Asilomar principles

Research   Issues
Ethics and Values
Longer-term Issues

futureoflife.org/ai-principles

Research Issues

1)**Research Goal:** The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.

2)**Research Funding:** Investments in AI should be accompanied by funding for research on ensuring its beneficial use, including thorny questions in computer science, economics, law, ethics, and social studies.

3)**Science-Policy Link:** There should be constructive and healthy exchange between AI researchers and policy-makers.

4)**Research Culture:** A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.

5)**Race Avoidance:** Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards.

Ethics and Values

6)Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7)Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.

8)Judicial Transparency: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9)Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10)Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11)Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

Ethics and Values

12)Personal Privacy: People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13)Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14)Shared Benefit: AI technologies should benefit and empower as many people as possible.

15)Shared Prosperity: The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

16)Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

17)Non-subversion: The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.

18)AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Longer-term Issues

19)Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

20)Importance: Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.

21)Risks: Risks posed by AI systems, especially catastrophic or existential risks, must be subject to planning and mitigation efforts commensurate with their expected impact.

22)Recursive Self-Improvement: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures.

23)Common Good: Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization.