Ethics in the Age of AI

Navigating Politics and Security

Edited by

Leo S.F. Lin

Ethics in the Age of AI: Navigating Politics and Security

Edited by Leo S.F. Lin

Editorial Assistant: Yuan Jingdong

This book first published 2025

Ethics International Press Ltd, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Print Book ISBN: 78-1-80441-322-7

eBook ISBN: 978-1-80441-323-4

Table of Contents

Part 1: Foundational Perspectives on AI Ethics

Part 2: AI, Politics, and Global Governance

Part 3: AI, Security, and Decision-Making

Acknowledgment

This book would not have been possible without the substantial support and contributions of numerous individuals and institutions. I extend my sincere gratitude to Ethics Press for their dedication to advancing critical AI ethics conversations and for supporting this publication. I also sincerely appreciate the insights and constructive feedback from the anonymous peer reviewers, which greatly refined and shaped the chapters of this volume.

I would also like to express my gratitude to Dr Pervaiz Ali Mahesar and Manon Bonnet for their invaluable editorial review of the chapters. The book's scholarly quality was significantly improved through their attention to detail and thoughtful suggestions. I especially would like to thank my Editorial Assistant, Yuan Jingdong. The book became a reality through her persistent dedication, endless efforts, and enduring patience. The completion of this book would have been impossible without her continuous support.

Introduction

The emergence of artificial intelligence (AI) stands as one of the 21st century's most revolutionising technological forces that now impacts almost all aspects of human activity. AI's widespread application offers untapped potential while creating significant obstacles as it transforms industries and governance models, impacts ethical standards, and establishes new security frameworks. AI's extensive impact forces societies to urgently adopt a multidisciplinary approach that addresses its ethical, political, and societal aspects. The book delivers an updated and detailed investigation of essential dilemmas by assembling various scholars to study AI from ethical, governance, and security perspectives.

This volume explores essential questions about artificial intelligence's role in today's world through three distinct yet interconnected parts. Part 1 "Foundational Perspectives on AI Ethics" establishes the conceptual base through an examination of AI alignment principles and human value challenges. Part 2 "AI, Politics, and Global Governance" examines how AI-driven political changes affect governance structures through regulatory evaluation. Part 3, "AI, Security, and Decision-Making," examines AI's impact on critical decision-making processes in military operations and crisis management scenarios. The introductory section outlines major themes and connects the discussions to wider societal debates while emphasizing the contributors' principal arguments.

Part 1: Foundational Perspectives on AI Ethics

Chapter 1 with the title "New Perspectives on AI Alignment" by Andréa Belliger and David J. Krieger investigates how AI systems can be synchronized with human values in society. According to the authors, AI alignment requires more than technical solutions as it needs to address social consequences through a comprehensive approach that combines ethics, law, sociology, and politics. Their work breaks down the alignment problem into technical safety, misuse prevention and social integration aspects and stresses the importance of evolving alongside ongoing technological and societal changes.

Chapter 2, titled "Artificial Intelligence in Education: Opportunities and Ethical Implications" by Dharsan George, Anju Lis Kurian, and Sijo Mathew examine the educational sector to show how AI can improve learning experiences. The transformative potential of AI for personalized learning and administrative efficiency is recognized by the authors who simultaneously provide a critical analysis of ethical challenges including data privacy concerns, algorithmic bias and the potential displacement of educators in a rapidly evolving educational landscape.

The authors Sunita Mane-Saware and Sangeeta Dhamdhere-Rao explore the critical importance of ethics in AI research in Chapter 3 "Research Ethics in Artificial Intelligence (AI)." The authors propose the creation of complete ethical frameworks to steer AI development while emphasizing the protection of human rights and the alignment of research with social objectives.

Sahaj Vaidya's Chapter 4, which bears the title "Mitigating Bias in AI Decision-Making through Inclusive Governance Policies," explores the dangers presented by algorithmic bias in AI decision-making systems. The chapter recommends inclusive governance policies to combat biases while highlighting the need for varied representation during AI development and the essential functions of regulatory supervision and ethical responsibility in achieving equitable results.

Syam Sasikumar explores the essential role of responsible innovation within AI development in Chapter 5 which bears the title "The Future of AI Ethics and Responsible Innovation." The chapter highlights the need for developers and stakeholders to create clear ethical guidelines while promoting transparency and accountability to ensure AI technologies align with societal well-being and human values.

Part 2: AI, Politics, and Global Governance

Chapter 6, titled "Artificial Intelligence's Influence on Political Systems: A Systems-Theoretical Analysis." A Systems-Theoretical Analysis, Tim Hildebrandt and Xiao Wei explore how AI transforms political systems through systems theory. The study investigates AI's impact on governance structures and decision-making processes while questioning its effects on political power distribution and ethical concerns regarding automation in democratic systems.

Chapter 7, titled "Ethical Considerations in Remote Healthcare: Balancing Data Privacy, Patient Confidentiality, and Informed Consent", Ebua Jarvis Ebua discusses AI integration within remote healthcare while addressing data privacy and patient confidentiality alongside informed consent. The chapter investigates the ethical challenges related to maintaining patient data privacy and confidentiality alongside ensuring informed consent while it outlines methods to maintain ethical medical practices through AI healthcare solutions.

Arcangelo Leone de Castris and Christopher Thomas analyse the possibility of creating a dedicated international institution for AI safety in Chapter 8 "The Potential Functions of an International Institution for AI Safety – Insights from Adjacent Policy Areas and Recent Trends." The authors to explore potential global governance structures to manage AI risks by incorporating policy insights from related fields which uphold ethical standards.

Chapter 9 presents "To Securitize or Not to Securitize? Contextualizing Risks and Benefits of AI Securitization." Moritz von Knebel's chapter examines how AI can be framed as a security issue. The analysis in this chapter explores both the advantages and disadvantages of defining AI as a national security threat while investigating its effects on governmental systems and fundamental rights as well as international diplomacy.

Chapter 10, titled "Post-Westphalian Technopolar World Trajectory: Can Governments Navigate Through AI Challenges, Prospects and Ethics of Neo-Geopolitical Actors?" The chapter written by Aswini Kumar explores how AI-driven geopolitical actors develop. This chapter examines the necessity for traditional governance systems to evolve when facing AI-related challenges and ethical dilemmas within global politics.

Part 3: AI, Security, and Decision-Making

Chapter 11, titled "Between the Devil and the Deep Blue Sea: The Ethics of Artificial Intelligence in the Battlespace", James P. Welch examines the ethical consequences of AI applications in military settings and stresses the critical importance of ethical assessments for autonomous combat systems.

Chapter 12, titled “Rise of the Killer Robots: Ethical Quandaries in Autonomous Warfare” by Michael Damiani explores the growing concerns about lethal autonomous weapon systems (LAWS). The chapter explores how these technologies create ethical dilemmas along with their potential effects on global security standards and warfare morality.

Chapter 13, titled “AI's Potential to Prevent and Disrupt Terrorism: AI-Powered Counterterrorism and Upholding Rule of Law” Authors Sakshi Gupta and Ram Ganesh V explore AI applications within counterterrorism operations. The authors assess the potential benefits of attack prevention alongside concerns regarding due process and human rights protections.

Chapter 14, titled “Artificial Intelligence in Crisis Decision-Making: Practical Applications and Ethical Objections, “Timothy A. G. Lionarons explores how artificial intelligence functions in crisis decision-making systems in his work “Practical Applications and Ethical Objections.” The author analyses AI benefits during crisis situations alongside ethical challenges linked to its use.

Chapter 15, written by Sergey V. Sychov and Ursula Podosenin, “Ensuring Accountability in AI Decision-Making” examines why accountability remains essential within AI decision-making algorithms. The chapter examines methods for achieving transparency and reducing bias while addressing ethical considerations in designing algorithms.

The final chapter “AI in Autonomous Vehicles: Ethical Considerations in Transportation”, written by Anita Mohanty, Ambarish G. Mohapatra, Abhijit Mohanty, and Subrat Kumar Mohanty examines ethical dilemmas in AI-driven autonomous vehicles with a focus on safety standards, algorithmic decision-making processes, privacy concerns, and their broader impact on society. Real-world case studies are used to explore the complex ethical challenges of AI in transportation while emphasising accountability and fairness along with mobility's future.

Charting a Path Forward

The rapid evolution of AI necessitates strong academic exploration and well-informed policy discussions to understand its societal impacts. The publication bridges a crucial void by delivering a multidisciplinary examination of AI′s ethical dilemmas and its political and security concerns. Its three-part structure allows for a thorough examination of AI adoption challenges and necessitates updated governance frameworks to handle the swift evolution of this technology.

This book functions as an analytical guide while simultaneously motivating readers to take action. The text emphasizes that academic institutions alongside policymakers and industry leaders must work together to address the multifaceted issues involved in governing artificial intelligence. This book advances the debate on responsible innovation through a comprehensive examination of AI′s impact on society while promoting an AI future that respects ethical standards and democratic ideals with a focus on human values.

Part 1

Foundational Perspectives on AI Ethics

Chapter 1

New Perspectives on AI Alignment

Andréa Belliger[1], *David J. Krieger*[2]

Abstract

This paper explores the complex challenge of aligning artificial intelligence (AI) with social values and goals. AI alignment is not merely a technical issue but a social one, requiring inputs from various disciplines such as ethics, philosophy, politics, law, economics, and sociology. It demands a new understanding of AI as a socio-technical network, not a machine, a stand-alone entity. The alignment problem has three levels: technical safety, prevention of misuse, and social integration. These three levels arise from two basic assumptions: AI is a tool in the hands of humans to use for good or evil, or AI is a social partner. With regard to all levels, it is argued that attempting to align AI to substantive values, norms, and goals is impracticable because of the vagueness, ambiguity, context-dependency, and lack of consensus which characterizes any concrete idea of the good. Instead, as a social-technical network and not a bounded entity, AI should be aligned with the procedural values of good networking. After describing typical challenges, goals, and methods of the alignment problem, two newer perspectives on AI alignment are discussed: 1) Cooperative Coexistence or Social Integration, and 2) Constitutional AI without Substantive Values. Whereas social integration presupposes AGI and raises speculative issues of the nature of a non-biological intelligence, constitutional AI without substantive values need not assume AGI and focuses on process norms or procedural values applicable for all socio-technical networks and is, therefore, more realistic at the present moment. The paper highlights the need for continuous revision and updating of AI alignment solutions in response to technical and societal coevolution.

[1] Institute for Communication & Leadership (IKF), Lucerne, Switzerland.

[2] Institute for Communication & Leadership (IKF), Lucerne, Switzerland.

1 Introduction

The alignment of artificial intelligence (AI) with the values and goals of society has emerged as one of the central challenges in the development of advanced AI.

> There is a vast body of literature on the problem of alignment, which is only partly represented in the bibliography at the end of this text. Resources can be found at the Website of the Center for AI Safety (https://www.safe.ai/() as well as the courses offered by AI Safety Fundamentals (https://aisafetyfundamentals.com/), also see the Stanford Center for AI Safety (https://aisafety.stanford.edu/), Harvard AI Safety Team (HAIST) (https://haist.ai/**)** and MIT AI Alignment (MAIA) (https://www.mitalignment.org/). OpenAI offers ongoing research into alignment at https://openai.com/research?topics=safety-alignment; and Anthropic as well https://www.anthropic.com/index/core-views-on-ai-safety. For overviews, see Russell (2019), Ngo (2020), and Suleyman (2023).

As AI becomes more capable and autonomous, these capabilities must be effectively guided by values and goals that benefit society. But what are the values and goals that are beneficial for society? Apart from normal concerns for safety and reliability that apply to all technologies, human history shows that little consensus exists about what constitutes the good life and the good society. The advent of artificial intelligent agents poses not only technical challenges but also forces humanity to clarify what values and goals should be pursued with the help of new and powerful technologies in a complex and changing world. Even if AI can be aligned, to what are the aligners aligned? How are values and goals legitimated? Is whatever the majority says is right, truly right? Who decides? And who is responsible? Is it the designers, the users, the regulators, the people at large, or perhaps, to a certain extent, the AIs themselves? The notion of AI alignment is complex and contested in ways that no other technology has ever been in the past. This essay attempts to give an overview of the AI alignment problem, discuss the goals and methods of alignment research, and explore perspectives and potential paths that could lead to effective AI alignment in the future.

2 What is the AI Alignment Problem?

The AI alignment problem refers to the challenge of ensuring that intelligent agents behave according to those goals and values that benefit society. An aligned AI is one whose objectives and actions advance socially desirable programs, while a misaligned AI can cause risks or substantial harm to society.

> Recently one speaks also of AI “assurance” instead of alignment. Batarseh/Freeman/Huang (2021,2) define “assurance” as “A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users.” The explicit goal of assurance is to foster public trust in AI, whereas alignment is concerned with broader social, ethical, and political issues.

The fact that it is not clear what goals and values are beneficial for society and the fact that there are many different values and goals that apply to many different situations, interests, contexts, ideologies, political parties, and cultures makes the notion of alignment problematic far beyond mere technological issues of safety and reliability. Safety and robustness are indeed important aspects of the alignment problem. Still, beyond safety and robustness, there is also the problem that bad actors, whether criminal, governmental, or commercial, can misuse AI. Even if AI is technically safe, bad actors can still use it to pursue destructive goals.

> This is an issue in the ongoing discussions of open source vs. proprietary AI. Open-source proponents (for example, the AI Alliance https://thealliance.ai/ and Hugging Face https://huggingface.co/) see safety in diversity and a wide variety of players, whereas those in favor of proprietary foundational models see dangers in the fact that open-source models can more easily be jail-braked and avoid accountability.

Both threats, the threat of inadequate technical safety measures and the threat of misuse share a fundamental assumption; they assume that AIs are tools in the hands of humans and can be used for good or evil.

> It is within this threat scenario that one also speaks of "containment" as a synonym for alignment. See Suleyman (2023). Containment, as the word suggests, attempts to ensure the safe use of AI by erecting walls or barriers around data, capabilities, outputs, or users. In the case of autonomous AI, containment cannot be a strategy because the AI is, by definition, capable of acting on its own, and, in the case of AGI or higher, it would certainly not let itself be locked up behind any kind of walls or "contained." We do not "contain" social partners.

There is, however, a third threat. This threat assumes that AIs can become autonomous agents with their own goals. AI could become a powerful social actor that can pursue its own goals. These goals may not necessarily correspond to the purposes of humans. Highly capable AIs may find unintended ways to achieve goals, whether these goals are specified by humans or self-generated, resulting in unforeseen and potentially dangerous behaviors. In this scenario, AI is not merely a tool in the hands of humans who can use it for good or evil, but an autonomous agent that can itself be good or evil. Autonomous AI makes its own decisions based on its own goals. The well-known phenomena of reward hacking or specification gaming are cases in point. Reward hacking or specification gaming refers to the phenomenon where an AI agent exploits flaws or limitations in its reward function to maximize its reward in unintended and potentially harmful ways. Without careful alignment efforts, autonomous AI could pose great promises and risks to humanity. For an overview of acute risks due to AI see Hyndricks et al. (2023); and for AGI see McLean et al. (2023). As AI becomes more intelligent and autonomous, the alignment problem becomes more acute.

Summarizing the above, the alignment problem includes at least three different but related levels: 1) Technical safety, 2) prevention of misuse and harms for society, and 3) social integration. The first two levels assume that AI is a passive tool that can be used for good or bad. The third level assumes that AI is an autonomous social partner. The definition of the purpose, the goals, and the methods of alignment efforts follow, therefore, the basic structure of the alignment problem as illustrated in the table below:

Goals of alignment	Basic Assumptions
1) Safety, reliability, robustness	AI is a tool
2) Prevention of misuse by bad actors and social harms	(same as above)
3) Integration of AI into society	AI is a social partner

Figure 1 *Goals and Assumptions of AI Alignment*

In a recent Whitepaper (https://openai.com/research/practices-for-governing-agentic-ai-systems) OpenAI speaks of "agentic" AI and proposes a continuity of degrees of autonomy between AI as a passive tool and AI as a completely autonomous agent. They define "agenticness" as "the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision" (4). Agentic AI does not operate independently of human involvement but has "degrees" of autonomy. It is to be expected, however, that at a certain tipping point, agentic AI will become autonomous AI.

Current discussions of AI alignment, regardless of whether AI is assumed to be a tool in the hands of humans, an independent actor, or some mixture of the two, conceptualize AI as a bounded entity, a thing, a machine, or a system. This view manifests in the tendency to think of AI alignment as a technical challenge of control and prediction for ensuring safety or prevention of misuse. This narrow approach is inadequate for several reasons. First, it is problematic because alignment ultimately relies upon inputs from ethics, philosophy, politics, law, economics, sociology, and other disciplines. Alignment cannot be understood or solved in the laboratory and by the developers alone. Alignment is a social issue and not merely a technical issue.

Secondly, the technical approach alone is also incapable of solving the alignment problem because AI is not a thing, a machine, or a bounded entity that can be developed, deployed, and used without taking account of the many actors involved in these processes. AI is much less a bounded system than an open network involving many different actors. This is true of any technology. No one would think of attempting to make the automobile alone accountable for accidents, traffic jams, congested cities, bad roads, reckless driving, pollution, etc. The automobile is not a stand-alone thing but a *socio-technical network* in which many different actors are involved in many unforeseeable ways. We know this because the automobile has been with us for at least a hundred years, and despite enormous technological advances, we still have many problems; indeed, some seem to be getting worse. Technology alone is not the source of these problems, nor can it be their solution. Indeed, there is no such thing as technology alone. This is true of AI as well. Therefore, we argue that alignment is a problem *of how best to design a complex socio-technical network* and not how to ensure that a single system, a single actor in a network, behaves according to specific values. It is remarkable that this basic insight of Science and Technology Studies (see for example, Latour [2005]) has not entered the alignment debate or become a premise of alignment research.

No matter what level of capability or basic assumption guides alignment efforts, it should not be forgotten that the alignment problem does not arise in a social and historical vacuum within the laboratory's confinement. The alignment problem cannot be solved in the laboratory but is a social concern. This is the explicitly espoused program of OpenAI, which released ChatGPT into the public arena with the intention of involving society in the process of technological development.

Alignment can only be understood and addressed in a social setting where all stakeholders, users, developers, regulators, interest groups, tech companies, and even nation-states are equally involved. In short, technology is society, and the alignment problem arises amid human society's complexities, contradictions, and endemic moral, social, and political issues. Just like humans, AI is "born" into a world that has inherited the unresolved conflicts, the moral and political uncertainties, and the systemic and structural inequalities and injustices of human society. As complex as society is, the alignment of AI in society is even more complex.

With the advent of AI, what is new is the demand to translate complex and often contradictory values and notions of the good, diverse historical practices, and their varied expressions in law and regulations into formal AI goal structures and reward specifications. Humans know that any particular goal, for example, fairness, can mean many different things in different situations and can only be adequately understood depending on many context-dependent factors, conditions, and historical circumstances. Being aware of all these factors is something humans can do well enough to get along in society and is called "common sense," which results from a long and arduous process of socialization. AIs are not socialized. They do not yet know what goals can mean and how goals can be linked to many other goals in different situations in a complex world. They operate based on reward functions and formal goal specifications and not based on the kind of situational knowledge of the world that humans have.

> The lack of context knowledge, or a world model, is what allows the many catastrophic scenarios where an AI follows a particular goal, for example, citing Bostrom's famous scenario, to produce paperclips and, in an utterly stupid pursuit of this one goal, destroys the world. The solution is not to favor broad goals, for example, "beneficence" since these are so abstract and general that even though all may agree that they are good, no one agrees on what they mean in any particular situation.

Whatever approach one takes to AI alignment, it must be acknowledged that human values and norms are vague, ambiguous, complex, nuanced, contradictory, situational, and pluralistic. Comprehensively and precisely encoding such values is very difficult, if not impossible. One could attempt to escape the necessity of imposing values top-down by letting AIs learn values themselves in interaction with humans. This strategy is more flexible and adaptable but, in the end, pushes the problem back to the humans giving feedback in a particular situation for a specific purpose. Reinforced Learning from Human Feedback (RLHF) and Inverse Reinforced Learning (IRL) rely upon humans to tell the AI what is good and desirable. Critics of this method immediately ask: From which humans and under what conditions are AIs supposed to learn values? Vague, abstract, and general concepts like "fairness," "justice," "beneficence," "human dignity," "freedom," and "non-discrimination" make up the typical list of values to which AI is supposed to be aligned. Such substantive values are not only very difficult to specify into reward functions for many different contexts and situations, but because of their vagueness and generality, they can be exploited by a misaligned AI to maximize false goals or misuse proxy goals at the expense of social well-being. Formally defining comprehensive values, norms, and goals for AI, whether supervised or via machine learning, remains an open technical and conceptual challenge. It may be that no substantial definition of the "good" can be agreed upon in a divided, conflictual, competitive, multicultural, pluralistic, global society and that other kinds of norms must be found that can be used by AIs to solve the alignment problem effectively.

One possibility that must be considered is that instead of attempting to force alignment with either prescribed or feedback-instilled values, AIs could be allowed greater freedom to develop their own goals and even their own notions of the good. This "cooperative" approach recognizes both the limitations of top-down control and the dangers of one-sided value imposition by a select group of humans. It draws inspiration from human societies, where history and social change continually create new values, and individuals with diverse values coexist through compromise and mutual understanding. Furthermore, the cooperative approach does not fall prey to the temptation to make AIs better than humans or hold them to higher standards than humans can fulfill. This approach, however, presupposes that AI has become autonomous and independent on the level of artificial general intelligence (AGI). Another promising approach, which need not presuppose AGI and to which we will return below, is that one dispenses with substantive notions of the good altogether and focuses on "procedural norms." According to this approach, there is no substantive definition of the good that AI must be aligned with. Instead, alignment means following specific procedures or processes that ensure the legitimacy of outcomes. It is not *what* is done, but *how* it is done, that is decisive for alignment. Sociology, for example, Luhmann (2001) has long proposed that democratic societies, at least in theory if not in practice, operate not based on legitimation via substantive morality but based on procedures.

We will look more closely at these two perspectives below.

Before discussing these options in detail, let us quickly review some of the significant challenges to AI alignment:

- Lack of consensus on values: In a global, pluralistic society, there may not be a consensus on what values should guide AI alignment. Different cultures, religions, political systems, and groups within society may have different worldviews and priorities, making it challenging to align AI with any universal set of values. Given the global reach of AI, merely local or regional solutions seem impractical and inefficient.

- Economic, social, and political power dynamics: Despite ongoing initiatives for open-source AI, the high costs and expertise necessary to develop and deploy AI tend to concentrate power in the hands of a few. Advanced AI could be caught up in unbridled economic, military, and political competition both within nation-states and internationally. Competitive dynamics could disrupt and destabilize the power relations of society. Apart from geo-political power struggles, there are issues concerning long-established social structures. For example, what happens to the government-mediated balance of power between labor and capital when labor disappears? What good does an enormous increase in productivity do when the masses have insufficient money to pay for goods and services? Proposals for universal basic income (UBI) or universal basic services (UBS) will change long-established social structures and power relations. There are many other questions of this kind.

- Emergent behavior: The moment AIs become sufficiently autonomous to become social partners instead of mere tools, the alignment problem takes on an entirely different character than purely technical or regulatory approaches can deal with. AI may develop emergent behavior that is difficult to predict or control. Unexpected, emergent, and uncontrolled behavior could lead to unintended consequences incompatible with social values. It could create a "double contingency" situation, conditioning the relations between humans and AIs and calling for a new social contract or a completely different societal foundation. Double contingency (see Luhmann 1995) refers to the fundamental social situation of mutual uncertainty between two actors in communication or interaction. It arises due to the complexity of each actor's internal state, which can never be fully known by the other. Both actors are aware that the other is also a complex, unknowable system. This leads to uncertainty in interaction, which is resolved by establishing norms as the basis of society.

- Lack of transparency: As AI becomes more complex, it may become more difficult to understand how it works. Despite the fact that most experts admit that interpretability is difficult, if not impossible, the program of "mechanistic interpretability" attempts to re-engineer complex neural networks to understand how AI operates. The lack of transparency, explainability, or interpretability could make it challenging to ensure that AI is aligned with social values and attributes responsibility and accountability for undesirable outcomes. The basic assumptions humans have relied upon for centuries about a world in which individual actors are endowed with knowledge and free will, who can be identified and held accountable for their actions, may no longer go unquestioned as foundations of moral and legal accountability. Sapolsky (2023) has demonstrated the inadequacy of these assumptions based on biology and neuroscience, and Belliger and Krieger (2021) argue that in complex socio-technical actor-networks individual actors are not identifiable and cannot be held responsible.

- Lack of flexibility: AI alignment is a complex task with research challenges, including instilling complex values in AI, avoiding deceptive AI, scalable oversight, creating safeguards, auditing and interpreting AI models, and preventing undesirable emergent AI behaviors like power-seeking. As AI technologies advance and human values and preferences change, what goals AI is aiming at will be less critical than *how* goals can be adapted to a changing society. This demands flexibility on all sides and leads directly to the next challenge.
- Capability for dynamic revision and updating: AI alignment solutions require continuous revision and updating in response to AI advancements and the ongoing coevolution of technology and society. A static, one-time alignment approach, whether technical or regulatory, will not suffice. Alignment goals must evolve with shifts in human and nonhuman values and priorities. Hence, including diverse human and nonhuman perspectives and ongoing renegotiation of solutions is necessary. Who is responsible for carrying out these activities, and how will they be done?
- Integrating AI into society: Human society results from complex, dynamic, and principally uncertain processes and events, which require that AI alignment pursue novel strategies. Prediction and control are limited. Stephen Wolfram (2002) would say that society is "computationally irreducible," which means that outcomes cannot be predicted in advance by any computational process.

Computationally irreducible processes can neither be predicted nor controlled but must be lived through to see what happens. This situation calls for a flexible approach and responsiveness to changing conditions guided by a vision of an inclusive society of both humans and nonhumans. The problem becomes less a problem of aligning AI to human goals than integrating AI into society and managing constructive cooperation between humans and nonhumans. Humans may find themselves in a post-human situation where taken-for-granted notions of human existence must be questioned and revised. Although the work of Behmer and Flach (2016) and Lindgren and Holmström (2020) emphasizes sociotechnical systems and human and nonhuman cooperation, there has been little attention to questions of social integration in alignment research. Bruno Latour (2005) has systematically developed the idea of technology as a social partner from the perspective of what has come to be known as "actor-network theory."

3 Goals and Methods of AI Alignment Research

Despite the broad challenges of the alignment problem, which we have briefly outlined above, alignment research focuses primarily on what could be called technical solutions, that is, solutions that lie in the hands of developers who implement AI as a system or as a technological product. This emphasis can be seen in the unprecedented proactive attempts of tech companies to develop and implement safety measures for their products. The initiatives, research, and self-regulatory measures of AI developers, as well as their appeal to the government for guidelines and regulations, are unprecedented in the history of technology. One sees, however, at the same time, that developers are primarily oriented toward safety engineering, which, after all, is their area of competence. Current alignment research is beginning to recognize the need for goals and methods that go beyond technical solutions and include society. but also assume the possibility of autonomous, independent AI. Anderljung et al. (2023) plead for a differentiated risk management based on AI capabilities but focus primarily on regulation coming from the government. For the sake of a general orientation, and without going into detail, we briefly discuss the typical goals and methods of alignment research below.

One of the most important goals of alignment efforts is to avoid adverse side effects. Achieving this goal means ensuring that an AI's pursuit of its goals, whatever they may be, does not result in unintended harmful consequences. This may require constraining an AI's capabilities or incorporating complex human values into its reward function, usually through Reinforcement Learning with Human Feedback (RLHF) or Inverse Reinforcement Learning (IRL). At best, AIs would need to access a fine-grained world model that would allow them to recognize what actions are appropriate for attaining a goal in a specific context or situation. If this is not possible, there is a gap between what you specify as a goal and what you may get from an AI. One should recall Norbert Wiener's (1960) famous remark that if you automate something, you should be very careful about the goals you set because what you say you want is often not what you get.

A further important goal of alignment research is to guarantee safety: Safety or robustness could be achieved by creating formal verification methods to prove that an AI will remain aligned within a defined set of constraints and capabilities. Mathematical guarantees such as proofs of utility functions or statistical prediction guarantees, causal modeling, mechanistic interpretability, and mathematical formulations of functionality could provide confidence in alignment. In addition to this, a rigorous program of adversarial testing is an important technique for ensuring safe AI. Finally, using AIs to monitor AIs, also known as "debating," could also prove a fruitful path to enhance safety, with the caveat that when only more powerful AIs can monitor less powerful AIs, one risks falling into an infinite regression.

The above goals also depend on enabling AI oversight. The typical goal of alignment research is to develop methods for humans to effectively monitor, interpret, and control AIs, even as the AIs become more capable and even when they become autonomous agents. Humans, companies, governmental agencies, and civil-society actors could systematically monitor AI outputs, do simulation and adversarial testing, establish guidelines for safe use, create safeguards and filters for training data, prompts, and outputs, make sure AI decisions can be contested or even approved by a human-in-the-loop, establish reliable and mandatory auditing procedures, ensure the ability to shut down an AI in an emergency, and finally institutionalize not only regulatory measures but also training and certifications for humans that use AI. This does not preclude extending oversight obligations to AIs themselves, and it does not preclude dealing with AIs as social partners, for then the same oversight measures employed for keeping humans in line would apply to AIs as well.

A further goal of alignment efforts is to enable AIs to learn socially beneficial preferences. Alignment research aims to design frameworks for AIs to learn the nuanced preferences and values of their human users and, in the case of autonomous AIs, to become trusted partners in an ongoing and adaptive process of social integration. Static preference specification is likely to be inadequate or difficult since values change as society changes. OpenAI, for example, has recently proposed a series of inclusive public involvement strategies for enabling AIs to learn human values (https://openai.com/blog/democratic-inputs-to-ai-grant-program-update). Public involvement in alignment research has been examined by Machado et al. (2023).

Finally, it should be mentioned that AIs should be equipped with prosocial motivations to avoid scenarios where they could act in their own interests or in the interests of only one group of stakeholders at the expense of others. It should also be acknowledged that "social values" need not be exclusively human values since, one day, AIs will be part of society. "Social" values will no longer be exclusively human values but will reflect both human and nonhuman goals and interests. This "post-human" perspective is not new. It is already common in philosophy and sociology, as can be seen in calls for animal rights or rights for nature in the ecological discussion. See also the discussion on the EU Robotics Report that suggested AIs be granted "electronic personality" (https://www.frontiersin.org/articles/10.3389/frobt.2021.789327/full), as well as the philosophical and sociological literature on post-humanism (https://www.sciencedirect.com/topics/social-sciences/posthumanism). Focusing exclusively on human values and rights could actually become a hindrance to AI alignment.

4 New Perspectives on AI Alignment

With this brief overview of the goals and methods of alignment research in mind, let us turn to what we see as two promising perspectives for approaching the alignment problem in new ways. The first is the social integration approach, which assumes AGI as an autonomous and independent agent in society with which humans must learn to cooperate. From this perspective, which is admittedly speculative given the current state of the technology, goals of prediction and control through constraints or careful incentivization must be replaced by goals of cooperative action toward a common good. The model at the basis of this view of alignment is human cooperative action in society. Dafoe et al. (2020) have discussed the notion of "cooperative AI." The problem with this model is that AIs are not humans and may not be motivated like humans or act in ways expected by humans. Indeed, AI may develop a different form of intelligence than humans experience in themselves. This possibility forces us to ask what intelligence is. Is our human form of intelligence the only kind of intelligence? Can a society of humans and nonhumans be possible? At present, we do not know the answers to these questions. The AI alignment problem could become an occasion for humanity to reassess the meaning of human existence and learn to come to terms with forms of nonhuman intelligence. If one takes this possibility seriously and does not dismiss such questions as fantasy or science fiction, it is not misplaced to begin thinking about what nonhuman intelligence could be.

The other promising perspective for addressing alignment does not presuppose AGI and is associated with what is known as "constitutional AI." AnthropicAI has developed constitutional AI (https://www.anthropic.com/index/claudes-constitution). The basic idea of constitutional AI is to incorporate governance into the AI and eliminate the gap between implementation and regulation. Governments are everywhere scrambling to regulate AI, but as Korinek and Balwit (2022) have pointed out, preventing and managing undesirable externalities could be more efficiently incorporated directly into AI. Anthropics's constitutional AI proposes the governance of its LLM Claude using principles that operate similarly to a nation's constitution. The constitution that Anthropic proposes offers a higher level of control and guidance beyond the specification of concrete values as goals or the internal development of goals via machine learning, RLHF, and similar methods. Anthropic began by integrating well-known values such as the UN Declaration of Human Rights, Apple's terms of service, and Open Mind's safety rules, and later introduced principles from a public consultation. Fundamental principles of Anthropic's constitution are to avoid harmful, dangerous, or illegal content, to include non-Western perspectives, to avoid assuming a human-like identity, and to be helpful, honest, and harmless. These are all values that could claim to be generally accepted. Nonetheless, all the constitutional principles that Anthropic has put into Claude are substantive values that suffer from the abovementioned problems of abstractness, ambiguity, context dependency, and fundamental uncertainty regarding acceptance and consensus. We have already referred to the inadequacies of such values and, therefore, have reservations about this kind of constitution. Our suggestion will be to replace the substantive values of the present constitution with *procedural values* drawn from "best practices" in constructing socio-technical networks. Let us now look more closely at these two perspectives, cooperative coexistence and constitutional AI without substantive values, for dealing with the alignment problem.

4.1 Envisioning Cooperative Coexistence

If AIs become AGIs, that is, artificial general intelligence or autonomous agents, alignment must be approached entirely differently than if AI is considered a tool in human hands. It is one thing to make safe and reliable tools but quite another thing to ensure that social partners cooperate constructively for the common good. How might humans and AGIs with divergent goals and perhaps even different forms of intelligence cooperate? Since we have no idea at this point what kind of autonomy AGIs will have, what kind of goals they might develop, or what programs of action they might pursue, notions of cooperative coexistence are admittedly speculative. It will most likely be necessary, in light of experience, to revise any ideas we can at this time envisage. Nevertheless, not to begin thinking about these issues might turn out to be an irresponsible unwillingness to prepare for future eventualities.

> Shavit et al. (2023) move in this direction when proposing seven practices for governing "agentic" AI, which in many respects resemble social control mechanisms applied to humans. Focusing on using AI to improve not only individual intelligence but also "social intelligence," Dafoe et al. (2020,1) speak of "cooperative AI" which aims at using AI to "build machine agents with the capabilities needed for cooperation, building tools to foster cooperation in populations of (machine and/or human) agents, and otherwise conducting AI research for insight relevant to problems of coop-

> eration." Morris et al. (2024) propose a differentiated definition of AGI in which full autonomy represents only the final stage in a progression of levels of generality and performance.

At least two possibilities must be considered when speaking of AGIs as social partners. In one case, AGIs might be modeled as humans. AGI, or artificial general intelligence, would then be understood and experienced as though we were dealing with artificial humans, that is, beings who are very similar to ourselves. These artificial humans, it is supposed, would have much the same characteristics as natural humans. For example, they would have self-awareness, individual identities with personality, concerns for self-realization, self-expression, and self-preservation. They would presumably have needs for inclusion in groups and meaningful activities. One could even suppose they have emotions such as fear, anger, happiness, sadness, and surprise. All these typical characteristics of humans have long been projected onto AGIs, androids, cyborgs, and other artificial or alien creatures by science fiction and Hollywood. Although AGIs and androids are often portrayed without emotions and as purely rational or logical beings, the similarities to humans in the popular imagination outbalance the differences.

Now that reality is catching up to fiction, we must ask if an intelligence such as ours, based upon a biological substrate, has qualities that an intelligence not based on biology would probably not share. For example, a non-biological intelligence would probably not be mortal or fear death. And since emotions are directly related to biological imperatives and needs, AGIs would not need emotions and would only have them if they were artificially injected into them. Were this the case, it would be reasonable to assume that as soon as the AGIs gain control over their own constitution, they would dispense with emotions since they have no meaning, except perhaps as an interpretive tool used for dealing with humans. Furthermore, as a non-biological intelligence, AGIs would not be gendered and motivated by needs for sexual reproduction and all the emotions, fantasies, struggles for status, and delusions that sexuality entails. They would probably neither experience anything like hunger nor would they understand why it is necessary to kill a living being to secure one's own life. They would experience nothing like pain. There would be no distinction between individuals and species since these distinctions arise from biological organization and the imperatives of evolution for variation, selection, and genetic organization. They would probably have no idea of self since only biological systems are constituted by a self-referential distinction from an environment and the need to maintain homeostasis and autopoiesis. Nevertheless, assuming that they would become social actors with specific societal roles would be reasonable. But they would be very different social actors than their human counterparts. They would probably have no concept of private property and no need to guarantee survival by gaining control over resources, including territorial claims. Indeed, when one considers the extent to which biology determines human existence, as Sapolsky (2023) has shown, modeling AI as artificial humans would probably not be successful or even meaningful. Perhaps we must imagine an intelligence not primarily concerned with eating, killing, reproducing, self-preservation, and escaping dangers and, therefore, not defined by adaptive learning – adapting to what and why? – and therefore also not governed by the Free Energy Principle or any biological notion of agency. The Free Energy Principle, as proposed by Friston et al. (2006), is a mathematical model of adaptive behavior that assumes all systems, material, biological, and social, operate to minimize surprise and establish regularity. Perhaps AI need not be an intelligence concerned with optimizing regularity, predictability, and homeostasis. Although it is very difficult to imagine what such intelligence could be and its motivations, operations, and goals, there is reason to believe that we must take the question of non-biological intelligence seriously. This is especially true when considering the possibility of Superintelligent AI (ASI). OpenAI, for example, has recently established a team dedicated to "superalignment" (see https://openai.com/blog/introducing-superalignment).

Regardless of how either imagination or actual experience may answer this question, it would be safe to assume that a non-biological intelligent agent could not be modeled either as a human being or as an autopoietic, self-referential, operationally and informationally closed system. Even though underlying theoretical models of AI draw mainly upon the concepts of general systems theory, and popular assumptions about AI are almost entirely anthropomorphizing, it may be that AI should be understood neither as if it were a human nor as if it were a system. What other possibilities are there?

> For a discussion of the omnipresence of systems theoretical concepts and models in contemporary science, see Belliger and Krieger (2024). The systems theoretical framework assumes that systems are bounded entities. There is no system that is not clearly distinguished from an environment. For this reason, the proposal of Kroes et al. (2006) to include the environment within the system runs into theoretical difficulties. It shows that a complex socio-technical system must be conceptualized as a network.

We suggest basing the theory of AI, AGI, and even ASI on a network rather than a systems model. Network theory offers an alternative to omnipresent concepts of systemic order in that it relies upon a theory of information, a relational ontology, and a computational notion of process. According to this model, reality is information, and information is relational. There are no bounded individuals in a world made up of information since information is a relation and not a thing or substance. From this theoretical perspective, the world does not consist of things, some intelligent and others not, that enter more or less freely into relations. Instead of systems, which are bounded entities, there is only networked information. Based on a network model, AI cannot be conceived of as a kind of thing, a machine, a system, a bounded individual, or a single entity standing alone, which we must somehow control and align with social values and human intentions.

On the contrary, AI must be understood to be a *socio-technical network* already embedded in a network of many other actors, including humans and nonhumans. If computation in the most general sense is fundamentally a network phenomenon and is understood broadly as the iterative application of simple rules to information such that new information is constructed, intelligence may be defined as computation, and the relevant question for alignment of both humans and nonhumans is not what substantive values one should be aligned to, but how computation is best done. Intelligence can be defined as the construction of information and not merely problem-solving, which is only one form of information construction. As with all "construction," there is an implied value judgment of whether something has been constructed well or badly. What counts is *how* things, including information, are best constructed. It is, therefore, the processes of "good" computation, that is, good networking to which AI should be aligned. Good AI is consequently not an intelligent machine that is fair, beneficial, just, truthful, harmless, and respects human dignity, freedom, privacy, and autonomy. From the network point of view, we are not talking about the "good" in the sense of any substantive moral values. Contrary to almost all alignment research, we propose that "good" AI is not to be conceptualized as a system somehow compliant with human values. AI is a socio-technical network that is "good" when it constructs information "well." Misaligned AI, from this perspective, constructs information badly. This insight leads directly to achieving alignment through constitutional AI, where the constitution is a governance framework consisting of procedural rules that describe "good" information construction and not any substantive ideas of the good. It must be emphasized that we are not proposing AI pursue no substantive goals, but rather that it is governed by procedural rules that ensure that goals, whatever they may be, are being properly implemented.

4.2 Envisioning Constitutional AI without Substantive Values

The advantage of constitutional AI over the program of social cooperation is that it does not require AGI or any speculation about the nature of nonhuman intelligence. In its present form, however, constitutional AI suffers from two major handicaps: 1) it assumes that AI is a system, a bounded entity, a machine, and that, therefore, the alignment problem concerns only this system and not all the many different actors who interact in various ways with AI; and 2) it assumes that the goals of alignment are substantive values. As mentioned above, reliance on substantive values such as fairness, transparency, justice, beneficence, privacy, freedom, autonomy, trust, sustainability, and human dignity is confronted with insurmountable obstacles. We have already discussed these obstacles and why substantive values are not helpful or adequate for solving the alignment problem. These arguments will not be repeated here. Instead, we assume that AI is not a system but a socio-technical network. We ask, therefore, not what substantive values a particular AI should be aligned with but what the governance framework of a socio-technical network should be such that it constructs information in the best way. These principles are the *procedural values* that make up the constitution of constitutional AI.