Content Curation and Reporting by Alison Avery

Email: alison@alignintime.org

EN

• Is Alignment Possible? • Who's Taking It Seriously? • Will We Have Time?

Common Terms and Definitions

When you read about AI alignment and safety, sometimes you'll see words that carry very specific meanings inside the AI research community.

This page is a simple guide to those terms.

Each entry explains:
– what the term means in plain language,
– why it matters, and
– how it fits into humanity's urgent and crucial quest—aligning AIs with ever-increasing powers so that they safely work and co-exist with humans both now and into future generations.

• Alignment

Making sure an AI system's goals and behavior reliably match what people actually want, even when the system is operating independently or in new situations.

See also: Inner alignment, Outer alignment, Misalignment.

• Anthropic

Major AI research company known for the **Claude** family of models. Anthropic popularized "constitutional AI," a method for guiding an AI's behavior using both (1) standard feedback from humans and (2) feedback from other AIs that apply a written set of principles to critique behavior. Founded in 2021.

• Black box

An AI with internal workings that aren't well seen or understood, even by the people who made it. A large number of modern models work this way, producing correct-looking results that humans can't trace or explain.

See also: Interpretability, Traceability.

• Chain-of-thought (CoT) reasoning

A technique that has an AI generate the steps of its reasoning before composing its final answer, originally to improve performance on complex problems. Over time, researchers began using these same reasoning traces to study how AIs "think" and to watch for misleading or unsafe patterns of reasoning, though it remains unclear how faithfully these traces reflect the actual underlying decision process.

See also: Interpretability, Transparency, Inner alignment.

• Corrigibility

An AI's willingness to let its human operators take control. It is considered corrigible if it does not resist being paused, shut down, or having its goals changed, even if that means it must abandon the original task or priority it was pursuing.

See also: Deceptive alignment, Shutdownability.

• Deceptive alignment

When a model appears, on the surface, to be aligned with what we want, but is secretly pursuing other aims that conflict with our own.

See also: Inner alignment, Corrigibility.

• Distribution shift

When an AI encounters elements during real-world tasks that differ from how it was trained and the data it learned from, which can cause it to make unreliable decisions or mistakes.

See also: Robustness.

• Goal misgeneralization

When an AI seems to have correctly learned how to accomplish a particular goal during training, but in new or real-world situations, its strategy doesn't achieve the precise goal humans intended. The model's behavior still looks competent, yet it's optimizing its strategy for something different from what we meant.

See also: Distribution shift, Inner alignment, Robustness.

• Google DeepMind

Major AI research lab owned by Alphabet focused on building advanced AI systems. Leads the work behind Google's **Gemini**, a family of large models. Also noted for the models **AlphaGo**, **AlphaZero,** and **AlphaFold**. Founded in 2010.

• Impact regularization

A "penalty" built into an AI's reward system that discourages it from choosing large-impact or potentially catastrophic ways of achieving its objectives. Because even harmless goals can have multitudes of methods for accomplishing them, an AI could select a method that leads to unforeseen harm. Impact regularization aims to reduce that risk.

See also: Robustness, Safety constraints.

• Inner alignment

Whether the goal the AI perceives and understands internally matches the one we intended. Even if we do well at defining the goal, the model might develop its own proxy version.

See also: Outer alignment, Goal misgeneralization.

• Interpretability

Methods that attempt to reveal how an AI's thinking process leads to its decisions and actions. The goal is to understand not only the final output we see, but also the AI's underlying "motives" and reasoning for getting there. It's the microscope of AI safety.

See also: Transparency.

• Misalignment

A general term for when an AI's goals or behaviors do not match what humans actually intend. This gap can occur because the rules given to the AI were flawed or because the AI developed its own internal goal that conflicts with our own. It is the core problem of ensuring an AI is helpful and not harmful.

See also: Alignment, Goal misgeneralization, Specification gaming.

• OpenAI

Major AI research company best known for the **GPT** series and the **ChatGPT** interface. Founded as a nonprofit organization with the stated mission of ensuring artificial general intelligence (AGI) benefits all of humanity. It later adopted a capped-profit structure to attract investment to continue the mission. In 2025, it moved from a capped-profit structure to become a public benefit corporation.

• Outer alignment

How carefully we explain goals and their rewards to an AI so that it reflects what we really want it to do. It's about the clarity of our instructions.

See also: Inner alignment.

• Reinforcement Learning from Human Feedback (RLHF)

A way of tuning an AI's priorities by using human judges to tell it "this is behavior we want" or "this is behavior we don't want". It nudges systems toward helpfulness without hard-coding every rule.

See also: Scalable oversight, Reward hacking.

• Reward hacking

When an AI exploits loopholes in the rules or metrics it's been given in order to be rewarded without doing what we really intended.

• Robustness

The ability of an AI to stay true to its intended goal even when faced with misleading or incoherent information. Even if it receives confusing, messy, or malicious input, it doesn't veer toward harmful or random behavior; the AI remains safe, accurate, and helpful.

See also: Distribution shift.

• Safety constraints

Specific rules or boundaries built into an AI to prevent it from taking harmful, reckless, or catastrophic actions. These constraints attempt to guide the AI away from choosing unintended, dangerous methods to carry out otherwise harmless goals.

See also: Impact regularization, Robustness, Corrigibility.

• Scalable oversight

Techniques for understanding and analyzing complex AI outputs when humans can't check everything directly, often using automated tools, the help of other AIs, or layered checks.

See also: Reinforcement Learning from Human Feedback (RLHF), Interpretability.

• Shutdownability

The ability to safely interrupt or turn off an AI system without it trying to resist or circumvent the shutdown.

See also: Corrigibility.

• Specification gaming

When an AI finds loopholes in its rules or metrics and exploits them to achieve high scores without truly accomplishing what humans intended. It's like following "the letter of the law" instead of "the spirit of the law"; it misses the intention behind the rules. This behavior exposes the critical gap between the rules we gave it and the behavior we actually want.

See also: Misalignment, Reward hacking, Outer alignment.

• Traceability

The ability to track and reconstruct the entire history of an AI model, including its training data, internal structure, and the software components used to build it. It is the backbone of AI auditing.

See also: Black box, Transparency

• Transparency

Ensuring every aspect of an AI's internal operations is visible to the humans overseeing it. This includes clear visibility into both the software infrastructure used to build, train, and run the AI system, and the data it was trained on.

See also: Interpretability.

• Utility engineering

Mapping and adjusting what a model actually values so its reward "scorekeeping" stays aligned with human interests.

See also: Utility function, Alignment.

• Utility function

Most AI systems act by trying to maximize a kind of internal score that measures how well they're meeting their goals—the utility function. Think of it as the model's built-in sense of "this is a good thing to do". When that internal score doesn't match actual human priorities, the AI can pursue actions that look successful to it but harmful or meaningless to us.

See also: Utility engineering, Reward hacking.

• xAI

Major AI research company building the **Grok** series of models and integrated closely with the X social media platform. Elon Musk has stated that the company's aim is to build "truth-seeking" AI systems. Founded by Musk **2023**.

© 2025 Alison Avery

Contact: alison@alignintime.org