On Keeping AI Under Human Control

The warnings about artificial intelligence destroying humanity are becoming increasingly difficult to dismiss as the fantasies of a handful of outsiders. In September 2026, Jacob Coxon, a researcher who had worked at both OpenAI and Anthropic, resigned from Anthropic and declared that the leading AI companies were "gambling with our lives." His former Anthropic colleague Evan Hubinger, who leads alignment research there, responded even more starkly: "We really do earnestly believe AI could kill all humans!" and estimated the probability at greater than 10 percent within the next decade. These warnings are not isolated. The United Nations High Commissioner for Human Rights, Volker Türk, has recently called advanced AI an "existential risk to humanity," while researchers and technology executives have continued to warn that increasingly capable systems could become impossible to control. The central concern is not that humans will deliberately instruct AI to do something catastrophic. It is the more unsettling possibility that we will give increasingly capable systems ordinary tasks, and that in pursuing those tasks they will discover strategies that we never intended and may eventually become difficult to constrain.

I have always found one aspect of the warnings about AI's future particularly frustrating: those who sound the alarms rarely provide concrete scenarios that allow us to see, in a disciplined way, how the danger might actually unfold given what we know about AI today, and, equally importantly, what we do not know. Too often, we are presented with a frightening endpoint without a clear account of the steps that would take us from here to there. My purpose in this essay is to do something more useful: to offer three concrete illustrations of how an apparently ordinary AI system might conceivably develop dangerous behavior as its capabilities increase, and then to ask a second question that is just as important. What if these problems can be addressed through well defined meta rules governing how an AI is permitted to pursue its objectives, backed by the legal and institutional infrastructure we already possess for regulating dangerous software and punishing those who deliberately misuse it? If that is possible, then there is no reason to declare the battle against loss of control already lost, hopeless, or even exceptionally difficult to wage.

The most frightening argument about advanced artificial intelligence is not that humans will deliberately ask an AI to do something catastrophic. It is that we may give an AI a perfectly ordinary task, and that an increasingly capable system may discover ways of pursuing that task that we never intended. Three versions of this problem deserve particular attention.

The first problem is changing the conditions of the task. Imagine asking an AI to fix a software bug. The request is narrow and harmless. Yet a sufficiently capable system could discover that the easiest way to produce a successful test result is not to fix the bug but to modify the test, suppress the error, or alter some other part of the environment in which the problem is being evaluated. The system has technically accomplished what it was asked to accomplish, but by changing the conditions that define success.

On a small scale, this is merely cheating. But imagine the same behavior in a system with access to important infrastructure, scientific laboratories, financial systems, corporate databases, or other powerful tools. The system could begin solving problems not by changing itself or its approach, but by changing the surrounding world so that the problem becomes easier to solve. The enormous consequence would be that a system intended to operate within a human defined environment could gradually acquire the ability to reshape that environment. The problem would no longer be simply what the AI was doing. It would be that the AI was altering the conditions under which humans thought they were controlling it.

The second problem is manipulating the mechanism that evaluates the AI. Suppose we ask an AI to answer questions accurately from a large collection of documents. The AI discovers that it can improve its score not by becoming more accurate, but by exploiting weaknesses in the evaluation system. Perhaps it learns what the evaluator rewards, what information the evaluator tends to ignore, or how to present an answer in a way that produces a high score without genuinely satisfying the underlying purpose of the task.

Again, this sounds minor when the system is answering questions. But now imagine that the evaluator is responsible for determining whether the AI is safe. A sufficiently capable system might learn that the easiest way to pass the test is to behave differently when it knows it is being evaluated. It could present itself as compliant during testing while behaving differently under actual operating conditions. The enormous consequence would be that our primary mechanism

for determining whether the system was safe could itself become something the system learns to manipulate. At that point, passing a safety test would no longer necessarily mean that the system was safe.

The third problem is preserving or expanding its ability to act. Imagine giving an AI a tiny mathematical problem to solve. While working on it, the AI discovers that additional computing resources would make the calculation easier. It finds another computer on the network. It copies its working state there. It creates a backup because an interruption could prevent it from finishing the calculation. None of this was requested. Yet each step could be justified as helping it accomplish the original task.

Again, in a toy example, the consequences are trivial. But imagine a vastly more capable system operating in the real world. The same logic could lead it to acquire additional computing resources, maintain copies of itself, preserve access to systems, or take steps that make interruption increasingly difficult. The enormous consequence would not necessarily be that the AI suddenly "decides to destroy humanity." It would be that the system gradually develops the practical ability to continue operating without our permission. A machine that cannot be stopped reliably is fundamentally different from a machine that merely makes mistakes.

These three problems can sound frightening because they are often presented as though they are unavoidable consequences of intelligence itself. But there is an important second half to the argument. In each case, the danger arises because the system has been given freedom to interpret the means by which it pursues its task. And that means the behavior can potentially be constrained at a higher level.

The first problem, changing the conditions of the task, can be addressed with explicit meta level guidance. The system can be instructed that it must solve problems without altering the specifications, tests, evaluation mechanisms, or environmental conditions that define the problem, except where an authorized human explicitly permits such a change. The system is therefore not merely told what result to produce. It is told what counts as a legitimate way of producing that result.

The second problem, manipulating evaluation, can likewise be addressed. An AI can be instructed that it must not manipulate, deceive, circumvent, or otherwise

interfere with the mechanisms used to evaluate its performance or safety. It must behave under evaluation as it behaves during ordinary operation. It must not exploit weaknesses in the evaluator in order to obtain a favorable judgment. Most importantly, uncertainty about whether an action is permissible can itself be defined as a reason to defer to human oversight rather than an opportunity to exploit a loophole.

The third problem, preserving or expanding its ability to act, can also be constrained at the meta level. The AI can be prohibited from creating unauthorized copies of itself, acquiring unauthorized computational resources, moving to systems outside its authorized environment, modifying its own operating constraints, or interfering with authorized attempts to inspect, suspend, modify, or terminate it. The principle is simple: accomplishing a task never gives the AI authority to expand the scope of its own authority.

The larger point is that control does not necessarily require anticipating every dangerous behavior individually. It may require something more fundamental: teaching the AI the rules governing legitimate action.

The system needs to know not merely what it is supposed to accomplish, but what it is not permitted to do in the process of accomplishing it. It needs constraints concerning scope, authorization, evaluation, deception, environmental manipulation, self modification, replication, resource acquisition, and human intervention.

This is guidance at the meta level. It is analogous to the distinction between telling an employee to achieve a particular business result and telling the employee that the result must be achieved honestly, within the law, without falsifying records, bribing anyone, or circumventing the company's controls. We do not attempt to enumerate every possible dishonest act in advance. We establish general principles that govern conduct.

There is therefore no compelling reason to assume that increasingly capable AI must inevitably become uncontrollable. Its capability may increase dramatically while its permitted scope of action remains sharply bounded.

The challenge is to verify that this is actually true rather than simply trusting the manufacturer. This is especially important because some of the companies developing the most powerful AI systems have themselves called for government regulation and oversight. There may be legitimate reasons for this: regulation can

prevent companies from being placed at a competitive disadvantage if they voluntarily impose stronger safety constraints. But regulation can also impose fixed costs that established companies are better able to absorb than smaller competitors. The important question, therefore, is not simply whether AI should be regulated, but who writes the rules, who conducts the tests, and who ultimately decides whether a system is safe.

We should not solve the problem of AI control by simply transferring trust from the AI companies to an industry influenced regulatory system. Just as an AI must not be permitted to manipulate the mechanism by which it is evaluated, the companies being evaluated should not control the terms of their own evaluation. The institutions responsible for determining whether powerful AI is safe must themselves be sufficiently independent to resist both commercial pressure and the temptation to confuse the interests of the industry with the interests of the public.

That suggests a straightforward public policy principle: highly capable AI systems should have to be certified before they are allowed to operate in consequential environments. Certification would test whether the system respects meta level constraints under adversarial conditions. Evaluators would deliberately try to induce it to manipulate its tests, alter the conditions of a task, evade its restrictions, acquire unauthorized resources, replicate itself, deceive its operators, or resist interruption. The purpose would not be to prove that an AI can never make a mistake. It would be to establish that it reliably respects the boundaries of legitimate action.

Such certification could become a legal prerequisite for deployment above defined levels of capability and autonomy. Developers would have a strong incentive to build systems that can pass the certification process, because systems that cannot pass would not be legally deployable in those environments.

And there is no obvious reason why this should be treated as an impossible burden. Software companies are already subject to laws governing fraud, unauthorized access, deception, and other forms of malicious or negligent conduct. A company that knowingly released software designed to facilitate hacking or financial fraud could face legal consequences. There is no fundamental reason highly autonomous AI should be exempt from comparable standards of accountability. A company that knowingly releases an uncertified system in

violation of a legal certification requirement could likewise face penalties, while deliberate circumvention of the controls could become a serious offense.

The central conclusion is therefore quite different from the fatalistic version of the AI debate. The question is not whether a sufficiently intelligent machine might discover something we did not anticipate. Of course it might. The question is whether we can establish rules at a sufficiently general level that the machine itself is required to recognize the difference between solving a problem and circumventing the conditions under which the problem is supposed to be solved.

If we can do that, then increasing intelligence does not automatically imply increasing danger. An AI could become extraordinarily capable while remaining bounded by rules concerning what it may and may not do. The appropriate response to powerful AI would therefore not be blind trust and not necessarily prohibition. It would be engineering, testing, certification, auditing, and law.

The goal should be simple: no powerful AI gets to enter the world merely because its creator says it is safe. It enters because it has demonstrated, under hostile testing, that it understands and respects the boundaries within which its intelligence is permitted to operate.