Asilomar AI Principles

阿西洛马人工智能原则 23 条

Research Issues 科研问题

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

研究目的：人工智能研究的目标，应该是创造有益(于人类)而不是不受(人类)控制的智能。

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, such as:

研究经费：投资人工智能应该有部份经费()用于研究如何确保有益地使用人工智能，包括计算机科学、经济学、法律、伦理以及社会研究中的棘手问题，比如：

How can we make future AI systems highly robust, so that they do what we want without malfunctioning or getting hacked?
如何使未来的人工智能系统高度健全(“鲁棒性”)，让系统按我们的要求运行，而不会发生故障或遭黑客入侵?

How can we grow our prosperity through automation while maintaining people’s resources and purpose?
如何通过自动化提升我们的繁荣程度，同时维持人类的资源和意志?

How can we update our legal systems to be more fair and efficient, to keep pace with AI, and to manage the risks associated with AI?
如何改进法制体系使其更公平和高效，能够跟得上人工智能的发展速度，并且能够控制人工智能带来的风险?

What set of values should AI be aligned with, and what legal and ethical status should it have?
人工智能应该归属于什么样的价值体系?它该具有何种法律和伦理地位?

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

科学与政策的联系：在人工智能研究者和政策制定者之间应该有建设性的、有益的交流。

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

科研文化：在人工智能研究者和开发者中应该培养一种合作、信任与透明的人文文化。

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

避免竞争：人工智能系统开发团队之间应该积极合作，以避免安全标准上的有机可乘。

Ethics and values 伦理和价值
6) Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

安全性：人工智能系统在它们整个运行过程中应该是安全和可靠的，而且其可应用性的和可行性应当接受验证。

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

故障透明性：如果一个人工智能系统造成了损害，那么造成损害的原因要能被确定。

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

司法透明性：任何自动系统参与的司法判决都应提供令人满意的司法解释以被相关领域的专家接受。

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.

责任：高级人工智能系统的设计者和建造者，是人工智能使用、误用和行为所产生的道德影响的参与者，有责任和机会去塑造那些道德影响。

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.

价值归属：高度自主的人工智能系统的设计，应该确保它们的目标和行为在整个运行中与人类的价值观相一致。

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.

人类价值观：人工智能系统应该被设计和操作，以使其和人类尊严、权力、自由和文化多样性的理想相一致。

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.

个人隐私：在给予人工智能系统以分析和使用数据的能力时，人们应该拥有权力去访问、管理和控制他们产生的数据。

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

自由和隐私：人工智能在个人数据上的应用不能充许无理由地剥夺人们真实的或人们能感受到的自由。

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

分享利益：人工智能科技应该惠及和服务尽可能多的人。

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

共同繁荣：由人工智能创造的经济繁荣应该被广泛地分享，惠及全人类。

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

人类控制：人类应该来选择如何和决定是否让人工智能系统去完成人类选择的目标。

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.

非颠覆：高级人工智能被授予的权力应该尊重和改进健康的社会所依赖的社会和公民秩序，而不是颠覆。

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

人工智能军备竞赛：致命的自动化武器的装备竞赛应该被避免。

Longer-term Issues 更长期的问题
19) Capability Caution: There being no consensus, we should avoid strong assumptions regarding upper limits on future AI capabilities.

能力警惕：我们应该避免关于未来人工智能能力上限的过高假设，但这一点还没有达成共识。

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.

重要性：高级人工智能能够代表地球生命历史的一个深刻变化，人类应该有相应的关切和资源来进行计划和管理。

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.

风险：人工智能系统造成的风险，特别是灾难性的或有关人类存亡的风险，必须有针对性地计划和努力减轻可预见的冲击。

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.

递归的自我提升：被设计成可以迅速提升质量和数量的方式进行递归自我升级或自我复制人工智能系统，必须受制于严格的安全和控制标准。

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.

公共利益：超级智能的开发是为了服务广泛认可的伦理观念，并且是为了全人类的利益而不是一个国家和组织的利益。