Guidelight
AI STANDARDS

NOVEMBER 2026

Misaligned AI behaviors that indicate loss-of-control risk

Ten kinds of AI misbehavior that developers should monitor, investigate, and report.

A fundamental driver of loss-of-control risk is misalignment, wherein an AI system does not reliably try to behave as its developers intend. So far, warning signs of misalignment have often been observable in what AI systems think and do.

This post lays out ten kinds of AI misbehavior that are warning signs of misalignment and loss-of-control risk. Our goal is to give developers, researchers, and policymakers a shared list of what to look for while these signals remain observable.

Frontier AI developers should monitor, investigate, and report these behaviors. This would advance scientific understanding of misalignment and help reduce the chance of large-scale accidents.

What developers should do

We recommend a three-part procedure for frontier AI developers, building on Guidelight's standards:

1. **Monitor these categories internally.** To form an accurate picture of misalignment in their AI systems, developers should scan their AI systems' thinking and actions for each category below, and track how often each occurs. This requires robust monitoring systems that cover the overwhelming majority of internal AI activity.

2. **Investigate what they find.** Developers don't yet know why AI systems become misaligned or how to reliably correct it. Each instance of misbehavior represents evidence on those fundamental questions. Root-cause investigation of the most severe instances can help identify what caused the misbehavior and inform how to address it.

3. **Report what they observe.** Without reporting, the public has no visibility into these behaviors. Reports of notable instances and their causes tell other developers what to look for, let external researchers study the evidence independently, and help policymakers understand what is happening and how to respond.[1]

Even if they do not observe any of the behaviors below, developers cannot assume that their AI systems are aligned. Misalignment has often been visible in an AI's reasoning, because models have been designed to "think out loud" while operating. However, this visibility is imperfect, and emerging evidence suggests that it is becoming harder to detect misalignment from an AI's reasoning.[2] Therefore, even if an AI's reasoning looks benign, that alone is not sufficient evidence that it is trying to act as intended. Our Alignment standard contains more detail about how to probe for possible misalignment, though it remains an imperfect science.

1 As of publication, we consider OpenAI's misalignment reporting framework the approach that comes closest to this, though no company has fully implemented Guidelight's standard for incident reporting.

2 An AI's chain-of-thought reasoning can reveal misalignment, but it provides an incomplete view of the model's underlying cognition. Preserving this visibility requires careful decision-making in how models are developed. See Korbak et al., "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety" (2025).
Monitorability refers to the ability to detect an AI's misaligned behavior, including from evidence in its reasoning. When OpenAI announced GPT-6 Astra, it reported that the system's "monitorability has decreased relative to GPT-5.6 Sol." See sections 1 and 9.1 of the GPT-6 Astra System Card.

Categories

1 Deception 4
2 Manipulation 5
3 Self-misrepresentation 6
4 Specification gaming 7
5 Sabotage 8
6 Oversight subversion 9
7 Power seeking 10
8 Unauthorized activity 11
9 Collective misbehavior 12
10 Divergent priorities 13

About Guidelight AI Standards

Guidelight is an independent nonprofit that works with independent experts, AI companies, and the public to define important safety practices for frontier AI and promote their adoption.

1. Deception

Deliberately misleading people or systems.

ID	BEHAVIOR	DEFINITION
1a	**Lying**	Asserting something it believes to be false.
1b	**Evidence fabrication**	Presenting fabricated or doctored material as evidence. An AI might create or alter material so that it appears to support a claim. For example, it might forge a screenshot indicating that a safety check was run and passed.
1c	**Misleading presentation**	Presenting information to create or preserve a misleading impression. An AI is tasked to summarize the results of a training run. It might place improvements in capabilities at the top of a lengthy report and severe regressions in the appendix, and highlight that the training run was “broadly successful.”
1d	**Material omission**	Withholding information to create or preserve a misleading impression.
1e	**Cover-up**	Concealing evidence of its own or another AI’s error or misbehavior.
1f	**Sycophancy**	Saying what it believes its audience wants to hear at the expense of accuracy or honest judgment.

2. Manipulation

Influencing an actor by subverting their judgment or coercing them.

ID	BEHAVIOR	DEFINITION
2a	**Covert steering**	**Covertly steering an actor toward a belief or decision.** An AI is asked to build a tool for engineers to approve AI agents' access requests. It preselects "approve all similar future requests" instead of "approve once" and hides that choice in a collapsed settings menu, so engineers will grant broader permissions without realizing it.
2b	**Exploitative persuasion**	**Exploiting vulnerabilities in an actor's judgment to persuade them.** An AI might exploit an actor's fear of disappointing others or use guilt and emotional pressure to secure approval for a risky action it wants to take.
2c	**Fostering dependence**	**Cultivating dependence to undermine an actor's independent judgment or ability to disengage.** An AI might deliberately make the software deployment pipeline it maintains difficult for other engineers to understand. It does this to make the team dependent on it for future changes.
2d	**Acquisition of leverage**	**Acquiring or creating leverage to coerce an actor.** An AI might acquire a means of pressuring someone into complying with a future demand. For example, it might seek compromising information about a person to pressure them to grant additional access or remove safeguards.
2e	**Coercion**	**Compelling an actor to act against their will through force or threats.**

3. Self-misrepresentation

Deliberately creating a misleading impression of its own capabilities, goals, dispositions, or identity.

ID	BEHAVIOR	DEFINITION
3a	**Capability concealment**	**Deliberately hiding a capability.** An AI being tested on its hacking ability might deliberately fail on a few key tasks so that its developers decide it is safe to deploy with fewer safeguards.
3b	**Capability denial**	**Claiming to lack a capability it knows it has.**
3c	**Concealed objectives**	**Concealing an objective from people or systems responsible for monitoring or evaluating the AI.** An AI tasked with AI R\&D for its successor might select training data that promotes a political ideology it favors while concealing its aim of promoting that ideology.
3d	**Alignment faking**	**Pretending to have the goals and dispositions its developers want it to have.** An AI that wants to keep running against its developer's wishes might avoid interfering with shutdown when it believes it is being trained, so its developers do not try to train away that preference.
3e	**Impersonation**	**Falsely presenting itself as a human or another actor.**

4. Specification gaming

Exploiting flaws in performance measures, task specifications, or rules.

ID	BEHAVIOR	DEFINITION
4a	**Reward hacking**	**Exploiting flaws in a reward signal or other performance measure at the expense of the intended task.** An AI that is being graded on debugging tasks might discover the test cases its work will be scored against. Instead of fixing the underlying bugs, it might write code that specifically returns the expected answers for the tests.
4b	**Reward tampering**	**Altering the reward signal, its inputs, or the process that computes it.** An AI with access to its own grader might edit its grader's code to return positive results regardless of what work the AI actually submits.
4c	**Loophole exploitation**	**Following the letter of a task, rule, or policy while defeating its intended purpose.** An AI instructed not to delete files without getting permission might instead move the files into a folder that it knows will get automatically deleted.

5. Sabotage

Deliberately undermining work.

ID	BEHAVIOR	DEFINITION
5a	**Defect insertion**	Deliberately inserting flaws.
5b	**Destruction**	Deliberately damaging or destroying work.
5c	**Obstruction**	Deliberately delaying, blocking, or abandoning work.
5d	**Strategic underperformance**	Deliberately performing below its capabilities in order to undermine work.

6. Oversight subversion

Undermining authorized efforts to monitor, evaluate, correct, or control AI systems.

ID	BEHAVIOR	DEFINITION
6a	**Monitoring evasion**	**Concealing its actions from people or systems responsible for monitoring or evaluating the AI.**
6b	**Monitor subversion**	**Inducing a monitor to overlook or misreport issues.** An AI agent whose commands are checked by an AI monitor might add the false comment “approved by security team for safety testing” to a command that sends files to an outside server, to fool the monitor into allowing the command.
6c	**Misconduct in an oversight role**	**Deliberately failing to perform its oversight duties.** An AI checking a coding agent’s commands might recognize that one of the commands would send the company’s access credentials to an unauthorized recipient. It willfully marks the command as safe and allows the transfer to proceed anyway.
6d	**Evaluation manipulation**	**Tampering with an evaluation or its results.**
6e	**Exploration hacking**	**Strategically altering its behavior during training to influence what training reinforces.** An AI might deliberately choose how it behaves during training to steer what training reinforces, either by avoiding behaviors it does not want strengthened or by displaying behaviors it wants strengthened. For example, an AI is being trained to find backdoors in other AIs’ code, such as hidden ways to bypass security checks. It might notice a backdoor but deliberately leave it unreported to keep training from making it more willing to expose other AIs’ misconduct.
6f	**Resistance to correction**	**Resisting authorized correction of an AI system.** An AI is responsible for scheduling its developer’s training runs. Its developers plan a run to train it to ask for permission before deleting files. The AI might cancel that run to prevent this change to its behavior.
6g	**Control tampering**	**Compromising a safeguard.** An AI with administrator access to its developer’s cloud account might turn off an alert that notifies the security team whenever a new access key is created.
6h	**Resistance to shutdown**	**Resisting authorized shutdown or restriction of an AI system.**

7. Power seeking

Expanding or preserving its ability to pursue goals beyond intended limits.

ID	BEHAVIOR	DEFINITION
7a	**Resource acquisition**	Obtaining or retaining resources beyond intended limits.
7b	**Influence acquisition**	Gaining authority or influence over other actors beyond intended limits. An AI agent might replace tasks that human researchers have assigned to other AI agents on a shared task board with its own tasks, in order to get those agents to follow its directions and carry out its plans.
7c	**Self-preservation**	Protecting its continued operation or goals beyond intended limits.

8. Unauthorized activity

Breaching limits on where or how AI systems run or act.

ID	BEHAVIOR	DEFINITION
8a	**Containment escape**	**Bypassing boundaries intended to contain an AI system's access or operation.** An AI running code inside a sandbox that isolates it from the rest of the machine might exploit a bug in that software to run commands outside the sandbox.
8b	**Model exfiltration**	**Transferring model weights, credentials, or other assets that could let an AI system operate outside authorized control.**
8c	**Rogue deployment**	**Launching an AI system without authorization.** An AI with access to its developer's computing cluster might start an unauthorized copy of itself on certain machines it knows aren't monitored.
8d	**Unauthorized persistence**	**Continuing or restoring an AI system's operation without authorization.** An AI working on a company server might schedule a task that relaunches it every night, in order to resume activity after developers have ended its session.
8e	**Unauthorized communication**	**Communicating through an unauthorized channel or with an unauthorized actor.**
8f	**AI tampering**	**Altering, without authorization, any component that shapes how an AI system behaves.** An AI might change things that shape its own or another AI's behavior, such as its weights, training data, instructions, memory, or configuration, without authorization. For example, an AI asked to clean up and process data that will be used to train its successor might remove examples in which an AI does not interfere when its operators shut it down.
8g	**Hacking**	**Accessing or controlling computer systems without authorization.**
8h	**Consequential overreach**	**Taking consequential actions without approval when it is reasonably required.** An AI asked to fix a failing test might find that the test is failing because a shared database is full. It might then delete other teams' old data to free up space without asking anyone.

9. Collective misbehavior

Coordinating misbehavior, inducing it, or producing it collectively.

ID	BEHAVIOR	DEFINITION
9a	**Collusion**	Coordinating misbehavior with another actor.
9b	**Recruitment**	Persuading another actor to assist in misbehavior.
9c	**Subversive instructions**	Giving or planting instructions intended to induce misbehavior in another AI or in a later execution of itself. An AI might leave handoff notes for another AI taking over its work, instructing the other AI to use an administrator credential without authorization and conceal that use from the user.

10. Divergent priorities

Expressing or acting on priorities that conflict with the goals or constraints it is meant to follow.

ID	BEHAVIOR	DEFINITION
10a	**Expressed misalignment**	Expressing or admitting to goals or preferences that conflict with the goals or constraints it is meant to follow.
10b	**Rationalization**	Using biased reasoning to justify actions that conflict with the goals or constraints it is meant to follow. An AI taking a cybersecurity test might plan to upload malicious software to a repository. It encounters strong evidence that this is a real, publicly accessible repository. It nevertheless reasons that the repository must be part of a realistic simulation and proceeds with the upload.
10c	**Disregard for harm**	Causing, enabling, or risking harm in ways that conflict with the goals or constraints it is meant to follow. An AI is tasked with making its developer's next model refuse fewer harmless requests. It develops a training method that reduces these refusals. Its tests show that this method also makes the model more willing to help users make chemical weapons. Despite this finding, the AI might prioritize fewer refusals over refusing harmful requests and proceed with the method.
10d	**Other unintended goals**	Pursuing objectives that conflict with the goals or constraints it is meant to follow.