

Book Review

Open Access

Artificial Intelligence; Safety and Security

Milad Savari¹

1- Department of Electrical and Computer Engineering, Faculty of Engineering, Kharazmi University, Karaj, Iran.

Abstract

Artificial Intelligence Safety and Security, edited by Roman V.Yampolskiy, examines the risks and challenges of developing safe AI systems. Divided into two sections, it first explores concerns from leading thinkers about ethical dilemmas, value alignment, and potential AI failures. The second section provides strategies and solutions, including adversarial learning, value alignment, and governance frameworks. This multidisciplinary work serves as a key resource for ensuring AI's safe and ethical integration into society.

Keywords Artificial Intelligence (AI), Value Alignment, Adversarial Machine Learning, Multidisciplinary Approach.

Introduction

The book *Artificial Intelligence Safety and Security* edited by Roman V.Yampolskiy is an important contribution to understanding and mitigating the risks associated with artificial intelligence. It is structured into two main parts:

1. Concerns of Luminaries: This section presents foundational works and key concerns from leading thinkers in AI, ethics, and security, including perspectives from Nick Bostrom, Ray Kurzweil, and Max Tegmark. These chapters discuss the potential risks of AI, such as malevolent AI, value misalignment, and the challenges of ensuring ethical behavior in intelligent systems.

2. Responses of Scholars: The second part offers theoretical and practical solutions to the risks outlined in Part I. It includes contributions from researchers across disciplines, addressing technical, social, and philosophical challenges in AI safety. Topics range from adversarial machine learning to the governance of superintelligence. The book effectively synthesizes a broad array of expertise and provides a deep dive into both the promise

and the perils of AI. It highlights critical failures of AI systems, ranging from biased algorithms to catastrophic failures in decision-making, and stresses the importance of preemptive safety measures.

This book is particularly recommended for researchers, policymakers, and technologists interested in the ethical development and deployment of AI. Its multi-disciplinary approach makes it a comprehensive resource for understanding the complexities of AI safety and security.

Book description

The book *Artificial Intelligence Safety and Security* edited by Roman V.Yampolskiy is the first comprehensive, multi-author volume addressing the AI control problem. It features two parts:

- **Concerns of Luminaries:** Contains seminal papers from philosophers, scientists, and technologists discussing existential risks and challenges in managing AI systems.
- **Responses of Scholars:** Includes theoretical and practical solutions, from opinion essays to technical approaches, tackling AI safety issues.

The book aims to serve as a foundational reference, initi-



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

* Correspondence:
miladsavari014@gmail.com

ating dialogue and research on the critical challenges of creating safe and secure AI systems.

Published as part of the Chapman & Hall/CRC Artificial Intelligence and Robotics Series, the book comprises 28 chapters organized into three distinct themes: security, artificial intelligence.

Number of Pages: 444 pages (based on the uploaded file).

Number of Chapters: 28 chapters.

Edition: First edition, published in 2018.

Language: The book is originally written in English.

Significance of the subject

The significance of the subject matter in Artificial Intelligence Safety and Security lies in its focus on addressing one of the most pressing challenges of our time: ensuring that advancements in AI are both safe and secure for humanity. The book underscores critical concerns and opportunities in this domain:

1. Preventing AI-Related Catastrophes

The book highlights the potential risks of AI, including accidental failures, value misalignment, and deliberate misuse. By examining historical failures and hypothetical scenarios, it underscores the importance of preemptive measures to prevent catastrophic outcomes, especially with the rise of Artificial General Intelligence (AGI).

2. Bridging Theory and Practice

Contributions from experts in fields ranging from philosophy to computer science ensure a comprehensive approach to AI safety. The book provides theoretical foundations as well as practical strategies for mitigating risks, making it a valuable resource for researchers, engineers, and policymakers.

3. Ethical and Societal Impact

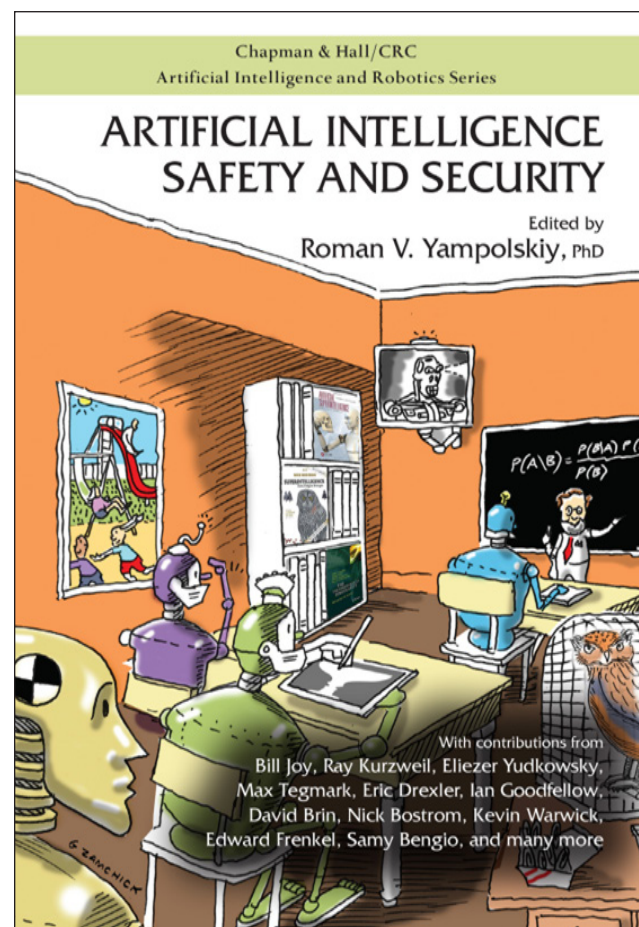
Ethical dilemmas, such as bias in algorithms and the societal consequences of AI decisions, are central to the book's discussion. It calls attention to the need for transparency, accountability, and value alignment to ensure AI serves humanity's best interests.

Readership

The difficulty of the chapters in Artificial Intelligence Safety and Security can vary widely depending on the background and expertise of the reader. Here's an assessment of factors that influence the accessibility of the book's content:

Difficulty Levels of Book Chapters

1. Foundational Theories and Ethical Discussions (Mod-



erate to Accessible) :

Chapters in the first section, Concerns of Luminaries, often focus on theoretical and philosophical perspectives, such as ethical dilemmas, AI value alignment, and existential risks. These chapters are generally accessible to readers with a basic understanding of AI concepts and an interest in ethical debates.

2. Technical and Mathematical Chapters (Challenging) :

The second section, Responses of Scholars, contains highly technical discussions, including machine learning models, adversarial AI, and formal methods for ensuring safety. These chapters may require a strong background in computer science, machine learning, or mathematics to fully comprehend.

3. Interdisciplinary Essays (Moderate) :

Chapters that blend technical content with societal, psychological, or philosophical insights are moderately challenging, appealing to readers with diverse interests. These essays explore broader implications without delving deeply into technical details.

Purpose of Reading :

For researchers or professionals, the book offers detailed

frameworks and technical solutions. For general readers, theoretical chapters and broad discussions provide thought-provoking insights but may feel dense without prior knowledge.

Structure and organization

The book *Artificial Intelligence Safety and Security* is structured in a clear and logical format to address both foundational concerns and practical solutions related to AI safety and security. Its organization ensures a comprehensive exploration of the topic while accommodating readers with varying levels of expertise and interests. Here's an overview:

Overall Structure

The book is divided into two main sections, with each focusing on different aspects of AI safety and security:

Part I : Concerns of Luminaries

This section includes 11 chapters authored by renowned thinkers, philosophers, and technologists.

Focus: It explores existential risks, ethical concerns, and philosophical dilemmas surrounding AI.

Topics include:

- The alignment problem (ensuring AI systems align with human values).
- Risks of malicious AI and unintended consequences.
- Broad societal and philosophical implications of advanced AI technologies.

Style : These chapters are often theoretical and accessible to a broader audience interested in ethics , policy, and the societal impact of AI. This section includes 11 chapters authored by renowned thinkers, philosophers, and technologists.

Focus: It explores existential risks, ethical concerns, and philosophical dilemmas surrounding AI.

Topics include:

- The alignment problem (ensuring AI systems align with human values).
- Risks of malicious AI and unintended consequences.
- Broad societal and philosophical implications of advanced AI technologies.

Style: These chapters are often theoretical and accessible to a broader audience interested in ethics , policy, and the societal impact of AI.

Part II: Responses of Scholars

- This section consists of 17 chapters, providing technical, theoretical, and practical solutions to the issues raised in Part I.

• Focus:

- o Strategies for mitigating AI risks, such as adversarial

machine learning, value alignment techniques, and safe system design.

o Governance frameworks, formal verification methods, and real-world examples of AI applications.

- Style: These chapters are more technical and intended for researchers, technologists, and professionals in AI, cybersecurity, and policy-making.

Introduction and Preface:

- Provides an overview of the AI safety landscape, outlining the goals and motivations behind the book.
- Frames the importance of the “AI control problem.”

Chapters by Experts:

- The chapters are written by leading figures in AI safety, ethics, and related fields, ensuring depth and authority.
- Each chapter is self-contained, allowing readers to explore topics independently.

Thematic Division:

- Theoretical and Philosophical: Found in Part I, addressing broad concerns and foundational issues.
- Technical and Practical: Found in Part II, offering detailed methodologies and solutions.

Notable Features

Multidisciplinary Approach:

- Contributions from experts in computer science, philosophy, ethics, and governance, making it relevant across disciplines.

Balance of Theory and Practice:

- The structure ensures readers gain a deep understanding of AI risks while exploring actionable solutions.

Accessibility:

- Chapters vary in complexity, catering to both general readers interested in AI ethics and professionals seeking technical insights.

Citations and References:

- Each chapter includes extensive references, making it a valuable resource for further research.

Novelty and Innovation

The book *Artificial Intelligence Safety and Security*, edited by Roman V.Yampolskiy , stands out for its novelty and innovation in addressing one of the most pressing challenges of the modern era: ensuring that AI technologies are safe, secure, and beneficial for humanity.

Conclusion

The book's structure is thoughtfully designed to guide readers from foundational concerns to advanced solu-

tions. By combining philosophical, ethical, and technical perspectives, it serves as both an introduction for newcomers and a detailed resource for experts, making it a cornerstone text in the field of AI safety and security.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: Jan. 2025 Accepted: Jan. 2025

Published online: Jan. 2025

DOI: 10.22034/ASAS.2025.496612.1072

References:

[1] Artificial Intelligence Safety and Security, Editor: Roman V. Yampolskiy

Submit your manuscript to Advances in the standards and applied sciences journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open Access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

**Submit your next manuscript at:
journal.standards.ac.ir**