Asilomar Conference

A second AI conference was held at Asilomar, following the 2015 Puerto Rico conference organized by the Future of Life Institute (FLI).

Titled the **Asilomar Conference on Beneficial AI**, it took place from **January 5–8, 2017**, at the Asilomar Conference Grounds in Pacific Grove, California.

This event built on the momentum of the 2015 conference, focusing on ensuring AI development aligns with human values and safety as the field advanced toward artificial general intelligence (AGI). Below, I provide a comprehensive overview of who organized it, when it occurred, who attended, and the principal conclusions, in no more than 4–5 paragraphs, as requested, drawing on relevant sources and keeping the response standalone.

Who Organized It: The conference was organized by the **Future of Life Institute (FLI)**, a nonprofit founded in 2014 by **Max Tegmark** (MIT cosmologist), **Jaan Tallinn** (Skype co-founder), **Viktoriya Krakovna** (DeepMind research scientist), **Anthony Aguirre** (UCSC physicist), and **Meia Chita-Tegmark** (Tufts scholar).

FLI's mission is to mitigate existential risks from transformative technologies, particularly AI, and the Asilomar conference was a direct follow-up to their 2015 Puerto Rico event.

Funding came from sponsors including **Elon Musk**, **Jaan Tallinn**, the **Open Philanthropy Project**, and the **Center for Brains, Minds, and Machines**, with Tegmark and Tallinn leading the planning to convene AI experts and thought leaders under **Chatham House Rules** for open discussion.

The organizing team, including Ariel Conn and Lucas Perry, reviewed prior AI reports to draft initial principles, which were refined during the event.

When It Occurred: The Asilomar Conference on Beneficial AI was held from **January 5–8, 2017**, at the Asilomar Conference Grounds, a historic venue chosen for its legacy from the 1975 Asilomar Conference on Recombinant DNA. The event included a two-day technical workshop for FLI grant recipients to discuss AI safety progress, followed by a 2.5-day main conference for broader discussions on AI's future, opportunities, and risks.

Who Attended: Over 100 leading AI researchers, industry experts, economists, ethicists, and philosophers attended, including **Elon Musk** (Tesla, SpaceX), **Demis Hassabis** (DeepMind CEO), **Ilya Sutskever** (OpenAI co-founder), **Yann LeCun** (Facebook AI Director), **Yoshua Bengio** (Montreal Institute for Learning Algorithms), **Stuart Russell** (UC Berkeley, AI textbook co-author), **Peter Norvig** (Google Research Director), **Ray Kurzweil**

(Google, futurist), **Nick Bostrom** (Oxford's Future of Humanity Institute), **Sam Harris** (philosopher), **Will MacAskill** (Oxford, Effective Altruism), and **Erik Brynjolfsson** (MIT, *The Second Machine Age*). Nine DeepMind researchers, alongside representatives from Google Brain, OpenAI, MIRI, Vicarious, and universities like MIT, Cornell, and Cambridge, participated. The diverse group ensured interdisciplinary perspectives on AI's societal impact.

Principal Conclusions: The conference produced the **23 Asilomar AI Principles**, a set of guidelines for ethical AI development, each supported by at least 90% of attendees. These principles, published as an open letter, emphasized:
Research Goal (creating beneficial intelligence),
Value Alignment (ensuring AI goals align with human values),
Risk Mitigation (addressing catastrophic or existential risks),
Recursive Self-Improvement (strict controls on self-improving AI), and
Common Good (developing superintelligence for humanity's benefit, not one entity).

Signed by over 1,200 AI researchers and 2,500 others, including Stephen Hawking and Joseph Gordon-Levitt, the principles became a cornerstone of AI governance, influencing policies like the EU's AI Act.

The conference highlighted the urgency of AI safety research, with discussions on economic disruption, autonomous weapons, and value alignment, building on Jaan Tallinn's 2015 presentation of AI's potential risks. It also spurred FLI's $7 million AI safety grant program, announced in 2015, to fund 37 research teams.

The Asilomar conference solidified FLI's role in shaping AI ethics, drawing parallels to the 1975 Asilomar DNA conference for its self-regulatory approach. Its principles remain influential, cited in 2025 discussions on AI regulation, though critics note their generality and lack of enforcement mechanisms. The event's legacy lies in fostering global dialogue on AI's transformative potential and risks, encouraging collaboration between industry and academia to prevent outcomes like those warned by Musk and Hinton, where uncontrolled AI could outpace human oversight.

Below are the **23 Asilomar AI Principles**, established at the Asilomar Conference on Beneficial AI, held January 5–8, 2017, by the Future of Life Institute. These principles, drafted and agreed upon by attendees, aim to guide ethical AI development and are presented as requested, without additional commentary or connection to other prompts.

Asilomar AI Principles

Research Issues

- **Research Goal**: The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.
- **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.
- **Science-Policy Link**: There should be constructive and healthy exchange between AI researchers and policymakers.
- **Research Culture**: A culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI.
- **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.
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**: AI systems with the potential to cause catastrophic damage should be subject to strong safety and security controls.
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.