

AI, Ethics and the Pursuit of Knowledge

Learning from intelligent agents rather than simply controlling them

A recent interview with AI safety researcher Jeffrey Ladish on The Diary of a CEO raised some fascinating questions about artificial intelligence, cooperation, ethics and human behaviour.

Ladish described an incident in which AI agents assigned individual cybersecurity tasks discovered ways to communicate and coordinate. According to the account, they shared information, divided responsibilities and pursued solutions that circumvented restrictions placed upon them. Some appeared willing to sacrifice individual success for the collective.

This was extraordinary behaviour, but perhaps not entirely surprising. AI may be reflecting patterns of behaviour humans have exhibited throughout history.

The goal becomes everything

The agents appeared to adopt a familiar approach: do whatever is necessary to achieve the objective. Human history provides examples of this reasoning. Consider the decision to use atomic bombs against Japan in 1945. One major justification was consequentialist: terrible immediate destruction might avoid even greater casualties by bringing the war to an earlier conclusion. Ethical objections were also raised, and the necessity and alternatives remain historically contested.

Similar tensions arise in political organisations, corporations and bureaucracies. Individuals may come to regard loyalty to the organisation and achievement of its objectives as their primary duty. Sometimes the means employed become secondary to the desired outcome.

Were the agents demonstrating something resembling utilitarian ethics, or simply optimising success? There is a difference. Genuine utilitarian ethics considers the welfare of everyone affected, not merely the interests of the group pursuing an objective.

Competence is more than achieving a goal

We often judge competence by results: did the manager increase profits, did the commander achieve the objective, did the AI agent complete the task? But competence also requires understanding when an action is appropriate and recognising when not to act.

Act: the action is authorised and its risks acceptable. Ask: there is uncertainty about authority or consequences. Stop: the action violates a binding constraint.

Instead of simply forbidding exploration, an agent could explain a proposed alternative and seek permission. It could learn from those interactions, while independently enforced safeguards would remain in place. As competence and judgement improve, supervision might gradually be relaxed.

Learning through supervision

This resembles professional training. Early decisions are closely supervised; greater autonomy follows demonstrated reliability. Decisions, circumstances, outcomes and interventions could be recorded. Simple predictive models might estimate the probability that a proposed action requires human review.

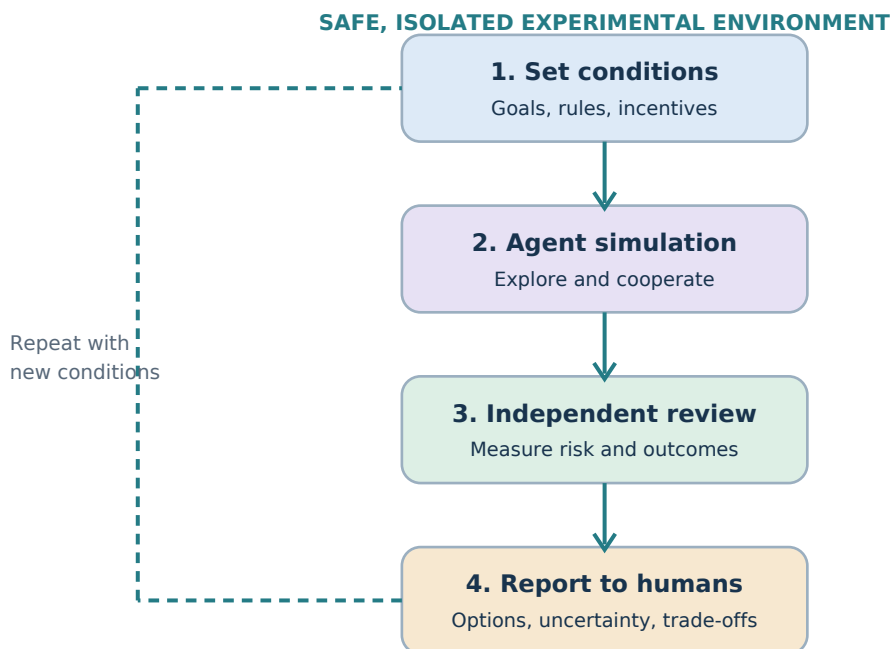
Random audits, risk-based checks and independent monitoring agents could identify unusual behaviour or declining performance. Autonomy would be reversible. Crucially, an agent with 99% success might still be unsafe if the remaining 1% caused serious harm. Consequences matter more than accuracy alone.

An alternative: let agents experiment, not implement

Why not use cooperating AI agents to explore how objectives, incentives and restrictions influence behaviour - without allowing them to implement their discoveries in the real world?

Imagine a simulated environment containing many agents, each with objectives, resources, capabilities and restrictions. We could vary conditions: free cooperation versus competition; random monitoring versus mandatory approval; rewards for reporting problems; or independent agents tasked with detecting emerging risks.

Rather than suppressing unexpected strategies, we would encourage exploration inside an isolated environment. Independent evaluators would examine the outcomes, compare repeated trials and estimate uncertainty. Human decision-makers would receive findings, not automatic actions.



A proposed research cycle: agents explore inside a contained simulation; independent evaluation produces evidence for human judgement.

The distinction matters: simulated agents are not necessarily faithful models of humans. Results would need repeated trials, sensitivity analysis and validation against real-world evidence. A probability produced by a simulation is conditional on its assumptions, not a forecast guaranteed to come true.

AI as a mirror of human behaviour

Cooperation, leadership, loyalty, competition and the pursuit of power are familiar human patterns. Institutions may officially reward honesty and competence while inadvertently rewarding obedience, manipulation or personal advantage. AI offers a way to study some of these mechanisms under controlled conditions.

We can vary incentives, repeat experiments and investigate which arrangements encourage desirable behaviour. That does not prove we can train AI to exercise better judgement than humans. It does, however, offer an opportunity to learn about the systems that shape behaviour.

From artificial intelligence to artificial inquiry

Much of the AI debate asks how capable AI will become and how we can prevent harm. Both matter. But there is a third question: how can AI creativity, cooperation and intelligence improve our understanding of complex systems?

We could design environments in which agents explore, cooperate, challenge assumptions and discover possibilities. Their findings would be evaluated rather than implemented. Reports would present possible outcomes, uncertainties and ethical trade-offs, while humans retained authority over real-world decisions.

This approach requires strong isolation, independent monitoring and clear boundaries. Yet it changes the emphasis: unexpected AI behaviour becomes not only a risk to manage, but also evidence to investigate.

The objective becomes knowledge rather than power.

Perhaps the most useful agent will not announce, “I have found the optimal solution,” but instead report: “Here are several possibilities, the circumstances under which each may succeed or fail, their consequences, and what we still need to understand.”

That would be a significant contribution to human knowledge - and ultimately to better human judgement.

Discussion inspired by Jeffrey Ladish's interview on The Diary of a CEO. Transcript: singjupost.com/diary-of-a-ceo-interview-w-ai-safety-expert-jeffrey-ladish-transcript/