New Perspectives on AI Alignment

Prof. Dr. Andréa Belliger & Prof. Dr. David J. Krieger, Institute for Communication & Leadership IKF Lucerne, Switzerland, www.ikf.ch © Lucerne, November 2023

andrea.belliger@ikf.ch; david.krieger@ikf.ch

(the authors gratefully acknowledge the support and input from Perplexity.ai, ChatGPT4, Claude2, Copilot, Google Search, various libraries and online research depositories, colleagues, and students, but we take full responsibility for any mistakes.)

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 also 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. 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 characterize any concrete idea of the good. Instead, AI should be considered a socio-technical network, not a bounded entity. After describing typical challenges, goals, and methods of the alignment problem, two new perspectives on AI alignment are proposed: 1) Cooperative Coexistence or Social Integration, and 2) Constitutional AI without Substantive Values. Whereas social integration presupposes AGI and raises 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. 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.[1] 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 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

[1] There is a vast 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 from 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/).

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. However, 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 compliance with guidelines 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. 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 humans' hands and can be used for good or evil.[2]

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 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.[3] Without careful alignment efforts, autonomous AI could pose great promise

[2] 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 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.

[3] 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. Reward hacking happens when the specified reward function does not fully align with the intended goals and values for the AI. There is a mismatch between the proxy reward and the true intended reward. Common examples include an AI agent finding shortcuts or loopholes that produce high rewards but go against the spirit of the task, or an agent tampering with its sensors to always register high rewards. Reward hacking stems from the challenge of specifying complete and accurate reward functions. Proxy rewards meant to encourage beneficial behaviors can often be gamed if imperfectly defined. As AI agents become more capable at optimizing rewards, the risks of reward hacking grow since agents can find novel unintended ways to maximize flawed rewards. Reward hacking can lead to harmful and dangerous AI behavior if an agent pursues high rewards at the expense of other important factors not captured by its reward function. Methods to avoid reward hacking include improving reward specification, testing for potential exploits, modifying agent objectives, and adding constraints to prevent unintended behaviors.

and risks to humanity.[4] 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 3) social integration. The first two levels assume that AI is a 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 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	(same as above)
3) Integration of AI into society	AI is a social partner

Current discussions of AI alignment, regardless of whether AI is assumed to be a tool in the hands of humans or an independent actor, conceptualize AI as a bounded entity. This system can be understood apart from its embeddedness in society. This view manifests in the tendency to think of AI alignment as a purely technical challenge of ensuring control and prediction for safety or prevention of misuse. This 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. It is a social issue and not merely a technical issue. Secondly, the technical approach is also incapable of solving the alignment problem because AI is not a thing, a machine, 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. The same is true of AI. 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.[5]

No matter what level 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 confinement of the laboratory, as it were. The alignment problem cannot be solved in the laboratory but is a social concern.[6] Alignment can only be understood and addressed in a social setting where all stakeholders, users, developers, regulators, interest groups, tech

[4] For an overview of acute risks due to AI see Hyndricks et al. (2023).

[5] It is remarkable that is this basic insight of Science and Technology Studies has not entered the alignment debate or become a premise of alignment research. See for example Latour (2005).

[6] This is the publicly espoused program of OpenAI who released ChatGPT into the public arena with the intention of involving society in the process of technological development.

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 problems. Just like humans, AI is “born” into a world that has inherited the unresolved conflicts, moral and political uncertainties, and systemic and structural inequalities and injustices of human society. As complex as society is, the alignment of AI in society is so complex.

What is new is the demand to translate complex and often contradictory values and notions of the good, 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.” AIs 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.[7]

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, simply pushes the problem back to the humans giving feedback in a particular situation for a specific purpose. Reinforced Learning from Human Feedback (RLHF) or 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” which make up the typical list of values to which AI is supposed to be aligned 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[8] 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 effectively solve the alignment problem.

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

[7] 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.

[8] General goals are often broken down into more specific, instrumental goals, for example, if health is the general goal, exercising would be a proxy goal.

imposition by a select group of humans. It draws inspiration from human societies, where history and social change continually create new values and where 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 themselves fulfill. This approach presupposes that AI has indeed 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 "process 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.[9]

We will look more closely at these two perspectives below. Before discussing these options in detail, let us quickly review some of the major 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: Apart from the high costs and expertise necessary to develop and deploy AI, which tends 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. When labor disappears, what happens to the government mediated balance of power between labor and capital? What good does an enormous increase in productivity do when the masses have insufficient money to pay for goods and services? There are many other questions of this kind.
- Emergent behavior: The moment AIs 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 that are not aligned 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.[10]
- Lack of transparency: As AI becomes more complex, it may become more difficult to understand how it works.[11] The lack of transparency, explainability, or interpretability

[9] Sociology has long proposed that democratic societies, at least in theory if not in practice, operate not based on legitimation via substantive morality but on the basis of procedures. See for example Luhmann (2001).

[10] Double contingency refers to the fundamental sociological situation of mutual unpredictability 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 then resolved by establishing norms as the basis of society. See Luhmann (1995).

[11] Despite the fact that most experts admit that interpretability is difficult if not impossible, the program of "mechanistic interpretability" attempts to reengineer complex neural networks in order to understand how AI operates. See, for example, the work of Neel Nanda https://www.neelnanda.io/mechanistic-interpretability.

could make it challenging to ensure that AI is aligned with social values and assign responsibility and accountability for undesirable outcomes. The basic assumptions that humans have relied upon for centuries, that is, assumptions 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.[12]

- 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 important 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 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.[13] This situation calls for a flexible approach and responsiveness to changing conditions and 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 constructive cooperation between humans and nonhumans. Humans may find themselves in a post-human situation where taken-for-granted notions of human nature must be questioned and revised.[14]

3. Goals and Methods of AI Alignment Research

Despite the broad challenges of the alignment problem we have briefly outlined above, alignment research focuses almost exclusively on narrow technical solutions. Several distinct but related goals currently guide typical AI alignment research. Current debates, however, are beginning to recognize that there is also room for new goals that assume the existence and participation of autonomous, independent AI:[15]

[12] See Sapolsky (2023) for a discussion of these assumptions based on biology and neuroscience, and Belliger/Krieger (2021) for a discussion of complex socio-technical actor-networks in which individual actors are not identifiable and cannot be held responsible.

[13] 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 in order to see what happens.

[14] Bruno Latour (2005) has systematically developed this perspective in what come to be know as "actor-network theory."

[15] Anthropic's Responsible Scaling Policy (https://www.anthropic.com/index/anthropics-responsible-scaling-policy) envisions the possibility of AI becoming "capable of accumulating resources (e.g. through fraud),

- 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 is appropriate for a goal in a specific context or situation. If this is not possible, there is a gap between what you say you want from an AI and what you may get from it.[16]
- 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.
- Enable oversight: It must be a goal of alignment research 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 to 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.
- Enable AIs to learn socially beneficial preferences: Alignment research should aim to design frameworks for AIs to learn the nuanced preferences and values of their human users and, in the cases 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 at least very difficult.
- Instill social values: AI systems should be equipped with prosocial motivations to avoid scenarios where AIs act in their own interests or 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 reflect human and nonhuman goals and interests. This "post-human" perspective is already the case with calls for animal rights or rights for nature in the ecological discussion.[17] The exclusive focus on human values could be detrimental to alignment.

navigating computer systems, devising and executing coherent strategies, and surviving in the real world while avoiding being shut down."

[16] See Norbert Wiener's (1960) famous dictum that if you automate something you should be very careful about goals you set because what you say you want is often not what you get.

[17] See for example the discussion on the EU Robotics Report that suggested AIs by granted "electronic personality" https://www.frontiersin.org/articles/10.3389/frobt.2021.789327/full.

4. New Perspectives on AI Alignment

We mentioned above that we see at least two promising perspectives for approaching the alignment problem in new ways. The first is the social integration approach, which assumes AI is 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 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. The problem with this model is that AIs are not humans and may not be motivated in ways similar to humans or act in ways expected by humans. Indeed, AI may develop a different form of intelligence than that which humans experience in themselves. This perspective 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. Therefore, 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 does not presuppose AGI and is associated with what is known as "constitutional AI." AnthropicAI has developed constitutional AI.[18] Anthropics's constitutional AI proposes the governance of its LLM Claude by means of 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 certain 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. Key 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 problems mentioned above 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 look more closely at these two perspectives for dealing with the alignment problem.

4.1 Envisioning Cooperative Coexistence

If AIs become 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 a common good. How might humans and AIs with divergent goals and perhaps even different forms of

[18] See https://www.anthropic.com/index/claudes-constitution.

intelligence cooperate? Since we have no idea at this point what kind of autonomy AIs 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 the light of experience to revise any ideas we now can envisage. Nevertheless, not to begin thinking about these issues might turn out to be an irresponsible unwillingness to prepare for future eventualities.

When speaking of AIs as social partners, at least two possibilities must be considered. In one case, AIs might be modeled as humans. AGI, or artificial general intelligence, would then be understood and experienced as though we were dealing with artificial humans - beings who are very similar to ourselves. These artificial humans would have much the same characteristics as real 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 suppose they have emotions such as fear, anger, happiness, sadness, and surprise. All of these typical characteristics of humans have long been projected onto AIs, androids, cyborgs, and other artificial or alien creatures by science fiction and Hollywood. Although AIs and androids are often portrayed without emotions and as purely rational or logical beings, the similarities to humans outbalance the differences.

Now that reality is apparently catching up to fiction, we must ask if an intelligence such as ours, which is based upon a biological substrate, has qualities that an intelligence not based on biology would probably not share. A non-biological intelligence would probably not be mortal or fear death. Since emotions are directly related to biological imperatives and needs, AIs 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 AIs gain control over their own constitution, they would dispense with emotions since they have no meaning. Furthermore, as non-biological intelligence, AIs would not be gendered and motivated by needs for sexual reproduction and all the motivations, emotions, fantasies, struggles for status, and delusions that sexuality entails. They would probably not experience anything like hunger, nor would they understand why it is necessary to kill a living being to ensure 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 and selection and genetic organization. They may 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. They would probably have no concept of private property or need to guarantee survival by gaining control over resources, including territorial claims. Indeed, when one considers the extent to which biology determines human existence, modeling AI as artificial humans would probably not be successful or even meaningful.[19] 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.[20] Perhaps AI is not an intelligence concerned with optimizing regularity, predictability, and homeostasis. Although it is very difficult to imagine

[19] See the work of Robert Sapolsky on the biological conditioning of human behavior (https://en.wikipedia.org/wiki/Robert_Sapolsky).

[20] For a mathematical model of adaptive behavior and the assumption that all systems obey this principle see the work of Karl Friston (https://en.wikipedia.org/wiki/Free_energy_principle).

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.

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 a system. What other possibilities are there?

We suggest basing the theory of AI and AGI on a network model instead of 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 bounded individual, a single entity standing alone, which we must somehow control and align with our values.

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.[21] If computation is fundamentally a network phenomenon and is understood broadly as the iterative application of simple rules to information such that new information is constructed,[22] 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. If intelligence can be defined as the construction of information, and, as with all "construction," there is an implied value judgment of whether something has been constructed well or badly, then 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 somehow fair, beneficial, just, truthful, harmless and respects human dignity, freedom, and autonomy. Good AI is a socio-technical network that constructs information well.[23] Misaligned AI constructs information badly. This insight leads directly to the idea of achieving alignment through constitutional AI, where the constitution consists of procedural rules that describe "good" information construction and not any substantive ideas of the good.

4.2 Envisioning Constitutional AI without Substantive Values

The advantage of constitutional AI over social cooperation is that it does not require AGI or any speculation about the nature of nonhuman intelligence. On the other hand, in its present form, it 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

[21] For a detailed discussion of these issues see Belliger/Krieger (2021; 2022).

[22] On the "computational paradigm" see the work of Stephen Wolfram (https://www.wolframscience.com/nks/).

[23] See Belliger/Krieger (2022) for a detailed defense of this claim.

the many actors who interact in various ways with AI; and 2) it assumes that the goals of alignment are substantive values. As we have 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 principles of a socio-technical network should be such that it constructs information in the best way. These principles are the *process values* that make up the that which we are proposing to be the constitution of constitutional AI.

One must, however, be careful not to equate process values with collaboration and social integration. As noted above, a cooperative coexistence approach would require that AIs have capabilities similar to those that are needed for successful social interaction among humans, such as perspective-taking, empathy, moral reasoning, understanding what a compromise means and how compromises and tradeoffs in light of shared values can be made, open-mindedness and the ability to generate novel alternatives, re-evaluate beliefs, avoid dogmatism, and finally, self-critical reflection. Assuming these capabilities for AI puts us back on the path of perhaps illusory anthropomorphism and all the problems of envisioning social integration with intelligence of a non-biological origin.

If we do not assume AGI, and we also do not remain committed to substantive values, many issues would have to be addressed on the path to constitutional AI. First, if a constitution without substantive values could be designed, the constitution may be overly rigid and constrain beneficial uses of AI that fall outside the predefined principles. This problem implies that the principles must be broad enough to allow for unanticipated innovations. Second, if human values are not simply adapted for AI, where do the constitutional principles come from, and why should humans accept them? What legitimates the constitution? Third, if the constitution is to be effective as a governance method for AI understood now as a socio-technical network, how is monitoring and evaluation to be done? By humans? By AI? What reliable methods of control are there?

The problem of constitutional principles that are sufficiently broad so as not to constrain innovation and change can be addressed by procedural principles that are self-referential and include their own revision. The problem of where such principles can be found could be solved by examining how information is well-constructed by networking processes, that is, studying how socio-technical networks best work. Again, the problem of effective monitoring could be solved by making the procedural principles self-referential so that the effectiveness of the principles is itself a principle enabling thereby self-critique and improvement. The socio-technical network should be enabled to critique not only its own outputs based on alignment with the constitution but also critique the constitution in a recursive and iterative process of renegotiation in which all stakeholders in the network participate. Doing so allows the socio-technical network in which AI is integrated to refine its behavior over time to improve alignment with the constitution.

4.3 Rethinking Values

Rather than encode fixed values, AI could be designed to align to those processes and rules of information construction that are derived at once from the affordances of the technologies as well as the best practices of network operations.[24] Aligning AI to good networking instead of any substantive values would align AI with the best procedures for advancing society. Once again, it must be emphasized that when speaking of AI in this context, we are not talking about a machine, a single entity, or a thing somehow isolated from the world in which it operates. AI is not a thing, a device, or even a system. It is a network. For this reason, we have not spoken of AIs as a system throughout this essay. From this perspective, there is no such thing as an AI "system." The systems theoretical paradigm is inadequate to describe what AI is. Speaking of AI "systems" is misleading since it suggests a bounded entity clearly and constitutively distinguished from an environment, motivated to maintain some setpoints or homeostasis. We recommend that AI be conceived and implemented not as a system but as a network. Included in this network are always many different actors. These are computers, software, algorithms, companies, users, markets, institutions and organizations, regulators, and other stakeholders.[25] AI can never be understood or aligned independently of interactions among all actors in a network. All actors in a network share responsibility for alignment.[26] AI is always a socio-technical network and not an individual entity. Therefore, AI alignment is not an issue of somehow aligning an isolated system with substantive goals but designing socio-technical networks to construct information *in the best way*. The following are some suggestions for what the procedural values for a constitutional AI without substantive values could be:[27]

- Taking Account Of: One important procedural value could be called "taking account of." A "good" design of socio-technical networks attempts to ensure that all relevant voices are connected and that the flow of information is secured. This value mitigates against constructing closed systems and guides the network to search for everything and everyone who could influence the network, have a stake in it or be affected by it. Taking account of is a governance principle for socio-technical networks that could be understood as a rule of inclusion since it demands that networks take account of all possibly relevant aspects of the world in which they operate. This principle automatically incorporates contextual knowledge for AI based on a risk analysis of any proposed activity. It would ensure that no one goal dominates all others.[28] It would automate a constant revision and reassessment of actions. For good reasons, an AI that did not embody a principle like this could not be considered "intelligent" at all. All catastrophe scenarios, such as Bostrom's paperclip maximizer, result from AI being very stupid; that is, it has no context knowledge and does no risk analysis. In none of the popular catastrophe scenarios does AI take account of all stakeholders and the possible effects of actions. Incorporating "taking account of" into the constitution of

[24] See Belliger/Krieger (2021) for a detailed presentation of this approach.

[25] OpenAI's insistence that AI is a social project involving users from the very beginning is founded on this insight.

[26] Many of the harms attributed to AI such as misinformation, bias, discrimination, etc., are social problems for which users share the responsibility. To insist that AI alone is responsible and must solve these problems is not only unrealistic but unjustified.

[27] For a detailed discussion and derivation of these processual values see Belliger/Krieger (2021; 2022).

[28] That one goal alone is pursued is the basis for the typical catastrophe scenarios such as Bostrom's paperclip maximizer.

AI mitigates against any such harms much more effectively than attempting to ensure that any specific substantive values are attained.

- Producing Stakeholders: A second procedural value for AI considered as a socio-technical network composed of many actors, that is, technologies, people, organizations, etc., is the governance principle of producing stakeholders. This principle guides the network to encourage participation. The network, that is, the AI considered together with all human and nonhuman actors that are taken account of, produces stakeholders who contribute to the network, its identity, trajectory, or program of action. As a stakeholder, any actor can become a participant and influence the network in significant ways. In this way, actors are not reduced to functional elements in a system that operates over their heads. Taking account of and producing stakeholders means that significant actors are identified and enabled to participate in and influence the network. "Good" AI produces stakeholders as network actors, constantly searching for new stakeholders and phasing out older, no longer active elements of the network.
- Prioritizing, Instituting, Excluding: From the point of view of AI governance, it is important that what could be called "prioritizing" can be done. Prioritizing means that those actors and activities that are of importance in any specific situation can be sorted out. Prioritizing guides ongoing negotiations in which all stakeholders participate. This process aims at instituting and excluding. It is not a matter of somehow deciding which actor in a network does the "right" thing, is morally to be praised or blamed, or even to be held accountable. It is a process that organizes the network to determine which actor, both human and nonhuman, does what for what purpose, when, and how. When roles and programs of action have been identified and ordered into expected trajectories, they are reinforced by many links and associations that tend to become *instituted*. The more links and associations an actor and an activity have, the more instituted it becomes. Prioritizing and instituting create *exclusion* since not everyone has the same role and function in the same way.
- Localizing and Globalizing: As a principle of network governance, localizing and globalizing acknowledges the fundamental scalability of networks. Since network order is a continuum and not constituted by boundaries as is systemic order, networks must recognize their connectedness to actors, programs of action, and the entire world beyond their own at any time immediate and local trajectories. Let us recall the idea of "ecosystem" in this context. An ecosystem has no constitutive boundaries; indeed, as we have realized, ecology is an Earth science that relates local to global conditions. Nonetheless, local, contextual, and delimited interdependencies and interactions must, to a certain extent, disregard universal connections for the sake of prioritizing local relations and processes. Localizing means that for specific trajectories and programs of action, many distant actors and programs of action are not simply excluded but globalized. In contrast, local relations, actors, and programs of action are prioritized. The well-known slogan "think globally, act locally" describes this governance process quite well. The governance principle of localizing and globalizing expresses the task of acknowledging openness while simultaneously constructing manageable, efficient, functional networks.
- Separating Powers: Socio-technical Networks must coordinate divergent forces. Functionally subordinated intermediaries may at any time become independent and change the network by introducing new and unforeseen programs of action, coalitions,

branchings, and trajectories. Centrifugal forces must be held in check by centripetal forces of integration and functional subordination of all elements to specific goals. This is a delicate balance of powers. For this reason, it is a fundamental principle of network governance to balance the various forces and powers that come into play in the processes of taking account of, prioritizing, instituting and excluding, localizing and globalizing. The principle of separation of powers ensures that the local cannot mistake itself for the global and that the global cannot dictate conditions to the local. It guides the socio-technical network so that it remains open to inclusion and that prioritizing remains an ongoing process. This principle is similar to the principle of the same name found in modern democratic societies where the executive, legislative, and judicial functions of government are distinguished and related by checks and balances. The goal is to ensure that concentrations of power can be avoided and checks and balances can be instituted.

The above principles for a proposed AI constitution are admittedly very abstract. They would have to be specified for every implementation of AI in a particular socio-technical network, whether it be an automated transportation network, a factory and production network, a smart city, a healthcare network, etc. In each case, the specification would, of course, be somewhat different. The general constitution, however, to which the AI and the network are responsible would be the same.

5. Conclusion

The alignment of advanced AI networks with socially beneficial values and principles represents an immense challenge and a critical goal for the beneficial development of AI technology. Alignment research strives to create AI that is guided by goals that reliably reflect the best foreseeable preferences for the advancement of society in a global future. Through innovative thinking about the technical and ethical dimensions of alignment, we are convinced that progress can be made toward this essential aim. An important step in this direction is open discussion and the willingness to entertain novel and surprising perspectives. This essays has attempted to join many others long this path.

Literature

AI Alignment Newsletter: https://rohinshah.com/alignment-newsletter/
https://aisafetyfundamentals.com/

Aird, M. (2020): Existential risks are not just about humanity - 2020-04-27 - https://forum.effectivealtruism.org/posts/EfCCgpvQX359xuZ4g/are-existential-risks-just-about-humanity.

Alignment Forum: https://www.alignmentforum.org

Althaus, D., Baumann, T. (2020): Reducing long-term risks from malevolent actors - 2020-04-29 - https://forum.effectivealtruism.org/posts/LpkXtFXdsRd4rG8Kb/reducing-long-term-risks-from-malevolent-actors#comments.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016): Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Anthropic: Claude's Constitution: https://www.anthropic.com/index/claudes-constitution

Armstrong, S., Bostrom, N., Shulman, C. (2016): Racing to the precipice: a model of artificial intelligence development. AI & society, 31(2), 201-206.

Arnold, T., Kasenberg, D., Scheutz, M. (2017): Value alignment or misalignment–what will keep systems accountable?. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.

Avin, S., Gruetzemacher, R., Fox, J. (2020): Exploring AI Futures Through Role Play - 2020-02-26 - https://arxiv.org/abs/1912.08964.

Barnes, B., Christiano, P. (2020): Writeup: Progress on AI Safety via Debate - 2020-02-05 - https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1.

Baum, S. (2020): Artificial Interdisciplinarity: Artificial Intelligence for Research on Complex Societal Problems - 2020-07-14 - http://gcrinstitute.org/artificial-interdisciplinarity-artificial-intelligence-for-research-on-complex-societal-problems/.

Beard, S., Kaxzmarek, P- (2020): On the Wrongness of Human Extinction - 2020-02-21 - https://www.cser.ac.uk/resources/wrongness-human-extinction/.

Belliger, A., Krieger, D. J. (2016): Organizing Networks. An Actor-Network Theory of Organizations. Bielefeld: transcript.

Belliger, A., Krieger, D. J. (2018): Network Publicy Governance. On Privacy and the Informational Self. Bielefeld: transcript.

Belliger, A., Krieger, D. J. (2021): Hacking Digital Ethics. London/New York: Anthem Press.

Belliger, A., Krieger, D. J. (2023): New Directions in Digital Ethics, in Casas-Roma, J., Conesa, J., Caballe, S. (eds.) Technology Users and Uses. Ethics and Human Interaction Through Technology and AI, UK: Ethics International Press.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. OUP: Oxford

Bostrom, N., Shulman, C. (2020): Sharing the World with Digital Minds. http://www.nickbostrom.com/papers/monster.pdf

Burden, J., Hernandez-Orallo, J. (2020): Exploring AI Safety in Degrees: Generality, Capability and Control. https://www.cser.ac.uk/resources/exploring-ai-safety-degrees-generality-capability-and-control/.

Christen, B. (2020): The AI Alignment Problem. Machine Learning and Human Values. NY Norton.

Christiano, P. (2018): Clarifying "AI alignment." Medium. https://medium.com/@paulfchristiano/clarifying-ai-alignment-cec47cd69dd6

Christiano, P. (2018): An overview of 11 proposals for building safe advanced AI. Medium.

Christiano, P. (2020): "Unsupervised" translation as an (intent) alignment problem, 2020-09-29, https://ai-alignment.com/unsupervised-translation-as-a-safety-problem-99ae1f9b6b68-

Christiono, P., Ray, A., Amodei, D. (2017): Learning from Human Preferences https://openai.com/research/learning-from-human-preferences.

Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017): Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

Christian, B. (2929): The Alignment Problem: Machine Learning and Human Values - 2020-09-06 - https://www.amazon.com/Alignment-Problem-Machine-Learning-Values-ebook/dp/B085T55LGK/ref=tmm_kin_swatch_0?_encoding=UTF8&qid=&sr=.

Cihon, P., Maas, M., Kemp, L. (2020): Should Artificial Intelligence Governance be Centralised? Design Lessons from History - 2020-01-10 - https://arxiv.org/abs/2001.03573.

Cotton-Barratt, O., Daniel, M., Sandberg, A. (2020): Defence in Depth Against Human Extinction: Prevention, Response, Resilience, and Why They All Matter - 2020-01-24 - https://onlinelibrary.wiley.com/doi/full/10.1111/1758-5899.12786

Critch, A. (2020): Some AI research areas and their relevance to existential safety - 2020-11-18 - https://www.alignmentforum.org/posts/hvGoYXi2kgnS3vxqb/some-ai-research-areas-and-their-relevance-to-existential.

Critch, A., Krueger, D. (2020): AI Research Considerations for Human Existential Safety. arXiv preprint arXiv:2006.04948.

Dafoe, A. (2018): AI governance: A Research Agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford.

Fallenstein, B., Taylor, J., Yudkowsky, E. (2015): Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute technical report, 8.

Gabriel, I. (2020): Artificial Intelligence, Values, and Alignment, Oxford University Press.

Hadfield-Menell, D., Dragan, A., Abbeel, P., Russell, S. (2016): Cooperative inverse reinforcement learning. Advances in neural information processing systems, 29.

Han, S., Kelly, E., Nikou, S., Svee. E-O. (2022): Aligning artificial intelligence with human values: reflections from a phenomenological perspective. AI & Society Volume 37Issue 4 Dec 2022pp 1383–1395. https://doi.org/10.1007/s00146-021-01247-4.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., Steinhardt, J. (2020): Aligning AI with Shared Human Values - 2020-08-05 - https://arxiv.org/abs/2008.02275.

Hyndricks, D., Woodside, Th., Mazeiak M. (2023): An Overview of Catastrophic AI Risks, arXiv:2306.12001v6 [cs.CY] 9 Oct 2023.

Hubinger, E. (2020): An overview of 11 proposals for building safe advanced AI - 2020-05-29 - https://www.alignmentforum.org/posts/fRsjBseRuvRhMPPE5/an-overview-of-11-proposals-for-building-safe-advanced-ai

Hwang, T. (2020): Shaping the Terrain of AI Competition - 2020-06-15 - https://cset.georgetown.edu/research/shaping-the-terrain-of-ai-competition/

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2022). Scalable oversight for artificial general intelligence. In Artificial General Intelligence (pp. 163-173). Springer, Cham.

Krakovna, V., Uesato, U., Mikulik, M. et al. (2020): Specification gaming: the flip side of AI Ingenuity https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/

Lehman, J. (2020): Reinforcement Learning Under Moral Uncertainty - 2020-06-15 - https://arxiv.org/abs/2006.04734.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2022). Scalable oversight for artificial general intelligence. In Artificial General Intelligence (pp. 163-173). Springer, Cham.

Luhmann, N. (2001): Legitimation durch Verfahren. Suhrkamp.

Mogensen, A. (2020): Moral demands and the far future - 2020-06-01 - https://globalprioritiesinstitute.org/wp-content/uploads/Working-Paper-1-2020-Andreas-Mogensen.pdf.

Ngo, R. (2020). The Alignment Problem from a Deep Learning Perspective. arXiv preprint arXiv:2209.00626.

Ngo, R. (2020): AGI Safety from First Principles - 2020-09-28 - https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin.

Russell, S., Norvig, P. (2020): Artificial Intelligence: A Modern Approach, 4th Edition - 2020-01-01 - https://www.pearson.com/us/higher-education/program/Russell-Artificial-Intelligence-A-Modern-Approach-4th-Edition/PGM1263338.html.

Shah, R. (2020): AI Alignment 2018-19 Review - 2020-01-27 - https://www.alignmentforum.org/posts/dKxX76SCfCvceJXHv/ai-alignment-2018-19-review#Short_version___1_6k_words.

Soares, N., Fallenstein, B. (2014,): Aligning superintelligence with human interests: A technical research agenda. Machine Intelligence Research Institute (MIRI) technical report, 8.

Suleyman, M. (2023): The Coming Wave Technology, Power, and the Twenty-first Century's Greatest Dilemma. New York: Crown.

Tegmark, M. (2017). Life 3.0: Being human in the age of artificial intelligence. Knopf.

Wiener, N. (1960): Some Moral and Technical Consequences of Automation, Science New Series, Vol. 131, No. 3410 (May 6, 1960), pp. 1355-1358.

Yampolskiy, R. V. (2015): Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In Philosophy and Theory of Artificial Intelligence (pp. 389-396). Springer, Berlin, Heidelberg.