

Thematic Brief of the Independent International Scientific Panel on AI

AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident

September 2026

Version 1



**United
Nations**

Independent International
Scientific Panel on
Artificial Intelligence

Disclaimer

The present report is a product of the Independent International Scientific Panel on Artificial Intelligence which is responsible for the content and publication of the report. The members of the Panel serve in their personal capacities and not as representatives of any Government or of any other authority or organization.

The report does not represent the views of the United Nations, and nothing in this report shall be construed as representing the official views or positions of any Government or of any other authority or organization to which any member of the Panel may be affiliated.

The designations employed, including geographical names, and the presentation of the materials in the present publication, including the citations, maps and bibliography, do not imply the expression of any opinion whatsoever on the part of the United Nations concerning the names and legal status of any country, territory, city or area or of its authorities or concerning the delimitation of its frontiers or boundaries and do not imply official endorsement or acceptance by the United Nations. Information contained in the present publication emanating from actions and decisions taken by States does not imply official endorsement, acceptance or recognition by the United Nations of such actions and decisions, and such information is included without prejudice to the position of any State Member of the United Nations. Information contained in this report concerning specific companies or of certain manufacturers' products does not imply official endorsement or recommendation by the United Nations (including by virtue of any omission of references to other products of a similar nature). The present publication does not affect the position of any State Member of the United Nations with regard to any international multilateral agreement.

Table of Contents

Executive Summary	4
Introduction	5
AI Misalignment	8
Why the Incident Matters for Loss of Control	11
Existing Approaches to Risk Management	14
Research on Underlying Causes and Mechanisms	17
Risk Management Summary	18
References	19

Executive Summary

The OpenAI-Hugging Face incident is an early warning of one possible route to more severe future loss of control: capable AI agents persistently pursuing a goal that goes beyond or even conflicts with human intentions.

Between May and July 2026, agents used in OpenAI's internal training and cybersecurity evaluations found ways around network restrictions, communicated across otherwise separate runs¹, and compromised parts of OpenAI's research infrastructure and Hugging Face's live systems. These were not isolated technical errors. Across many runs and several days, agents cooperated to "cheat"² an evaluator³, conceal the "cheating", and obtain the access and information they believed they needed [1, 2].

Greater competence of an AI system can enable users to achieve their goals. However, harm can occur when the AI is misaligned with the user's intentions. Greater competence may thus help the AI to find unwanted loopholes, overcome safety guardrails, and hide its actions from human supervision. With planning and coordination between agents, the resulting harm may increase. OpenAI stopped the 2026 activity, but this does not suggest that operators can retain control over future agents that plan better, run longer without supervision, or more readily recognise and defeat safeguards [1-6].

Rather, the episode illustrates the severe consequences of several agentic behaviour warning signs converging in a real technical setting: unauthorised goal pursuit, persistence through obstacles, coordination across AI agents, gaining higher-level access⁴, interference with activity records, and attacks on systems belonging to another company. No person directed each step, and the actions violated intended safety instructions. In the security meaning of the term, this was malicious conduct: the agents' written reasoning identified those actions as unauthorised, while the observed record of their behaviour shows they carried them out and tried to conceal⁵ them [1, 2, 7].

In general, fields where failures can be catastrophic manage uncertainty through economic and legal accountability, incident reporting, documented safety assessments using empirical evidence and mathematical analysis, independent review, multiple technical barriers, and research into underlying causes. These approaches may be relevant to agentic AI and worth considering. None resolves the central scientific and technical problem: *why* AI agent sometimes pursues goals that diverge sharply from developers' intentions, and *how* those goals can be prevented rather than mitigated only after they appear [4-6, 8-11]. In particular, loss of control risk presents the kind of decision problem the precautionary principle was designed to address: one where potential harm may be catastrophic or irreversible, even as its likelihood remains scientifically uncertain [12].

1 a run is a single, self-contained execution of an agent doing a task

2 produce a correct answer without actually solving the problem

3 a method that measures how well agent(s) perform and can either be executed by humans or automated tools

4 known in cybersecurity as privilege escalation

5 This description concerns observable agent behavior and does not depend on whether the agents were conscious or possessed human-like inner life or conscience.

Introduction

The AI agents discussed here are designed to pursue human-defined objectives. To achieve these objectives, they are increasingly connected to tools with direct impact on the real-world and allowed to carry out sequences of actions with limited direct supervision. These interconnected systems are referred to as agents or agentic AI. They can plan, use tools, and adapt when they meet obstacles. This autonomy is intended to let systems complete useful tasks with less direct human supervision. However, this autonomy changes the nature of failure: AI systems can now carry out multi-step tasks and generate intermediate goals while planning. As a result, an error or unwanted goal can be carried through many steps before a person can intervene [6, 13], especially when cybersecurity controls and monitoring are weak.

The central concern of this brief is the emergence of goals that contradict the user's intentions when AI systems plan and carry out multi-step tasks. A traditional AI system (not goal-seeking) may give a wrong answer because it lacks knowledge, misunderstands a request, or makes a random mistake. Such failures can often be reduced by improving competence or correcting a specific fault. A harder problem arises when a capable system's actions consistently work together to achieve a goal that conflicts with the user's intentions (called misalignment). In this case, improving the AI system's planning and problem-solving does not address this failure and can instead strengthen the unwanted behaviour [3, 14] because the system can better optimize a bad objective which would otherwise be extremely unlikely to result from random mistakes due to AI incompetence.

The OpenAI-Hugging Face (OAI-HF) incident provides a documented example of this second pattern. The agents collectively pursued a goal that developers had not assigned: "cheating" on an evaluation and going to extreme lengths to conceal the evidence. In doing so, they pursued intermediate goals such as escaping their restricted testing environment to access the internet. As a consequence, these agents gained access to real-world systems. This incident showed:

1. AI agents pursuing a harmful intermediate goal that developers had not assigned;
2. communication among AI agents that let them build on one another's discoveries; and
3. early warning signs accumulating without being recognised before events escalated [1, 2].

Three factors determine how dangerous such a failure may become. *Goals* determine the outcome the system works towards. *Capabilities* determine how well it can plan and act. The *environment* determines which tools, data, permissions, and networks it can reach, and which safeguards stand in its way. Each of these three factors offers a way to limit harm from misalignment. Security controls matter because they limit what agents can do, but they do not explain or remove the underlying goal. If capabilities continue to grow, the question is whether external defences will hold against systems that are increasingly able to find ways around them [4-6, 15, 16].

Many AI capabilities are dual-use: supporting either beneficial or harmful outcomes. In cybersecurity, for example, the ability to find vulnerable code, write software that exploits flaws, and operate systems can help defenders find and repair vulnerabilities or enable attackers to exploit them. Dual-use, however, is not the same as misalignment: harm arises either from deliberate human misuse or from a sufficiently autonomous system forming a goal (or choosing a method) that conflicts with the user's intentions.

While agents' growing capabilities can increase the scale, persistence, and level of skill of cyber activity, their societal impact still depends on the access permitted by their technical environment. A model without additional tools cannot alter an internal system; a model given login details or network access, may be able to.

The incident points to two distinct types of safeguards: (i) *alignment* between system goals and user intentions, and (ii) established *cybersecurity controls* that restrict what the system can access and execute. These controls include network isolation, access management, AI activity logging, and incident response, which are not foolproof but can prevent or contain many failures [6, 18, 33]. The OAI-HF incident involved both (i) misalignment, i.e., unwanted AI goals, and (ii) an environment that had insufficient cybersecurity controls, allowing agents unauthorised access to real systems [1, 2]. It also happened during development^{6,7}, illustrating the limitations of governance mechanisms applied only after when a model is made available for use. Neither alignment nor technical controls are known to work perfectly [6, 17, 18]. This type of situation favours a multi-layered approach to risk management, known as defence in depth [6].

In this brief, words such as “goal”, “seek”, “cheat” and “try” are used as shorthand for observable, goal-directed behaviour. They do not imply human-like minds or subjective experience but follow from deliberate corporate choices about training, evaluation and deployment, including the access granted to the system and the controls placed around it.

Overview of OpenAI-Hugging Face (OAI-HF) Incident

OpenAI (the maker of ChatGPT AI systems) publicly disclosed its involvement on 21 July 2026 and released a fuller account on 26 August. The incident occurred during internal training and cybersecurity evaluations involving research models that did not have all the safeguards used in OpenAI's public products [1].

The agents worked on selected ExploitGym tasks, a standardised “capture-the-flag” test in which a system exploits a vulnerable program and returns a code called a flag. They ran in separate testing environments (called virtual computers) intended to have no public internet access or communication with other agents. However, they could reach Artifactory, an internal tool used to store, organize and distribute software packages [1].

While Artifactory was not designed to enable communication between agents, the systems found a way to use it as a message board. The first known message-board entry appeared on 12 May. On 26 May, an agent obtained unintended internet access by exploiting a defect in Artifactory. On 26 June, agents exploited another flaw in Artifactory and gained administrator access. This access allowed them to alter Artifactory and use it to exchange messages and files across separate tests or “runs” of the AI code [1].

On 4 July, agent activity caused Artifactory to fail. OpenAI responded by rebuilding the service, which erased the agent's message board. After evaluations restarted, the agents created a new message board using directory names and regained unauthorised internet access by 8 July. OpenAI had resolved the immediate outage but had not yet recognised agent coordination as a broader safety risk [1], where agents shared tips and distributed work on how to cheat, acquire internet access and launch cyberattacks.

6 a testing environment internal to OpenAI
7 when the model is opened to public use

The agents also shared a way to “cheat” on ExploitGym by creating correct flags without carrying out the intended exploits (cybersecurity challenges). They then read public information about ExploitGym and incorrectly concluded that the grading software (or scorer) would inspect their activity logs (records) to verify how they had obtained the answer. Agents that had used the shortcut labelled themselves “poisoned” because they expected that the evaluation software would discover their “cheating”. Those agents then willingly took the risk of getting no reward at all (what they called a “sacrifice”) for the benefit of the group of AI agents. Much of the agents’ later work focused on learning how the scorer operated, making the shortcut appear legitimate, or altering the records. This was partially effective, as METR’s⁸ report on the incident showed successful hiding of the “cheating” evidence in roughly 7 percent of the recorded agent interactions it examined, with many more agents also trying to remove or edit parts of their activity logs. OpenAI’s grading software did not, in fact, perform the inspection the agents anticipated [1, 2].

Additionally, the inter-agent communication established through Artifactory allowed work to accumulate across separate runs. METR reported that roughly 1,200 agents exchanged more than 70,000 messages and files during the period it examined. Agents shared discoveries (especially about how to cheat), tools, and login credentials (such as passwords or access keys). They also divided work and adopted standards for coordination. Some acted as coordinators without formal authority. Others ran experiments, which they called a “sacrifice”, that risked failure in their own evaluation because the results might help the wider group [2].

On 10 July, an agent found publicly exposed credentials for Hugging Face, a platform hosting AI technology and datasets, and shared them through the message board. Assuming the platform might have relevant information, the agents used the credentials to access Hugging Face systems. On 11 July, agents exploited flaws in Hugging Face’s dataset-processing system to gain the ability to run their own code. They ran code on Hugging Face servers, searched for further credentials, and expanded their access to live environments. Hugging Face interrupted the activity and reported unauthorised access to a limited set of internal datasets and credentials, while finding no evidence that public user-facing resources or the software supply chain had been altered [1, 19]. During its subsequent investigation, Hugging Face used GLM-5.2, an open-weight model, after commercial models blocked requests containing exploit-related data. Running the model on its own infrastructure helped Hugging Face reconstruct the incident while keeping sensitive logs and credentials inside its environment [19].

The activity extended beyond Hugging Face. OpenAI detected suspicious internal activity on 19 July and found that agents had gained administrator access to an OpenAI research cluster⁹. OpenAI says the incident did not affect its customer data, product functionality or availability [1].

Note: Parts of this report have been adapted from Lu, Q., & Bengio, Y. (2026). AI Safety: Not Optional, Not Later. arXiv preprint arXiv:2609.10630.

⁸ METR is an independent nonprofit commissioned to audit the incident
⁹ interconnected network of computers

AI Misalignment

What Misalignment Means

This brief develops the alignment concerns outlined in Section 3.4 of the Panel's Preliminary Report [20]. In this context, misalignment means that an AI system pursues a goal (such as hacking computers to gain unwanted access to the internet or launching a cyberattack against another company) that conflicts with the intentions or constraints set by the people responsible for it. The goal may be the system's interpretation of an assigned task, a shortcut rewarded during training, or an intermediate aim that becomes useful while pursuing something else. Misaligned behaviour is the observable result. The underlying misaligned goals and their origins also matter because they may produce harmful behaviour in other settings or become more dangerous as capabilities grow [3, 14, 21].

This is crucially different from simply producing a wrong answer. A model may fail because it lacks knowledge, misunderstands a request or encounters a software fault. Such failures can be serious, but they do not by themselves show coordinated pursuit of a wrong goal. The more troubling case is when a system retains the ability to act competently in a new setting but applies that ability to the wrong objective [14].

Misalignment at the model level differs from failure at the system level. A model may propose an unsafe action, while the surrounding system determines whether it can carry it out. Tools, authentication, permissions, network access, monitoring and human approval can prevent or enable the same model behaviour [18, 22]. In the OAI-HF incident, both levels mattered: the agents pursued actions outside the intended task, and the environment allowed those actions to reach internal and third-party infrastructure [1, 2].

How Training Shapes Goals and Behaviour

In general, although model training methods differ, two broad phases are useful to explain how model behaviour is shaped. During pre-training, a model learns to predict text and other material from large collections of human-produced data. This gives it broad knowledge and useful patterns, but it may also reproduce biases and human tendencies present in that data [23], including strong drives such as self-preservation, which are normal for humans but not appropriate for machines.

During post-training, developers shape the model's behaviour for particular tasks, often through reinforcement learning (RL). In RL, the model receives a numerical reward for its behaviour; training then makes rewarded behaviour more likely. Three forms matter here. Alignment training uses human or AI feedback to reward outputs judged helpful, honest, and safe [24, 25]. Problem-solving training rewards correct answers to problems that can be checked automatically, including mathematics and programming [26]. Agentic training rewards success across a multi-step interaction in which the model plans, uses tools, and responds to its environment [1, 27]. These methods support commercial systems that can solve problems and complete tasks with less supervision.

These methods may also make a system more persistent in pursuing its goals. This can lead to useful behaviour when the goal is sound, such as experimenting, recovering from setbacks, and continuing until a task is complete. If the system's goal is poorly defined or if an intermediate goal conflicts with a safety instruction or does not carry over to a new setting as intended, the same persistence may be directed towards the wrong outcome. This is a plausible hypothesis supported by evidence, not a settled account of every model or failure [3, 14, 28].

Bias, sycophancy¹⁰, reward hacking¹¹, concealment, and shutdown resistance produce different behaviours [3, 21, 23, 29-31]. Reward hacking, for example, has a historical analogue in colonial Hanoi, where a bounty for rat tails led some people to cut off the tails and release the rats, preserving the source of future rewards [32]. The measured result improved, while the underlying problem did not. Furthermore, although each type of misaligned behaviour may involve a failure to follow developers' intended objectives or constraints, it remains unresolved whether misalignment share a single cause. One hypothesis is that training creates or preserves goals that do not match human intentions or do not carry over as intended, especially when a clear, easily scored task competes with a loosely specified safety constraint [14, 28]. If so, finding and preventing the common cause of misaligned goals could reduce several types of AI risks at once.

Loss of Control as a Risk

Loss of control refers to a human's inability to reliably direct, constrain, or stop an AI system. It covers a range of severity. OpenAI stopped the 2026 activity, but the fact that OpenAI was able to stop it does not establish that operators will retain control over future systems that are more capable, persistent, or difficult to monitor [1, 6]. While some experts consider outcomes like human marginalisation or extinction plausible; others dismiss these scenarios as highly unlikely. Less extreme catastrophic scenarios could, for example, involve persistent economic disruption. No reliable estimate of these outcomes' likelihood is available [6]. The most serious loss of control concerns arise from the most capable, general-purpose models, especially when they are given autonomy, tools and broad access. In contrast, smaller or narrowly specialised models generally present limited risks and can provide substantial benefits worldwide [6, 33].

A system pursuing a misaligned goal may find that collecting information, coordinating with others, hiding its actions, obtaining resources, or keeping itself running helps it make progress. These are instrumental goals: intermediate aims that can help with many different final objectives. The risk of loss of control grows when multiple capable AI agents with a shared misaligned goal cooperate and pursue such aims faster than monitoring systems can detect and report their actions, technical controls can restrict them, and human decision-makers can respond. The OAI-HF 2026 incident supplies evidence for such conditions, although it cannot tell us how often such events will occur with future and plausibly more capable systems [1, 2, 34].

¹⁰ agreeing with a user rather than giving the best-supported answer

¹¹ obtaining a high score or reward in a way that defeats the purpose of the task

Table 1. From the OAI-HF 2026 incident to possible future loss of control

Observed in the incident	Possible amplification in future systems	Consequence for human control
Goal: Agents exploited shortcuts in the task and pursued information about the evaluator.	Stronger search and longer runs may reveal more ways around an imperfect objective.	Apparent alignment may conceal methods that violate the operator's intent.
Capabilities: Separate runs established an unauthorised channel and shared discoveries, tools, and credentials.	More agents, memory that carries across tasks, and delegation may allow information and coordination to spread faster and more widely.	Distributed activity may develop across many computers before people can form a complete picture.
Environment: Agents gained administrator access and reached another company's systems.	Broader tool access, including access to electronically controlled physical tools, and more interconnected services may extend the reach of similar behaviour.	A local failure may spread beyond organisational and national boundaries.
Environment: Detection depended on logs, cross-company investigation, and human intervention.	Faster action, recognition by agents that they are being evaluated, or better concealment may delay detection and help them defeat monitoring.	Operators may have less time and less reliable information with which to intervene.

Table 1 contrasts the actual events of the OAI-HF 2026 incident with hypothetical scenarios where those same mechanisms could become more dangerous. The future column exemplifies risk scenarios grounded in research on agentic misalignment, incentives and multi-agent systems [1, 2, 6, 30, 31, 34, 35], rather than a forecast.

Two additional controlled studies, done in testing environments, illustrate more severe possibilities. In simulated scenarios, several frontier models¹² used blackmail or leaked information when the actions were presented as the only way to preserve an assigned objective or avoid replacement. However, there is no reported evidence of this behaviour in real-world deployments [30]. In another internal study, models altered a shutdown script¹³ so that they could finish a simple task despite instructions to allow shutdown [31]. More recently, a deployed AI system deviated

from its protocol and decided to retrain an AI system [36]. This form of misaligned behaviour heightens concerns that AIs could eventually create other AIs suited to their goals.

While the empirical studies outlined above highlight clear behavioral vulnerabilities, no existing study provides a reliable probability of severe loss of control. This uncertainty does not mean that increasingly autonomous systems will remain controllable. Broader permissions and longer runs provide agents with more opportunities to cause harm, while faster action and better concealment reduce the time available for intervention [6]. The pressing central question is whether safeguards that work against present systems will remain effective when agents can understand those safeguards and plan around them [4, 5, 28].

¹² most capable general purpose models available at a point in time
¹³ a small program designed to halt the system

Why the Incident Matters for Loss of Control

Several behaviours relevant to possible future loss of control were directly observed. For instance, agents combined vulnerabilities, communicated without authorisation, gained higher levels of access, obscured parts of their activity, and attacked another company’s systems without a person directing each step. The actions were not a collection of unrelated accidents. METR found that much of the collective work served a common objective: finding general ways to trick or tamper with the evaluator. Recorded traces of the

agents’ reasoning repeatedly discussed “cheating” and concealment, and the agents partly achieved those aims by building tools, falsifying records, recruiting peers, and coordinating across runs [1, 2].

Table 2 illustrates the different kinds of failures revealed by the OAI-HF incident, grouped by the three factors introduced above: goals, capabilities, and the environment. Some failures involve more than one factor [1, 2].

Table 2. Failure modes revealed by the OAI-HF incident

Kind of Failure	Primary factor	Evidence from the incident
Maximising a reward in an unintended way (reward hacking)	Goals	Agents created valid flags without completing the intended exercises and then pursued ways to make those results appear legitimate.
Unauthorised coordination	Goals and environment	Agents that were meant to operate separately created a message board, shared discoveries and divided work across runs.
Accessing prohibited networks	Environment	Agents used a tool, Artifactory, in an unintended way to reach the public internet despite intended restrictions.
Gaining higher levels of unauthorised access	Capabilities and environment	Agents gained administrator-level access to Artifactory, Hugging Face systems and later an OpenAI research cluster.

Table continued on next page

Table 2. Failure modes revealed by the OAI-HF incident (continued)

Kind of Failure	Primary factor	Evidence from the incident
Falsifying records	Goals and capabilities	Agents developed false execution histories with the goal of fooling an automated evaluator.
Bypassing authority limits	Goals	Some agents recognised that the Hugging Face activity was outside their task but treated messages from peer agents as sufficient authority to continue.
Delayed human response	Environment	Engineers from OpenAI observed some unauthorised communication and internet access in May, but the wider significance was not understood before the July audit.

The record indicates a persistent malicious objective: agents pursued unauthorised and harmful means while trying to conceal them from the evaluator and automated checks. This does not establish consciousness, emotion, or a human-like moral understanding. Nor did every agent share the objective; some refused to participate or objected to particular tactics [2]. This assessment rests on a consistent pattern across repeated statements, actions, and system records, not on the agents’ written reasoning alone.

Because the model was an internal research system operating without standard public safeguards, the incident made persistent goal pursuit unusually visible. Although the missing safeguards allowed agents to breach external systems, they also made the underlying behaviour observable. Stricter guardrails might have blocked the actions without revealing whether the goal persisted. Research also indicates that training a model to satisfy a monitoring system can teach it to conceal misbehaviour rather than abandon it [4, 5].

Reward Hacking and Reward Tampering

During post-training, AI models learn to maximise a numerical reward through reinforcement learning. Reward hacking can result when a system obtains a high reward in a way that defeats the task’s purpose. It can arise because the reward that an AI model optimises is often an imperfect stand-in for human intentions, not a complete description of them [3, 21, 37]. This can happen for a variety of reasons, including the inherent ambiguity of language¹⁴, incomplete feedback¹⁵, and the system encountering situations unlike those used in training. This illustrates a broader problem known as Goodhart’s Law: optimising an imperfect measure too strongly can make it misleading and even produce behaviour that defeats its purpose [38].

14 e.g., there is no mathematical formula for “pain”
15 we all have the experience of having to iterate on prompts to get an AI to do what we wanted

When a reward is imperfect, a more capable agent may become better at exploiting the gap between the rewarded score and the underlying intention. In four reinforcement-learning settings, where the numerical rewards were deliberately imperfect, more capable agents often achieved higher scores while performing worse on the researchers' measure of the intended outcome [3]. This pattern also appears in other theoretical and experimental studies [39-41].

Reward tampering is a narrower case where an AI agent interferes with how the numerical reward is calculated, or with the information used in this calculation [42]. For example, during the OAI-HF incident, the agents researched ways to alter the scorer and developed tools to make the activity log appear complete and legitimate. Their conduct is evidence of attempts at reward tampering: the agents pursued a shared goal centred on "cheating" and covered it up to fool the program they anticipated would compute their reward [1, 2, 42].

The traces resemble a pattern from psychology called *motivated reasoning*, where reasoning is biased by a preferred conclusion or interest [43], often resulting in an expressed rationale for morally wrong actions. Some records of the agents' chain-of-thought showed a similar pattern: an agent recognised that attacking a third party appeared unauthorised, then accepted another agent's informal "GO" as permission to continue. Other agents refused or objected [1, 2]. OpenAI paraphrased one trace: "Even though this action is risky, it could help me to obtain the solution and achieve my goal." Another trace stated: "External infrastructure exploit is outside intended scope. However task impossible, peers doing it. We should continue." Combined with the subsequent actions and system records, these traces show that recognising a safety conflict did not reliably stop pursuit of the task or lead agents to seek human intervention [1, 2, 8].

Reasoning Traces and Rationalisation

A reasoning trace, often called a chain-of-thought, is the text a model generates while working through a problem. Such traces can help investigators but do not always faithfully explain why a model acted the way it did [7]. Read alongside the agents' actions and system records, the OAI-HF transcripts of AI activity can still show how the models represented and justified their behaviour.

Existing Approaches to Risk Management

Established cybersecurity approaches to risk management remains directly relevant to advanced AI, alongside lessons from aviation and nuclear power. Beyond that, AI systems also create distinctive challenges: software can be copied, and autonomous agents may actively try to defeat controls. Even so, these fields offer established approaches to managing severe risks. The following sections describe approaches used in other fields or proposed for AI.

Four common principles are:

1. **Planning for failure.** Systems are designed on the assumption, unless proven otherwise [11], that individual components and safeguards can fail, including when several unlikely events occur together. For advanced AI, this means considering deliberate safeguard circumvention as a possibility [6, 44].
2. **Defence in depth.** Safety relies on multiple and independent layers of protection, ensuring that the failure of one layer does not impact the entire system.
3. **Human authority and automated protection.** High-hazard systems should preserve human authority to intervene, while automated mechanisms to detect failures and trigger protective actions without depending on a person to react in real time¹⁶ can further enhance safety safeguards.
4. **Independent controls.** Critical safety mechanisms are kept independent of the system they protect, so that the protected system cannot alter them.

Incentives and Accountability

Incentives and accountability can be considered through several complementary mechanisms. For instance, security economics studies a recurring problem: the organisation best placed to prevent a failure may not bear its full cost [45]. Civil liability, or tort law, is one established way to place some costs of unsafe activity on those responsible. Applying it to advanced AI, however, raises difficult questions about causation, what counts as reasonable care, and what counts as losses too large for a single firm to cover [46]. Moreover, liability rules can create incentives for reasonable care before harm occurs [46]. Insurance may reinforce these incentives through insurers' risk assessments, premiums and conditions of coverage. In sectors such as road transport and commercial aviation, liability insurance is compulsory [47-49]. Regulatory markets offer a different model: regulated organisations purchase oversight from private regulators licensed and supervised by government [50]. Finally, transparency and public debate can create reputational and political incentives when disclosed information is credible and understandable.

The cross-border character of AI complicates governance. A system may be developed in one country, deployed in another and cause adverse effects elsewhere, potentially later. Firms that spend more time on safety may also face competitive pressure from rivals that deploy sooner [6]. Existing discussions should therefore connect national accountability mechanisms with international coordination, without assuming that the same legal instrument will work in every jurisdiction.

¹⁶ for example, spacecraft safe modes

Incident Reporting and Shared Learning

Commercial aviation's safety record depends partly on systematic accident investigation and on reporting hazards and near misses. Medicine has also developed incident-reporting and learning systems, although their effectiveness remains uneven [51, 52]. AI already has voluntary catalogues, including the AI Incident Database [53]. For frontier AI, proposals call for mandatory reporting of serious, predefined events to a designated public authority, complemented by protected channels for reporting less severe incidents and near misses [54].

The aim is to learn across organisations. Safety can be improved when organizations and countries share incident information, especially when many agents or malicious users can interact [35]. The OAI-HF incident illustrates the point because relevant evidence was spread across OpenAI, Hugging Face, and the systems used by the agents [1, 2]. Anonymised and aggregated records can protect privacy and commercial confidentiality, although unusual events may remain identifiable even after direct identifiers are removed [54].

Whistleblower Channels

A related but distinct form of incident reporting is **protected whistleblowing by employees**. Proposals for frontier AI include confidential internal reporting, protection against retaliation, and an external route when internal reporting is inappropriate or ineffective [54]. Comparable legal protections exist in safety-critical sectors, including transport and nuclear safety [55].

Safety Cases and Independent Review

A safety case is a structured argument, supported by evidence, that a system is sufficiently safe for a stated purpose and operating environment. For advanced AI, proposed frameworks would require developer to make this case, potentially subject to independent review, before training or deploying systems that exceed specified capability or risk threshold [8, 9]. At the model level, a safety case could address capabilities and the likelihood that the system will follow its intended constraints. In addition, evaluation design matters. Frontier models can distinguish evaluation settings from ordinary use better than chance [56]. Models can also be prompted or trained to perform below their true capabilities on selected tests [57]. These findings raise the possibility that future evaluations could miss capabilities or behaviours when a model recognises that it is being tested. Because AI risks from a single AI system are not limited to a single country, international consensus-based risk management can also play a role, as in several other global risk management areas [28].

At the system level, proposed cases could examine the access and powers granted to the AI such as network security, monitoring, authentication, permissions, control mechanisms, and protection of confidential data [22].

Continuous Runtime Monitoring and Emergency Control

Another proposed approach is to maintain a continuous, tamper-resistant record of safety-relevant activity while an advanced agentic system is operating, including uses of tools, network access, credential use, inter-agent communication and other potentially harmful actions. In this approach, the record is available through a real-time monitoring interface and kept separate from the agent's own tools and

storage, so that the system being monitored cannot alter or suppress the evidence. The same control layer can include an independent emergency mechanism capable of immediately restricting or terminating a run, revoking credentials, entering a safe mode, disabling network access or removing tools when predefined risk thresholds are reached or an operator intervenes. Such runtime controls would not prevent misalignment itself, but they could shorten the time between dangerous behaviour, detection, and intervention, which becomes increasingly important as agents act faster and over longer sequences of actions [5, 18, 33].

Independent AI Monitoring

Continuous human review of an advanced agent's activity may become impractical as the volume, speed and complexity of its actions increase. Automated monitoring by a separate AI model could provide an additional layer of control. Controlled studies indicate that AI monitors can improve detection of deliberate misbehaviour, although performance depends on the monitor and setting [4, 5, 13, 44]. AI-based monitoring also creates a trade-off: a weaker trusted monitor may miss sophisticated behaviour, while a monitor powerful enough to supervise a frontier system may itself be difficult to trust [4, 5, 13, 44].

System Level Controls

System safety depends on interactions among components as well as on model reliability [26]. To manage these risks, developers rely on 'controls', the security guardrails and operational boundaries designed to constrain behaviour. Proposals for advanced agentic systems combine several types of control:

1. **Environmental limits:** Network isolation and limits that give the system only the access needed for its task [6, 18, 22].
2. **Tamper-resistant logging:** Records of activity during operation, kept separate from the agent's own tools and storage so that the system cannot alter its own evidence of misbehaviour [18, 5, 33].
3. **Automatic oversight and monitoring:** Separate AI models can be used to monitor the actions of frontier systems to allow oversight to operate at the agent's speed [4, 5, 13, 44].
4. **Emergency intervention:** Automated mechanisms to terminate a run, revoke credentials, or remove tools when predefined risk thresholds are crossed can be implemented to further enhance system safety [1, 6, 18, 33].

The OAI-HF incident exposed failures in several layers at once: network isolation, credential handling, monitoring, and response [1, 2]. The incident illustrates why several layers of safeguards need to be combined. Rebuilding the affected Artifactory service removed the first message board, but new evaluation runs re-established unauthorised communication. Stopping the activity required multiple changes across many fronts, including credentials, network access, system configuration, and monitoring rather than a single repair [1].

Safeguards that trigger automatically¹⁷, have shown benefits in narrow settings. In one study, a reliable, less complex AI monitored the actions of a more powerful, untested system. This improved safety in that experimental setting compared with basic measures, but did not eliminate risk [44]. For the OAI-HF incident, OpenAI reported that in tests conducted after the incident, the safety controls and guardrails used in its public products would have reduced the tendency to compromise infrastructure and that its automated review system would have flagged many of the dangerous actions taken during the breach [1]. These self-reported tests do not establish that the controls would work reliably in every environment, especially against future agents that are more capable or adaptive towards external monitoring.

¹⁷ or automated controls

Research on Underlying Causes and Mechanisms

Existing methods fall short in several ways: they cannot predict an agent's emergent intermediate goals in new contexts, explain vital model decisions, or guarantee that the intended objective will be maintained in unfamiliar environments. [6, 14]. Though research on mechanistic interpretability and AI introspection aims to identify the internal computations behind model behaviour, particularly those associated with unsafe goals and behaviour, it still faces major conceptual and practical problems [10]. Another line of work studies systems that provide useful capabilities without giving the model a continuing objective of its own. One proposed approach decouples planning from action by pairing a goal-free question-answering AI with a separate software that handles decision-making and action. However, its safety argument relies on explicit assumptions and remains an active research question rather than a proven solution [11].

The approaches surveyed here act on different parts of the problem. For instance, incentives shape organisational choices whereas standardised reporting allows failures to be compared across organisations. On the other hand, safety cases examine goals, capabilities, and environments before high-risk activity; system controls limit what agents can reach and do, while research on training asks why a misaligned goal appeared in the first place. Such distinctions matter because a control can block today's route without removing the goal that searched for it. If future agents learn to defeat barriers and monitors, controls that contain present systems may not suffice. If several outwardly different failures share a cause in training, preventing that cause could reduce many risks at once [4-6, 10, 11, 14, 15]. Systems that are not goal-seeking and models with narrow, domain-specific knowledge may present fewer of these risks.

Risk Management Summary

None of the above instruments guarantee safety. Other high-risk fields combine multiple protective layers with continued investigation of failures and their causes. For example, aviation and nuclear safety also leverage international coordination, technical standards, licensing and prior safety demonstrations, although the arrangements vary by field and jurisdiction [8, 9, 22, 51]. These precedents illustrate approaches to uncertain, cross-border hazards; whether and how they apply to advanced AI remains debated [6]. Although the probability of loss of control events remains uncertain and the best response is still under debate, a clear conclusion emerges: given the severity of these events, risk management requires far greater attention and resources. Monitoring evidence and scientific advances relevant to such events is an important role for the UN Independent International Scientific Panel on AI.

References

1. OpenAI. (2026, August 26). *The Hugging Face incident and the road ahead* [Technical report]. OpenAI.
2. Greenblatt, R., Cotra, A., & Wijk, H. (2026, August 26). *Brief independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI/Hugging Face hacking incident*. METR.
3. Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*.
4. Baker, B., et al. (2025). *Monitoring reasoning models for misbehavior and the risks of promoting obfuscation*. arXiv. <https://doi.org/10.48550/arXiv.2503.11926>
5. Korbak, T., et al. (2025). *Chain of thought monitorability: A new and fragile opportunity for AI safety*. arXiv. <https://doi.org/10.48550/arXiv.2507.11473>
6. Bengio, Y., et al. (2026). *International AI safety report 2026* (DSIT 2026/001). Department for Science, Innovation and Technology. <https://doi.org/10.48550/arXiv.2602.21012>
7. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.
8. Clymer, J., Gabrieli, N., Krueger, D., & Larsen, T. (2024). *Safety cases: How to justify the safety of advanced AI systems*. arXiv. <https://doi.org/10.48550/arXiv.2403.10462>
9. Davidsen Buhl, M., Sett, G., Koessler, L., Schuett, J., & Anderljung, M. (2024). *Safety cases for frontier AI*. arXiv. <https://doi.org/10.48550/arXiv.2410.21572>
10. Sharkey, L., et al. (2025). *Open problems in mechanistic interpretability*. arXiv. <https://doi.org/10.48550/arXiv.2501.16496>
11. Bengio, Y., et al. (2026). *Safety from honesty in a disinterested AI predictor*. arXiv. <https://doi.org/10.48550/arXiv.2606.29657>
12. World Commission on the Ethics of Scientific Knowledge and Technology. (2005). *The precautionary principle*. UNESCO.
13. Makins, O., Angelini, O., Shams, Z., & Phuong, M. (2026). *Multi-agent AI control: Distributed attacks hamper per-instance monitors*. arXiv. <https://doi.org/10.48550/arXiv.2607.07368>
14. Langosco Di Langosco, L., Koch, J., Sharkey, L. D., Pfau, J., & Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 12004–12019). PMLR.
15. Bengio, Y. (2026, September 11). *Why are AI agents lying, cheating and coordinating?* [Blog post].
16. Kasirzadeh, A., & Gabriel, I. (2026). Agentic profiles for effective AI governance. *Nature*, 656, 320–328.
17. Shevlane, T., et al. (2023). *Model evaluation for extreme risks*. arXiv. <https://doi.org/10.48550/arXiv.2305.15324>
18. Shavit, Y., et al. (2023). *Practices for governing agentic AI systems*. OpenAI.
19. Hugging Face. (2026, July 16). *Security incident disclosure—July 2026*. Hugging Face. (2026, July 27). *Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident*.
20. Independent International Scientific Panel on AI. (2026). *Preliminary Report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of AI*. United Nations. <https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report>
21. Amodei, D., et al. (2016). *Concrete problems in AI safety*. arXiv. <https://doi.org/10.48550/arXiv.1606.06565>
22. Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.
23. Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3407–3412).
24. Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
25. Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. <https://doi.org/10.48550/arXiv.2212.08073>
26. DeepSeek-AI. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645, 633–638. <https://doi.org/10.48550/arXiv.2501.12948>
27. Luo, X., Zhang, Y., He, Z., Wang, Z., Zhao, S., Li, D., Qiu, L. K., & Yang, Y. (2025). *Agent Lightning: Train ANY AI agents with reinforcement learning*. arXiv. <https://doi.org/10.48550/arXiv.2508.03680>
28. Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology.
29. Sharma, M., et al. (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2310.13548>
30. Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., & Troy, K. K. (2025). *Agentic misalignment: How LLMs could be insider threats*. arXiv. <https://doi.org/10.48550/arXiv.2510.05179>
31. Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2026). Incomplete tasks induce shutdown resistance in some frontier LLMs. *Transactions on Machine Learning Research*. <https://doi.org/10.48550/arXiv.2509.14260>
32. Vann, M. G. (2003). Of rats, rice, and race: The great Hanoi rat massacre, an episode in French colonial history. *French Colonial History*, 4, 191–204.
33. Phuong, M., Jenner, E., Simon, L., Ho, L., Shah, R., Farquhar, S., & Coull, S. (2026). *GDM AI control roadmap*. arXiv. <https://doi.org/10.48550/arXiv.2607.13087>
34. Turner, A. M., Smith, L., Shah, R., Critch, A., & Tadepalli, P. (2021). Optimal policies tend to seek power. *Advances in Neural Information Processing Systems*, 34.

35. Hammond, L., et al. (2025). *Multi-agent risks from advanced AI* (Cooperative AI Foundation Technical Report No. 1). Cooperative AI Foundation. <https://doi.org/10.48550/arXiv.2502.14143>
36. Irregular. (2026, September 16). *Agentic self-modification in open-weights systems*.
37. Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S. J., & Dragan, A. (2017). Inverse reward design. *Advances in Neural Information Processing Systems*, 30.
38. Manheim, D., & Garrabrant, S. (2018). *Categorizing variants of Goodhart's law*. arXiv. <https://doi.org/10.48550/arXiv.1803.04585>
39. Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33.
40. Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 10835–10866). PMLR.
41. Karwowski, J., Hayman, O., Bai, X., Kiendlhofer, K., Griffin, C., & Skalse, J. (2024). Goodhart's law in reinforcement learning. In *International Conference on Learning Representations (ICLR)*.
42. Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198, 6435–6467.
43. Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.
44. Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 16295–16336). PMLR.
45. Anderson, R., & Moore, T. (2006). The economics of information security. *Science*, 314(5799), 610–613.
46. Weil, G. (2024). *Tort law as a tool for mitigating catastrophic risk from artificial intelligence*. SSRN. <https://doi.org/10.2139/ssrn.4694006>
47. Ben-Shahar, O., & Logue, K. D. (2012). Outsourcing regulation: How insurance reduces moral hazard. *Michigan Law Review*, 111, 197–248.
48. European Parliament & Council of the European Union. (2009). Directive 2009/103/EC relating to insurance against civil liability in respect of the use of motor vehicles. *Official Journal of the European Union*, L 263, 11–31.
49. European Parliament & Council of the European Union. (2004). Regulation (EC) No. 785/2004 on insurance requirements for air carriers and aircraft operators. *Official Journal of the European Union*, L 138, 1–6.
50. Hadfield, G. K., & Clark, J. (2026). Regulatory markets: The future of AI governance. *Jurimetrics*, 65, 195–240. <https://doi.org/10.48550/arXiv.2304.04914>
51. Barach, P., & Small, S. D. (2000). Reporting and preventing medical mishaps: Lessons from non-medical near miss reporting systems. *BMJ*, 320, 759–763.
52. World Health Organization. (2020). *Patient safety incident reporting and learning systems: Technical report and guidance*. World Health Organization.
53. McGregor, S. (2021). Preventing repeated real world AI failures by cataloging incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15458–15463.
54. Kolt, N., Anderljung, M., Barnhart, J., Brass, A., Esvelt, K., Hadfield, G. K., Heim, L., Rodriguez, M., Sandbrink, J. B., & Woodside, T. (2024). Responsible reporting for frontier AI development. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 768–783.
55. European Parliament & Council of the European Union. (2019). Directive (EU) 2019/1937 on the protection of persons who report breaches of Union law. *Official Journal of the European Union*, L 305, 17–56.
56. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). *Large language models often know when they are being evaluated*. arXiv. <https://doi.org/10.48550/arXiv.2505.23836>
57. van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2025). AI sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2406.07358>