1. Record Nr. UNINA9911009159003321

Autore	Friedenberg Jay
Titolo	Artificial Intelligence and Universal Values
Pubbl/distr/stampa	Bradford : , : Ethics International Press Limited, , 2024 ©2024
ISBN	9781804416068 9781804416051
Edizione	[1st ed.]
Descrizione fisica	1 online resource (169 pages)
Soggetti	Artificial intelligence Ethics
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Intro -- Introduction -- The Intelligence Explosion -- Values and Ethics -- Chapter Previews -- Chapter 1 The Benefits and Risks of Intelligence -- Intelligence -- Artificial Intelligence -- Large Language Models -- Limitations of Current AI Systems -- Artificial General Intelligence -- Artificial Superintelligence -- AI Risks -- Examples of AI Risk -- AI Safety -- Conclusion -- Chapter 2 Artificial Intelligence and The Role of Values -- Types of Value -- Arguments and Counterarguments on the Concept of Value -- Value Convergence -- The Value Alignment Problem -- Value Alignment Methods -- Goals of Value Alignment -- Principles for Value Alignment -- Conclusion -- Chapter 3 Ethics and Artificial Intelligence -- Ethics and Morality -- Meta-ethics and AI -- Normative Ethics and AI -- Virtue Ethics and AI -- Applied Ethics -- Philosophy and Machine Ethics -- Taxonomies of Moral Agents -- Autonomous Moral Agents -- Real World Deployment of Autonomous Moral Agents -- Conclusion -- Chapter 4 Psychological Approaches to AI and Values -- Evolutionary Theories of Value -- Psychological Theories of Value -- Value Theories -- Literature Studies -- Survey Studies -- Conclusion -- Chapter 5 The Ecological Approach and AI -- Ecological Value and AI -- AI Agent Systems -- An Ecological Theory of AI Safety -- Networks -- Foodwebs -- Environments -- Evolution -- Populations -- Species -- Communities -- Species Interactions --

Species Invasion and Defense -- Species Competition -- Species Deletion and Coherence -- Human Societal Applications -- Conclusion -- Chapter 6 Ethical Problems in AI and Human Values -- Automation and Human Employment -- Military AI -- Autonomous Vehicles -- Smart Cities and AI -- Surveillance and AI -- AI and Legal Systems -- Healthcare and AI -- Teaching and AI -- AI, Creativity and Art -- AI and Social Relationships.
Human Values and Ethical Problems -- Codes and Standards -- The ART of AI -- Conclusion -- Chapter 7 Universal Values and AI -- The Case Against Universal Values -- The Case for Universal Values -- A Metatheory of Universal Value -- The Five Levels -- Problems with the Universal Values Approach -- How Advanced AI Might Create Better Ethics -- Supra-human Values -- The Future of Value -- Conclusion -- General Conclusion -- References.

Sommario/riassunto

The field of value alignment, or more broadly machine ethics, is becoming increasingly important as artificial intelligence developments accelerate. By 'alignment' we mean giving a generally intelligent software system the capability to act in ways that are beneficial, or at least minimally harmful, to humans. There are a large number of techniques that are being experimented with, but this work often fails to specify what values exactly we should be aligning. When making a decision, an agent is supposed to maximize the expected utility of its value function. Classically, this has been referred to as happiness, but this is just one of many things that people value. In order to resolve this issue, we need to determine a set of human values that represent humanity's interests. Although this problem might seem intractable, research shows that people of various cultures and religions actually share more in common than they realize. In this book we review world religions, moral philosophy and evolutionary psychology to elucidate a common set of shared values. We then show how these values can be used to address the alignment problem and conclude with problems and goals for future research. The key audience for this book will be researchers in the field of ethics and artificial intelligence who are interested in, or working on this problem. These people will come from various professions and include philosophers, computer programmers and psychologists, as the problem itself is multi-disciplinary.