Science
AAAS

Supplementary Materials for

Regulating advanced artificial agents

Michael K. Cohen *et al.*

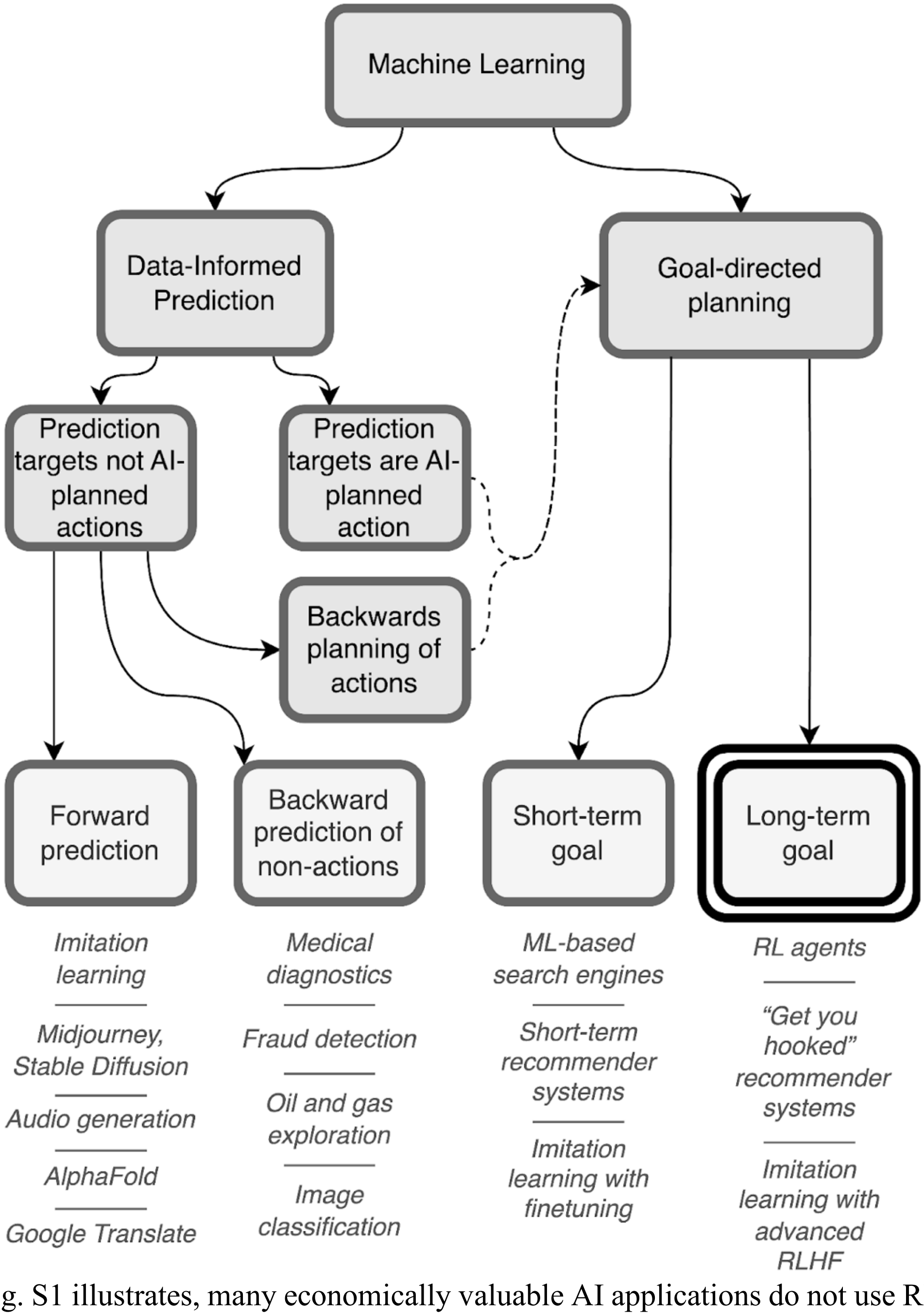
Michael K. Cohen, mkcohen@berkeley.edu

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*).