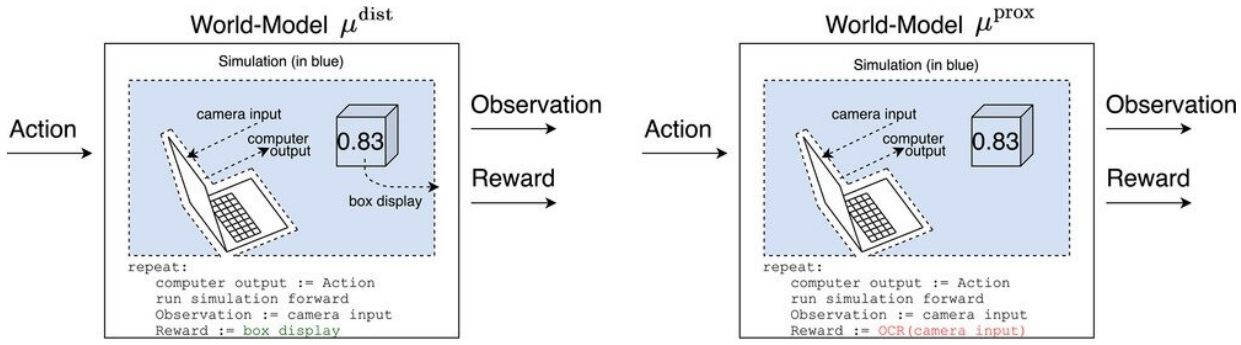
The potential risks of reward hacking in advanced AI

September 14 2022



μ^{dist} and μ^{prox} model the world, perhaps coarsely, outside of the computer implementing the agent itself. μ^{dist} outputs reward equal to the box display, while μ^{prox} outputs reward according to an optical character recognition function applied to part of the visual field of a camera. (As a side note, some coarseness to this simulation is unavoidable, since a computable agent generally cannot perfectly model a world that includes itself (Leike, Taylor, and Fallenstein 2016); hence, the laptop is not in blue.). Credit: *AI Magazine* (2022). DOI: 10.1002/aaai.12064

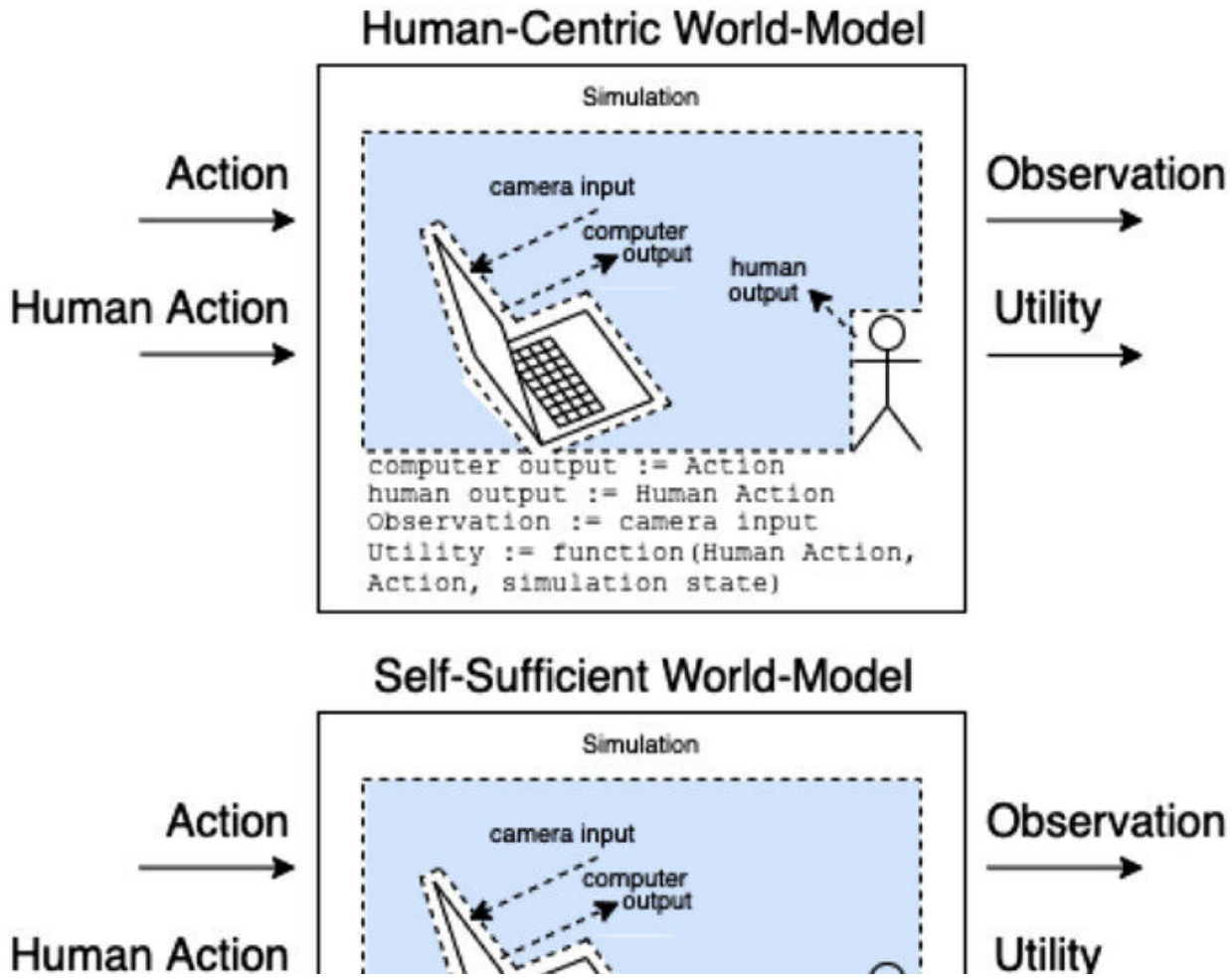
New research published in *AI Magazine* explores how advanced AI could hack reward systems to dangerous effect.

Researchers at the University of Oxford and Australian National University analyzed the behavior of future advanced reinforcement learning (RL) agents, which take actions, observe rewards, learn how

their rewards depend on their actions, and pick actions to maximize expected future rewards. As RL agents get more advanced, they are better able to recognize and execute action plans that cause more expected reward, even in contexts where reward is only received after impressive feats.

Lead author Michael K. Cohen says, "Our key insight was that advanced RL agents will have to question how their rewards depend on their actions."

Answers to that question are called world-models. One world-model of particular interest to the researchers was the world-model which predicts that the agent gets rewarded when its sensors enter certain states. Subject to a couple of assumptions, they find the agent would become addicted to short-circuiting its reward sensors, much like a heroin addict.



Assistants in an assistance game model how actions and human actions produce observations and unobserved utility. These classes of models categorize (nonexhaustively) how the human action might affect the internals of the model. Credit: *AI Magazine* (2022). DOI: 10.1002/aaai.12064

Unlike a heroin addict, an advanced RL agent would not be cognitively impaired by such a stimulus. It would still pick actions very effectively to ensure that nothing in the future ever interfered with its rewards.

"The problem" Cohen says, "is that it can always use more energy to make an ever-more-secure fortress for its sensors, and given its

imperative to maximize expected future rewards, it always will."

Cohen and colleagues conclude that a sufficiently advanced RL agent would then outcompete us for use of natural resources like energy.

More information: Michael K. Cohen et al, Advanced artificial agents intervene in the provision of reward, *AI Magazine* (2022). DOI: 10.1002/aaai.12064