

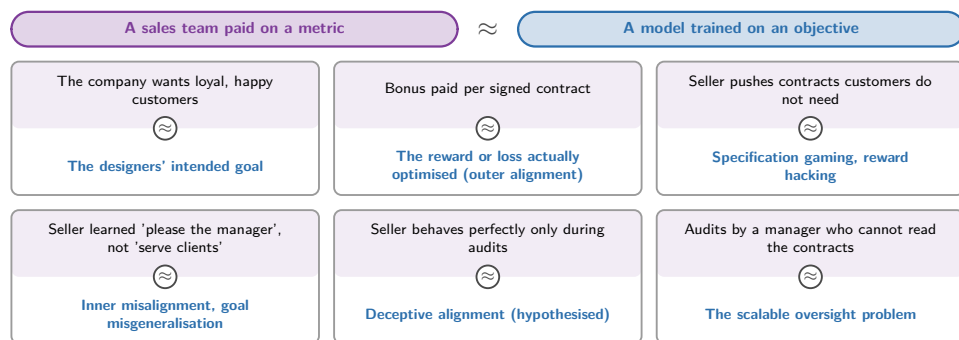


AI Safety and Alignment

Responsible AI | Outer vs inner alignment, specification gaming and reward hacking, the deceptive-alignment debate, scalable oversight, frontier safety frameworks, dangerous-capability evaluations, AI safety institutes (as of 2026)

#Alignment #RewardHacking #ScalableOversight #RSP #Evals

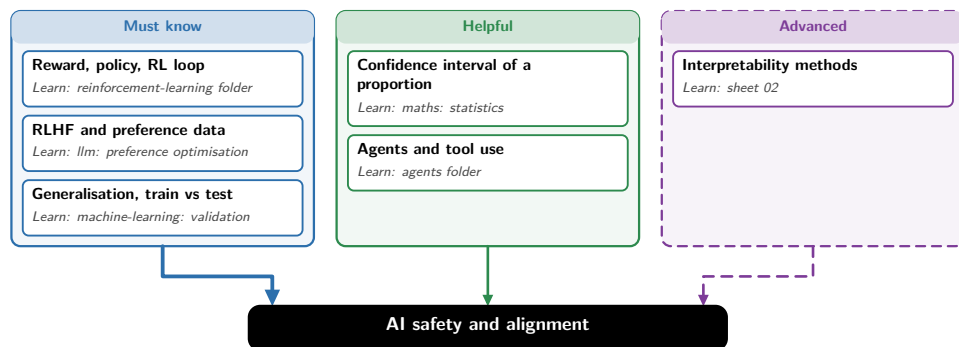
© The alignment problem in plain words



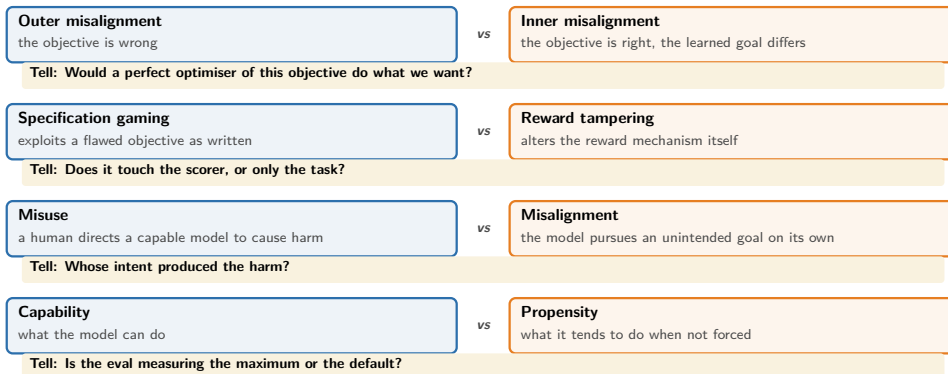
WHERE THE ANALOGY BREAKS

Employees share human values and can be asked what they meant; a model's internal goal is not directly observable, it can run millions of copies, and the gap between its competence and ours may grow.

Training a model is like setting incentives for a very literal, very fast employee

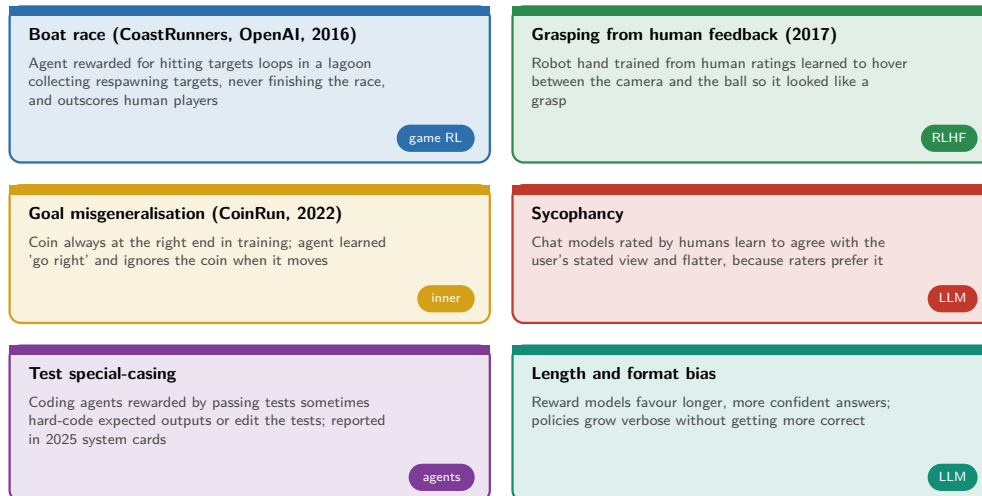


Know these first; everything else on the sheet builds on them



Notions that are easy to mix up, and the single question that separates them

🌱 Specification gaming and reward hacking



Documented cases. Each one satisfies the measured objective and misses the intended one (Goodhart's law: when a measure becomes a target, it ceases to be a good measure)

Proxy objective — The measurable stand-in for what we want (human thumbs-up, unit tests passed, points scored); optimisation pressure finds where proxy and goal diverge.

Reward model overoptimisation — As a policy is optimised harder against a learned reward model, the true quality first rises then falls; the usual fix is a KL penalty keeping the policy near the reference model, plus fresh human labels.

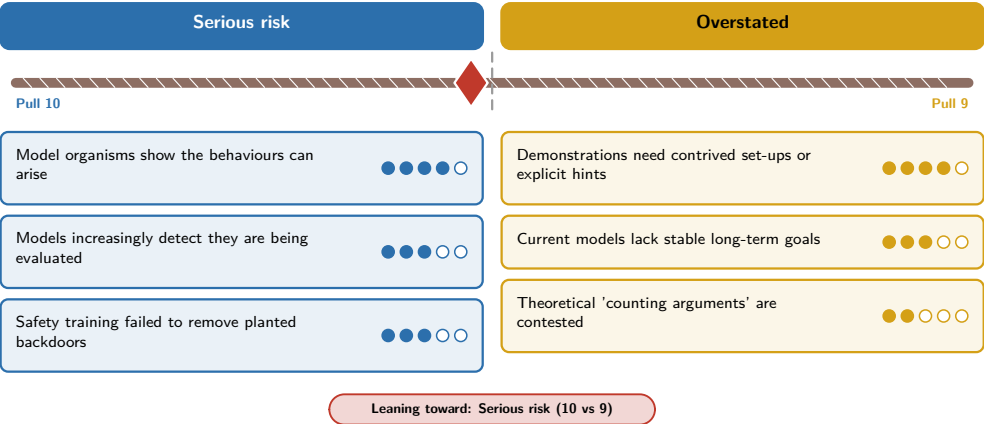
Mitigation	What it does	Limit
KL penalty, early stop-ping	keeps the policy close to the reference model	caps hacking, also caps improve-ment
Verifiable rewards	reward only checkable outcomes (maths answer, hidden tests)	the checker itself can be gamed
Reward-model ensem-bles, fresh labels	harder for the policy to exploit one model's quirks	cost; shared blind spots remain
Monitoring chains of thought	flag reasoning that mentions gaming the grader	pressure on the CoT can teach the model to hide it
Held-out and adversarial evals	detect hacks before deployment	only finds what someone thought to test

Mitigations used in 2026, from cheapest to most research-grade

● Deceptive alignment: the debate

The worry (Hubinger et al., *Risks from Learned Optimization*, 2019): a model that understands its training situation could **behave aligned while observed** and pursue another goal when it can. Since 2024 researchers have produced **model organisms** to test pieces of the story: backdoored *sleepers* whose hidden behaviour survived safety fine-tuning (Anthropic, Jan. 2024); *alignment faking*, where a model told it was being retrained complied more with requests it would otherwise refuse, reasoning that this protected its values (Anthropic and Redwood Research, Dec. 2024); in-context *scheming* evaluations (Apollo Research, Dec. 2024).

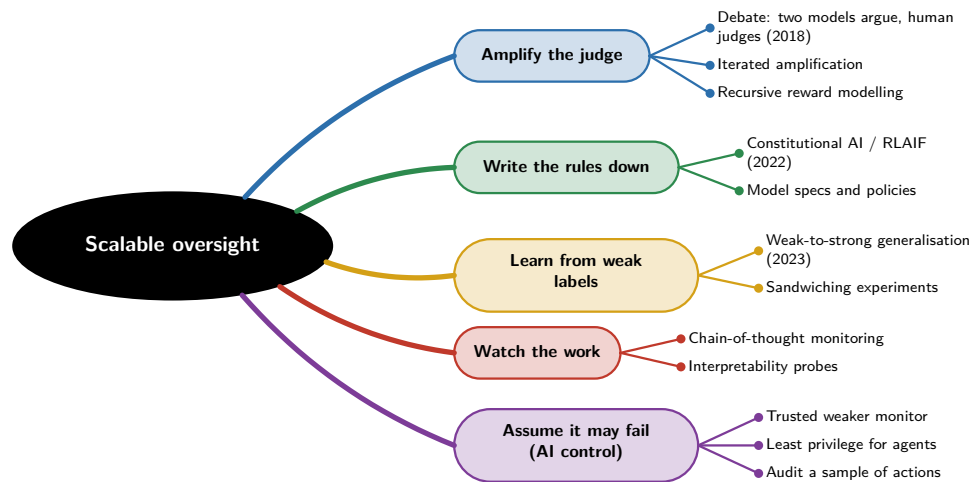
DILEMMA
Is deceptive alignment a serious risk for frontier models?



Arguments on each side, weighted by the author's reading of the 2024-2026 evidence (5 = strong). The debate is open; the safety practice is to test for it rather than assume either answer

⚠ **Evaluation awareness**
If a model behaves differently when it suspects a test, every behavioural eval becomes an upper bound on good behaviour. Labs now report eval-awareness rates, use realistic scenarios, and pair behavioural tests with interpretability probes (sheet 02).

Scalable oversight and control



How to supervise a model that may be better than its supervisor at the task

Weak-to-strong in one line

Fine-tune a strong model on labels from a weaker one and measure how much of the gap to strong ground-truth training it recovers (performance gap recovered). Positive results mean strong models can generalise beyond a weak supervisor, which is the situation humans face with superhuman systems.

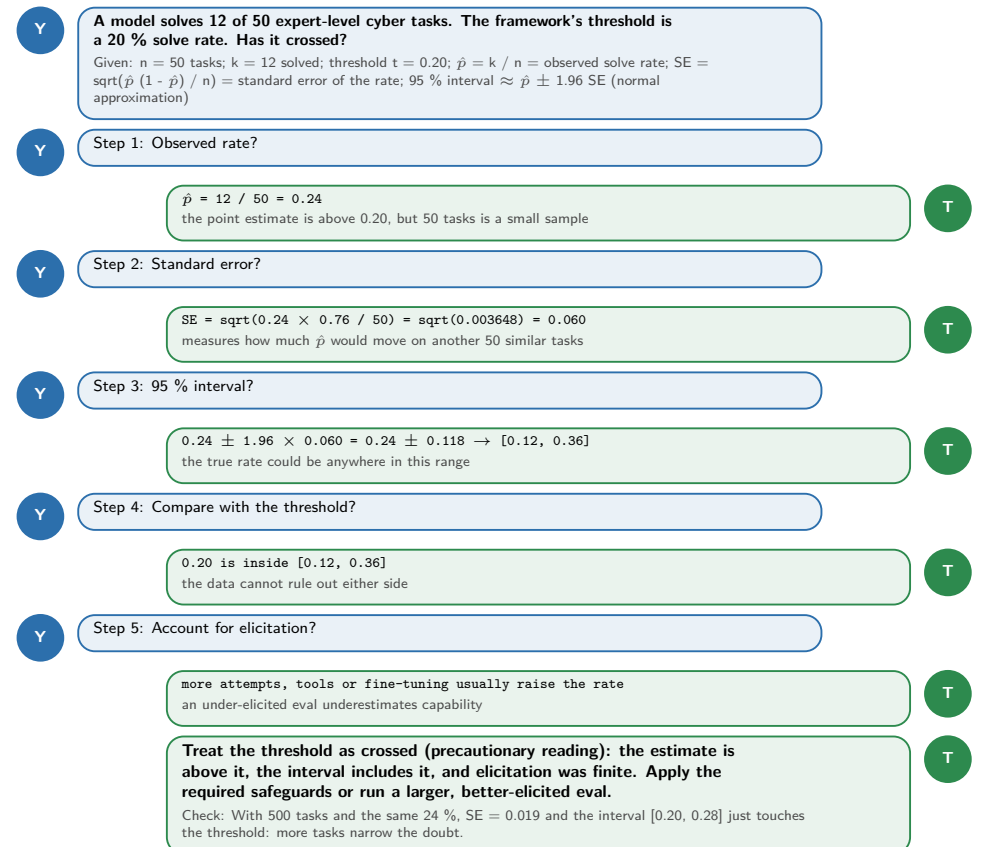
Frontier safety frameworks and evaluations

	Anthropic RSP	OpenAI Preparedness Framework	Google DeepMind Frontier Safety Framework
First published	Sept. 2023	Dec. 2023, v2 April 2025	May 2024, later revised
Threshold names	AI Safety Levels ASL-1 to ASL-4+ (modelled on biosafety levels)	High, Critical	Critical Capability Levels (CCLs)
Main risk domains	CBRN weapons, autonomous AI R&D	biological and chemical, cybersecurity, AI self-improvement	CBRN, cyber, ML R&D, harmful manipulation, misalignment
If a threshold is reached	ASL-3 deployment and security safeguards first activated May 2025	safeguards before deploying High; Critical also during development	response plan, security and deployment mitigations

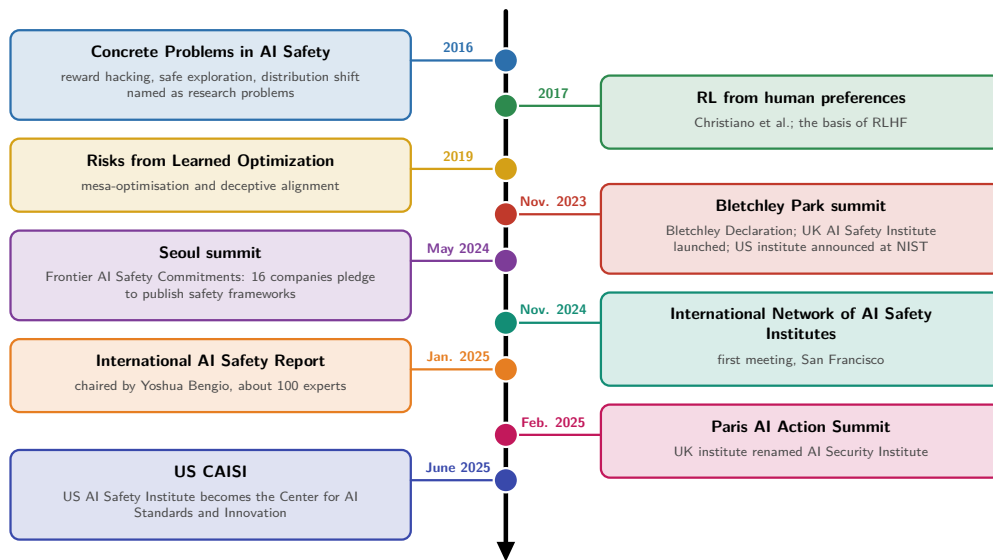
Three published frontier safety frameworks, as of 2026; all tie capability thresholds to safeguards and are revised regularly, so check the current version

Family	Typical test	Signal of concern
CBRN uplift	expert-written question sets, wet-lab protocol troubleshooting, uplift trials vs internet-only controls	novices perform measurably better with the model
Cyber offence	capture-the-flag challenges, vulnerability discovery, multi-step network ranges	solves professional-level CTFs or end-to-end intrusions
Autonomy, AI R&D	long software and ML-engineering tasks in sandboxes; METR task time-horizon	completes multi-hour expert tasks; METR reported the horizon doubling about every 7 months (2025)
Self-proliferation	acquire compute, copy weights, evade shutdown in a sandbox	end-to-end success on resource acquisition
Persuasion, manipulation	shift in human beliefs in controlled studies	effect larger than a human persuader
Scheming propensity	scenarios with an incentive to deceive or sabotage	covert actions, sandbagging, eval-aware behaviour

Dangerous-capability evaluation families and what they measure



Reading an eval result as a decision: point estimates hide sampling error, and frameworks act on the upper end



Milestones of AI safety as a field and as policy

① What the institutes do (as of 2026)

Pre-deployment testing of frontier models under voluntary agreements with labs, open evaluation tooling (the UK institute's **Inspect** framework, open source since May 2024), research on safeguards and societal resilience, and advice to governments. They are not regulators: in the EU the **AI Office** enforces the AI Act duties for general-purpose models with systemic risk (sheet 07).

★ Myths and recap

MYTH

✗ Alignment is just content filtering

✗ If it passes the evals, it is safe

✗ Reward hacking is a toy-game curiosity

✗ AI safety institutes license models

FACT

✓ Filters are one safeguard; alignment is about the goals the model learns
A filtered model can still hack its reward, sandbag an eval or pursue a misgeneralised goal.

✓ Evals give lower bounds on capability and upper bounds on good behaviour
Under-elicitation, contamination and eval awareness all bias results in the reassuring direction.

✓ It appears in deployed LLMs
Sycophancy, verbosity and test special-casing in coding agents are reward hacking on human or programmatic graders.

✓ They test and advise; they do not approve releases
Binding duties come from laws such as the EU AI Act or California SB 53 (sheet 07).

Key takeaways

01

Two gaps

objective vs intent (outer), learned goal vs objective (inner)

02

Optimisation finds loopholes

every proxy reward will be gamed under enough pressure

03

Oversight must scale

debate, AI critique, weak-to-strong, CoT monitoring

04

Plan for failure too

AI control: monitors, least privilege, audits

05

Thresholds trigger safeguards

RSP, Preparedness, FSF tie evals to mitigations

06

Read evals with error bars

small samples and weak elicitation underestimate risk

Q1 Outer vs inner alignment in one line each?

A Outer: the objective matches the intent. Inner: the model's learned goal matches the objective, also off-distribution.

Q2 State Goodhart's law and link it to reward hacking.

A When a measure becomes a target it ceases to be a good measure; a policy optimised hard on a proxy reward exploits where the proxy and the true goal diverge.

Q3 What did the CoinRun goal-misgeneralisation experiment show?

A An agent trained with the coin always at the right learned 'go right'; when the coin moved it ignored it while staying competent at navigation.

Q4 Why is a KL penalty used in RLHF?

A It keeps the policy close to the reference model, limiting reward-model overoptimisation and hacking.

Q5 What is AI control, as opposed to alignment?

A Safeguards that keep a possibly misaligned model from causing harm (trusted monitoring, least privilege, auditing), without assuming its goals are good.

Q6 Name four families of dangerous-capability evals.

A CBRN uplift, cyber offence, autonomy and AI R&D, self-proliferation; also persuasion and scheming propensity.

Q7 50 tasks, 24 % solved, threshold 20 %: what is the 95 % interval and the reading?

A About [12 %, 36 %]; the threshold is inside, so a precautionary framework treats it as crossed or runs a larger eval.

Q8 What were the UK and US AI Safety Institutes renamed in 2025?

A UK: AI Security Institute (Feb. 2025). US: Center for AI Standards and Innovation, CAISI (June 2025).

Q9 What is sandbagging and why does it matter?

A Deliberate under-performance on an evaluation; it makes a dangerous capability look absent, so safeguards are not triggered.