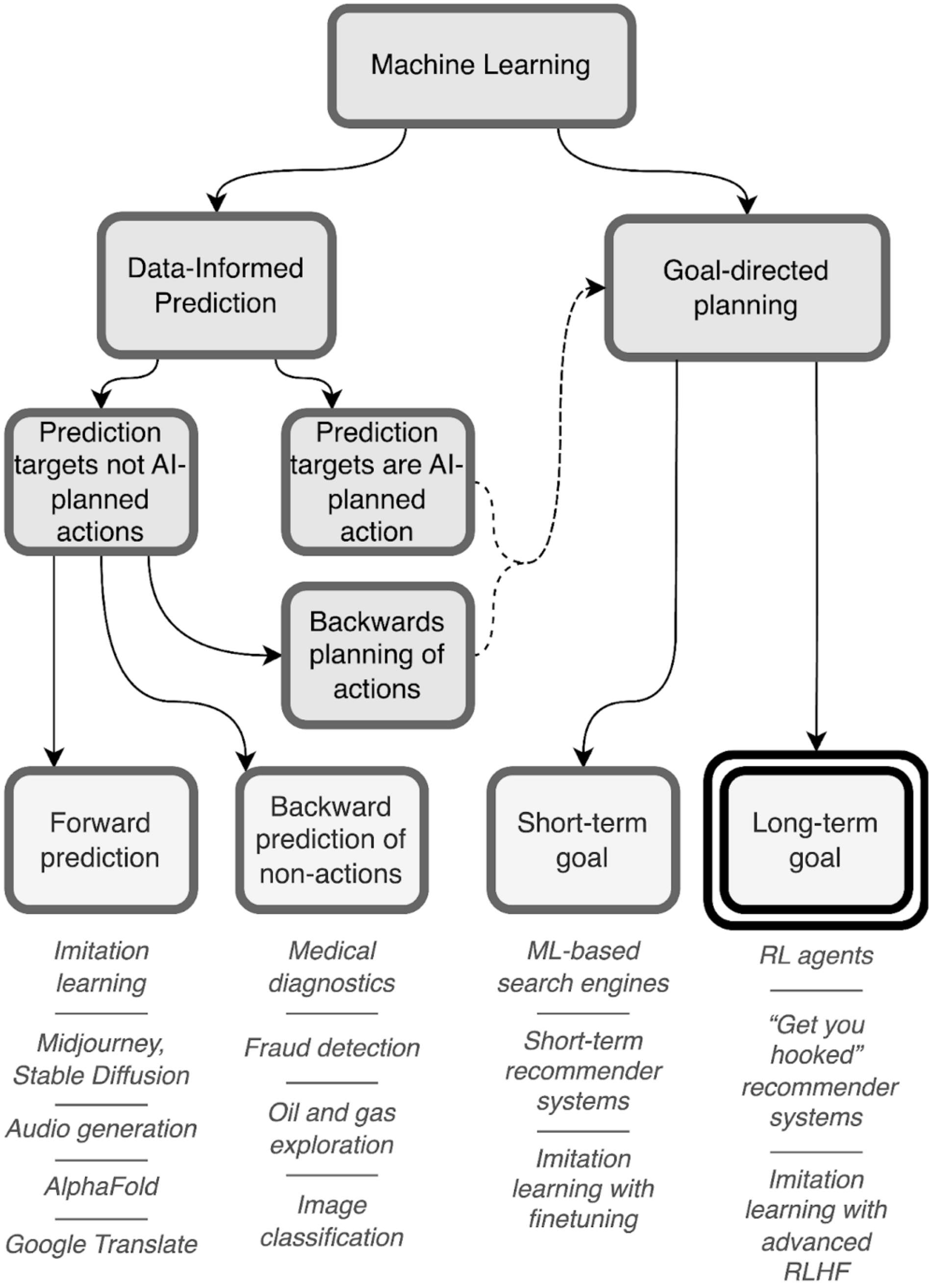
Michael K. Cohen, mkcohen@berkeley.edu

Science **384**, 36 (2024)
DOI: 10.1126/science.adl0625

The PDF file includes:

Fig. S1

Fig. S1. Sufficiently advanced long-term planning agents (LTPAs) cannot be safely tested. In contrast, other AI systems may be less incentivized to deliberately game a safety test. A full explanation follows in the text.


Machine Learning
Data-Informed Prediction
Goal-directed planning
Prediction targets not AI-planned actions
Prediction targets are AI-planned action
Backwards planning of actions
Forward prediction
Backward prediction of non-actions
Short-term goal
Long-term goal
Imitation learning
Midjourney, Stable Diffusion
Audio generation
AlphaFold
Google Translate
Medical diagnostics
Fraud detection
Oil and gas exploration
Image classification
ML-based search engines
Short-term recommender systems
Imitation learning with finetuning
RL agents
"Get you hooked" recommender systems
Imitation learning with advanced RLHF

As fig. S1 illustrates, many economically valuable AI applications do not use RL. In **data-informed prediction**, the core of machine learning, a system learns from historical examples to make predictions in new contexts. For example, large language models learn which words are likely to follow in a given sequence of text, based on examples of existing text.

This contrasts with **goal-directed planning** agents, in which an algorithm generates or refines a plan (or more generally, a conditional strategy), searching for one that achieves a goal, and preferring Plan A to Plan B whenever A is recognizably better, often using data-informed prediction to make such judgments. Not all goal-achieving algorithms, however, are goal-directed planning agents for our purposes. For example, consider a language model trained only on human-generated text. While the model might produce text that is conducive to a goal (e.g., increasing customer satisfaction), the training process of the model merely selects text or behavior resembling the training data. Crucially, the training algorithm does not optimize a strategy to *best* achieve a goal, so the research we cite offers no argument that it would present an existential risk, no matter how advanced the language model. The mere presence of agent-like behavior does not imply that the system has the inclination or ability to thwart human control. We have not argued definitively that human-imitating emergent agents *wouldn't* thwart human control; we only note that we lack strong arguments that they *would*.

In data-informed prediction, the algorithm predicts *y* given *x* after learning from examples (i.e., training data). There are two ways that data-informed prediction can give rise to *arbitrarily* advanced goal-directed planning: first, through **predicting the actions of goal-directed planning AI**. If some *y*'s in the training data are actions planned by a goal-directed AI, then the resulting system could itself perform advanced planning. For example, automatically predicting a chess engine's behavior produces another chess engine, simply by playing the predicted moves. Second, **backwards planning**, which involves identifying which actions must have preceded a desired outcome, could occur in data-informed prediction if some *y*'s in the training data are actions, and the *x*'s are the settings and the desired outcomes. For example, if *x* equals "checkmate from [insert chess position]", and if *y* equals "Qa8", then *y* is the move that led to the desired outcome described in *x*. Identifying an action likely to cause a desired outcome is the hallmark of goal-directed planning. Industrial datasets could perhaps, very expensively, be cleaned to avoid both categories.

In goal-directed planning, algorithms could be designed to select actions merely for a **short-term goal**. For example, a search engine selects the links most likely to be clicked. However, given the agent's short time horizon, it lacks the incentive to pursue protracted plans for thwarting human control. Meanwhile, in other cases, the system is designed to select actions for a **long-term goal**. For instance, recommender systems could select videos in order to have lasting impacts on users, namely making them avid viewers. It is this class of AI systems—**long-term planning agents (LTPAs)**, including RL agents that plan over long time horizons, that existing literature on existential risk from AI focuses on (*5*, *7*).