Review Article

Disentangling Fiction from Frameworks: A Critical Review of Instrumental Convergence in AI Ethics Risk

Rachel Rishita

Henry M. Jackson High School, 1508 136th St SE, Mill Creek, WA, 98012, United States

ABSTRACT

As Artificial Intelligence continues to evolve at a rapid pace, the concerns regarding its capabilities circling around existential risk have increased at a similar rate. The concept of instrumental convergence primarily – the theory that intelligence AI agents may become indifferent to the programming and commands of humans and prioritize self-preservation – have alerted individuals, as this could endanger humanity. However, this paper critically evaluates the extent of the validity of instrumental convergence as a real-world AI safety concern by arguing that it overextends AI's capabilities and is driven by anthropogenic assumptions. AI agents lack the fundamental characteristics such as intrinsic motivation, agency, and the capacity for self-generated terminal goals which would otherwise drive the theory of instrumental convergence. Moreover, other barriers to AI alignment are discussed in this paper, specifically the orthogonality thesis and the tricky notion of aligning complex human values with agents. Counterarguments to popular X-risk narratives are presented to support the conclusion that instrumental convergence is overstated. My findings suggest that a more accurate public and professional understanding of AI's limitations and capabilities would bring us closer to achieving a safe and effective relationship between humanity and AI.

Keywords: Artificial Intelligence; Machine Learning; Instrumental Convergence; X-Risk; Terminal Goals

INTRODUCTION

The public has grown increasingly aware of the prevalence of Artificial Intelligence in our daily lives. From acknowledging the rise of machinery and technology, to directly being able to converse with it as an AI ChatBot. Needless to say, many have suspected that AI could have additional risks which may even cause widespread harm against humanity. Although there is no doubt the possibility of a rogue AI or piece of technology that may harm society, would the effects come to such drastic measures as to hurt other humans? Is the story presented above an accurate depiction of AI capabilities or mainly a ruse created by those misunderstanding agents?

The rapid emergence of Artificial Intelligence, popularized mainly by generative AI, has sparked

Corresponding author: Rachel Rishita, E-mail: rachelrishita22@yahoo.com.

Accepted August 17, 2025
https://doi.org/10.70251/HYJR2348.34311318

controversial debates on its capabilities and implications. Today's version of AI, after being built upon from past machine intelligence explorations and workshops, primarily consists of image recognition/ processing, humanoid robots, chatbots used to converse and negotiate with each other, and humans. Many are excited about the opportunity of efficiency, imagination, and discovery that revolves around partnering the globe with AI. However, some argue that this newfound technology, which even experts cannot control well, can be dangerous. AI alignment is the problem of ensuring that a powerful AI has goals and behaviors aligned with human welfare.

There are many barriers to alignment; in this paper, we discuss instrumental convergence and how the possibility it presents is unrealistic. Instrumental convergence is the idea that sufficiently advanced intelligent systems with a wide variety of terminal goals – ultimate values an AI is to pursue – would pursue very similar instrumental – self-preservation – goals. The AI's terminal goal might be something it creates rather than the original goal it was programmed to achieve. Such terminal goals may include the following (Table 1):

Table 1. Terminal Goals and Their Descriptions

Goal	Description
self-preservation	preventing itself from being shut off or destroyed
goal integrity	avoiding having its goals changed by an external entity
tech advancements	improving the speed/efficiency of its computers

Illustrates a comprehensive list of common terminal goals which may be used to explain instrumental convergence and its meaning.

Although the goals seem positive, the concern lies in the possibility of an agent exceeding human capabilities and intelligence, thus making it unpredictable and unstoppable when or if it causes harm to human beings. A group of people who believe in the possibility that such goals of instrumental convergence would result in human extinction or an incredibly damaging global event is titled "The X-Risk" Group. To make instrumental convergence a real possibility, an AI agent would essentially need to have the following characteristics of human beings: selfishness, self-preservation, and the idea to produce actions with its own personal identity. The only way an agent could prioritize its well-being is to leave its robotic nature and prior programming behind and begin to yearn like a human being. Thus, the X-Risk concern of instrumental convergence is overstated, since AI lacks intrinsic desires, self-preservation instincts, and agency unless explicitly programmed with such tendencies. In other words, the worries of AI gaining superintelligence and the terminal goal of resource acquisition are very unlikely. These beliefs stem from a phenomenon titled "anthropomorphism", which suggests the attribution of human characteristics to inanimate objects, concepts, or non-human entities, in this case it would be the belief of selfishness from an AI agent.

The belief of anthropomorphism is much more dangerous than it seems, especially when discussing aligning AI values with human interests. By associating agents as fellow human beings we overestimate its capabilities which only involve pattern recognition, confuse it to have moral rights or responsibilities distracting real accountability from its creators, form excessive trust in agents, and mask the data collection and biases in its training for friendly or independent traits. All of these features would not only slow us down in figuring out the partnership between man and machine but also hurt society into believing that they've gained a human-like ally instead of a tool needed to be controlled. Thus, the purpose of this study is to debunk the instrumental convergence theory and state the dangers of believing such theories.

BACKGROUND

For a small quantity of background on the major areas of AI advancement that have been significant throughout the years, the three major areas include generative AI systems, machine learning, and facial recognition.

Generative AI Systems – as the name suggests – can generate images, audio, video, and other content when prompted by a user and can be used in various industries such as education, government, law, and entertainment. Reaching 100 million users, advanced chatbots, virtual assistants, and language translation are examples of AI systems currently being used by a significant fraction of the public, thus gaining global recognition. Receiving backlash from AI skeptics, many are concerned by the ethical dilemmas faced by generative AI, such as replicating the work of artists and authors, generating code for more effective cyberattacks, and even creating new chemical warfare compounds to

increase destruction.

Going hand in hand with the first application of AI, machine learning is being used in fields that require advanced imagery analysis, mainly to assist the medical diagnostics process. Benefits of this development include identifying hidden or complex patterns in data, detecting diseases earlier, improving treatments, more consistent analysis of medical data, increased access to care, and more. However, due to the limitations and bias often found in data, machine learning is used to a certain extent to reduce the dangers and ineffectiveness of its diagnosis and inequalities for specific patient populations.

Law enforcement often uses facial recognition to support criminal investigations, video surveillance, and match travelers to their passports at airport security checkpoints. Providing benefits in terms of security, facial recognition can also be abused to cause harm to others. "Despite improvements, inaccuracies, and bias in some facial recognition systems could result in more frequent misidentification for certain demographics" (10). With AI's innovative nature, several benefits are connected to its uses and capabilities. Being able to identify common and complex patterns in medical data and conduct efficient and repetitive actions, AI could improve the field of healthcare by assisting researchers in developing cures and treatments for resistant illnesses.

Although the benefits outweigh the cons, there are clear dangers to the three most common types of Artificial Intelligence mentioned above. For instance, generative AI can spread false information, fabricate false evidence to frame others, inappropriate pictures of others, and other media that can endanger individuals. In a more ethical sense, generating artwork and literature using generative AI can compromise the hard work and creativity of artists and writers and remove the emotional element of art. Facial recognition may be more efficient; however, when misused, it can be used for widespread surveillance, biometric data collection, and storage, and lacks transparency, leaving users unsure what facial data is being used for. Finally, the topic of machine learning and how it relates to AI alignment concerns, such as instrumental convergence, is the main discussion in this paper. Machine learning is how AI can learn and adapt by using new algorithms and existing models. With the fear of machine learning being too powerful, possibly powerful enough to create a superhuman AI that no longer aligns with human values, a problem of AI alignment stems, alerting critics falling under the safety, doubts, ethics, and capabilities category.

According to the article "Why the arguments against AI are so confusing", the group doubting AI's capabilities as well as the group wary about the actions of AI being ethical often fall under the "doubt" and "capabilities" group, with one side of the capabilities group suspecting that AI is trained by observing existing works and through machine learning adjusts itself to mimic a human's style and the "doubt" group not trusting AI to conduct work in general. Similarly, this argument is connected to the concern of human workers losing job opportunities because of a machine mimicking how they work more efficiently and error-free (11). To directly connect these two concerns, it is clear that others in the workforce and those concerned with the well-being of humanity are worried about AI "stealing" a person's skills and taking a job a human would have been hired for. A different direction others have looked at the topic of AI's machine learning is that there is a possibility that with a finite number of existing resources for AI to learn off of, it would eventually observe all work there is to learn and refrain from getting more intelligent; or, if the work observed by AI is biased, then that could lead it to become biased as a result. One example directly provided by the article recalls how, during the GPT-2 era, AIs thought that Japan was more prosperous than the US due to the model being trained primarily on "American writers' impression of the richer parts of Japan," (7) even though the opposite was true in terms of average prosperity. Others are worried that AI might give its primary holders too much power over others; these holders could include private corporations and governments. To directly connect these groups together, they may form similarities in thought processes with the ethics group, which discuss the ethical dilemma of incorporating AI into society. Such specific concerns include the unemployment rate that arises as a result of AI replacing workers, AI art and media circulating the digital world stealing work and opportunities for human artists as well as possibly spreading false information, and the negative environmental effects of AI being put to light recently.

Although the above concerns are valid, unfortunately they are not the focus of this work, instead this paper prioritizes concerns of the "safety" group – otherwise known as the X-Risk group – and its concerns on artificial intelligence alignment.

The concern of AI safety includes a group of people wary about the possibility of AI becoming too intelligent or unsupervised and causing harm to human beings. This group is typically known as the "X-Risk" group,

fearful of the consequences of neglecting AI alignment with human values and knowledge of consequences, worry that a superintelligent AI would be able to develop sub-goals or ulterior motives which override its own and ignore the need to protect humanity. Although hinting towards a malicious AI, the X-risk group actually worry about an indifferent AI with different goals and suffixes instead of a completely dystopian view of an AI supervillain.

To prevent a risky AI agent, experts have been working towards AI alignment, which – as previously mentioned earlier – works to match AI's capabilities with human values in order to prevent obscuring humanities interests with AI's rogue behaviors. While the simplest rules of robotics lay out the obligations an agent has to preserve human health, obey orders, and protect its own existence while following the previous rules, the issue of alignment is much broader than the ethical guidelines stated. Before diving into the flaws of instrumental convergence, we must gain knowledge on the major barriers to AI alignment.

CONCEPT CLARIFICATIONS

In order to fully understand this paper as well as distinguish between related ideas, it is important to define and differentiate between the main topics discussed, including goal-directedness, instrumental goals, anthropomorphism, etc. To begin, we must look at the difference between goal-directedness in AI and agency; in the context of Dan Dennet's "Intentional Stance", goal-directedness refers to the pattern of behavior where a system acts towards achieving a goal (5). On the other hand, agency refers to the system which initiates such actions to occur (5). A similar idea can be placed by Russel & Norvig's "Artificial Intelligence: A Modern Approach", where agency is understood as "anything that can be viewed as perceiving its environment through sensors and action upon that environment through actuators" (6). Alternatively, goal-directedness is the driving force behind the actions conducted (6). So in other words, goal-directedness is the "why", and agency is the "how".

Moving on, it needs to be emphasized that terminal and instrumental goals are terms that cannot be used interchangeably. According to philosophers Nick Bostrom and Steve Omohundro, terminal goals are the intrinsic values an agent wants to achieve, while instrumental goals are subgoals used to achieve the terminal goals. Thus, seemingly similar, different structures (2, 3).

Finally, clarifying a topic which comes up later in this paper is anthropomorphism, and how it differs from technical agency. Anthropomorphism arises when an inhuman object or thing is given human characteristics – like emotions or behavior which it cannot act upon; technical agency however refers to the ability of a system to act and make decisions, either based on its own instrumental goals or based on its pre-programmed logic or algorithms. Key differences include the nature of the attribute, focus, and implications (7).

BARRIERS TO AI ALIGNMENT

For more context, we briefly discuss the additional barriers to AI alignment, including the fragility/complexity of human values and the orthogonality thesis. In later sections, we will focus on instrumental convergence in detail.

AI alignment requires that we align AI values with human ethical principles. Such values and principles can include empathy, protecting other human beings, justice, morality, etc. This process involves "translating abstract ethical principles into practical technical guidelines and ensuring that AI systems remain auditable and transparent" (9); however, defining said principles and values prove difficult since they vary amongst different cultures and ideals. For example, a differing experience regarding socialization may shape the way AI may learn to interact with humans, or how AI agents situated in different parts of the globe may embody different religious identities if its owner wishes to have it adhere to a religious standard.

Additionally, based on the creator's viewpoint of politics or history the way the agent distributes information may vary. Nonetheless, a few statements can be agreed upon by most of the population, such as hurting someone is considered wrong and punishable. On a technical note, "operationalizing these values to make them explicit, traceable, and verifiable" (9) is still something technical professionals are investigating. In order to visualize such a process, readers should consider an example of an AI system being used to diagnose patients in a hospital, by aligning itself with the values of human doctors the AI should strive to navigate patient autonomy and privacy, remaining transparent whilst explaining the thought behind its recommendations and conclusions to patients.

However, even here, "value alignment also demands active participation from various stakeholders, including patients, doctors, and regulations, to ensure that the

system is sensitive to the needs of the people it serves" (9). In order to perfect AI as a tool to stand alongside human beings in fields where implementing human values and morality is crucial, audits are needed to maintain value alignment throughout the AI system's life cycle, and independent and internal assessments "ensure that AI systems continuously align with ethical standards and societal norms" (9). To continue with the discussion of alignment, red lines in AI that bring light to the ethical boundaries that AI does not cross must provide clear moral boundaries to ensure that AI systems do not go rogue and cause harm. Forming normative criteria for AI, establishing consequentialism, forming simple but capable AI methods for solving challenging abstract problems that model the real world, and finally, natural consequentialist problem-solving methods are current problems that plague the mission to align agents with human values.

Regarding AI, the Orthogonality Thesis states that "an agent can have any combination of intelligence level and final goal, that is, its final goals and intelligence levels can vary independently" (8). Three key points include that an AI's cognitive abilities do not inherently determine its objectives, thus pursuing goals that are trivial, harmful, or beneficial depending on its programmed objectives; the thesis underscores the importance of AI alignment and is a concept of instrumental convergence. A novel idea involves a hypothetical AI designed to create paper clips; if that is its sole programmed purpose, then it might take up all available resources or turn to unintended and harmful consequences, such as environmental degradation or human harm, in pursuit of its narrow goal: making paper clips. With its similarities, the orthogonality thesis does not differ from instrumental convergence as much since it serves as a supporting piece rather than its own idea.

Although the above concerns are valid, we strictly focus on the unrealistic expectations of the failure of AI alignment presented by instrumental convergence.

INSTRUMENTAL CONVERGENCE

As previously discussed, instrumental convergence is an X-Risk concern or concept involving an AI agent sufficiently advancing its system and intelligence to create terminal goals that interfere with human interests. Common concerns include self-preservation, goal integrity, resource acquisition, tech advancements, and cognitive enhancements. After carefully considering the possibility, an AI agent programmed to follow the basic rules of robotics should not be able to learn human-like traits and intelligence to the best of a programmer's ability. This argument I have formed is a mention or contribution to the anthropocentric fallacy. Anthropomorphism – AI gaining human-like feelings and characteristics – is an over-exaggeration of an AI's capability, and contradicts the difficulties of AI alignment mentioned above, stating that connecting AI to human values and preventing the hypothetical situation created by the Orthogonality Thesis is still to be solved. Overall, the risks of instrumental convergence are overstated because AI lacks intrinsic desires, self-preservation instincts, and agency unless explicitly programmed with such tendencies. To be clear, this paper describes the topic of instrumental convergence in the situation of an AI creating dangerous sub-goals that override its data.

According to Adriana Placani's "Anthropomorphism in AI: hype and fallacy", the main problem with anthropomorphic language are that "it asserts an out-of-place human-centric perspective that conceals the reality of how these networks work, as well as their limitations" (1). In other words, the human need or habit to sometimes put a human connotation on an inanimate object overrides how AI agents work. One example of anthropomorphism includes language models such as ChatGBT, Bing Chat, etc. With the power of generating human-like responses to almost any problem a user can identify, many patients often prefer chatbot responses over physician responses due to the higher quality and empathy. According to the article, users noted that the chatbot's dialogue came across as more empathetic than human physicians. The instance of humans anthropomorphizing here is assigning empathy – a human characteristic – to a chatbot, instead of understanding a patient's recorded systems and taking large amounts of information to convert it to a single response. With these false narratives, increased fears and hopes may arise inside people who misinterpret how AI systems connect with human beings. Such beliefs often come from the generalization and exaggeration of AI's abilities, our understanding of AI alignment and how agents operate decreases.

Anthropomorphism is more common than people notice and is mainly regarded as either a factual error or an inferential error. Although the implications of a phrase as simple as "my car is mad at me today" may seem harmless, connecting human emotions with AI and assuming that agents can obtain the human-like ability to override their original program and goals can "distort moral judgments about AI". The moral judgments that the author of the article mentioned above goes in-depth

about are judgments regarding moral character, status, responsibility, and trust; however in this paper we will focus on moral character and status.

Moral character involves a collection of qualities that differentiate one individual from another. This may include how a human thinks, feels, behaves, their characteristics, and most importantly, their beliefs and values, which are considered most beneficial to society. One criterion of possessing moral character excludes non-human beings from the possibility of understanding human traits and values. Since aligning AI with such traits and values is still ongoing and may not be possible for a long time, the exclusion also includes AI agents. However, with the problem of anthropomorphism, humans categorizing AI as empathetic or friendly, our moral judgments become more distorted and put AI as a subject of "moral character evaluations". This causes AI agents to be mistakenly open to being categorized as good, evil, empathetic, courageous, bad, trustworthy, etc. These characteristics assigned to AI would impact the way humans predict their actions. For example, when people are illustrated or thought of as evil, they are predicted to harm others. So if we were to perceive AI's future/current actions this way and assume its motives and actions in an emotional manner, our interactions, dispositions, expectations, and attitudes toward AI would change forever.

Moral status is defined as someone who understands that there are "certain moral reasons or requirements concerning how it is to be treated for its own sake", basically meaning that possessing moral status is to have moral obligations. The author mentions Matthew Lioa's "Ethics of Artificial Intelligence", which defines the qualifications for having moral status as being alive, conscious, able to feel pain, able to desire, and capable of rational agency. Neither of these characteristics can be given to an AI agent for apparent reasons. If an agent cannot possess moral status that allows it to conduct actions that matter morally for its own sake, it cannot go through with the resource acquisition associated with instrumental convergence, which protects itself and causes harm to human beings.

To conclude, the instrumental convergence theory that AI would be able to reject alignment, program, and the goals it was given to prioritizing self-preservation is false under the anthropomorphism fallacy, which clearly states that this theory does not apply to an agent under moral judgment does not qualify as an intelligent living being. Additionally, moral judgment shows that AI cannot make its own decisions, which would cause it to act selfishly or possess human traits such as evil or cruelty. Suppose AI alignment was achieved in such a way that AI agents would be able to possess and understand human values. In that case, the topic of anthropomorphism may seem more realistic; however, currently, with little progress towards that goal, the X-Risk concern of instrumental convergence is impossible and uses society's misinterpretations of AI to guide our understanding of it further back.

COUNTERARGUMENTS TO AI X-RISK ARGUMENTS

Investigating a series of holes in the basic argument for X-Risk from superhuman AI systems would allow for a more comprehensive understanding of an AI agent's capabilities and add to the argument created later on by the possibility of instrumental convergence. The article "Counterarguments to the basic AI x-risk case" counters the claims of superhuman AI systems being "goal-directed", sufficiently superior to humans to overpower humanity. The possibility of AI having bad goals would lead to destruction, and if these goal-directed agents are built, they will lead to adverse outcomes.

To begin, the author refutes the argument that superhuman AI systems will be goal-directed by first stating how "goal-directedness" is a vague concept, and to understand its meaning, applying the definition "utility maximization" would be wiser (12). Generally, the term is used in many ways and does not always imply aggressive, world-optimizing behavior. Especially in economic, training, and coherence-based pressures, "goal-directedness" might be favored. Utility maximization refers to an AI agent making decisions to maximize its perceived satisfaction or "utility" based on a defined utility function, aiming for the best possible outcome given its goals and constraints. This allows for the possibility of the argument mentioned since utility maximization would make a large class of goals deadly according to the article, since a system that maximizes a utility function tends to pursue convergent instrumental goals like acquiring more resources and underlining a classic AI risk concern.

However, it is important to note that since true utility maximization is impossible, most systems, including humans, are pseudo-agents (pursue goals in limited, context-specific ways without consistent or coherent long-term planning). These systems can conduct practical tasks, thus being goal-directed, without being incredibly dangerous, compared to utility maximization,

which allows for high autonomy and strategizing, which can be risky and unpredictable. Not all goal-directed systems are dangerous since most may remain weakly agentic or pseudo-agentic, which is much more viable and safer. The assumption that AI will eventually use utility maximization to its full potential and become risky and unpredictable is a theory that has not yet been proven.

Under the argument that an AI agent could become more goal-oriented, this does not raise the possibility that their goals could be catastrophically bad. Taking the way humans work into this situation, human values vary and there is no single point in "value space" to hit; in other words, human values do not have a specific attribute or direction which does not differ between other humans, Thus, if an AI agent does possess the same trait or fall somewhere in the human value cluster. Their values can vary the same way and might be no worse than those of a typical human you disagree with.

The author states that we do not need perfect alignment to make this happen, just an alignment "close enough". Training on many data trends to produce minor value differences – the author refers to as "d" – can avoid significant risks and make " d " likely small and safe (12). Essentially, the argument that a goal-directed AI would be bad does not consider the possibility of the goals being malicious, that minor errors in value alignment should still be safe, and that real danger requires both harmful goals and long-term ambition, which is uncommon.

To summarize, this article's counterarguments to the X-Risk case can relate to why instrumental convergence is unlikely or unrealistic. By stating that weak agencies can be safe, bad goals are not inevitable, perfect alignment is not needed, and true danger requires multiple factors, it can be proven that AI becoming destructive is both unlikely and unrealistic.

To add on to the counterarguments from a different perspective which incorporates the fact that AI is not "living" enough to make instrumental convergence a reality, Johannes Jaeger's "Artificial intelligence is algorithmic mimicry: why artificial 'agents' are not (and won't be) proper agents" argues that AI systems do not have intrinsic goals or agency, thus lacking the ability to conduct X-Risk theories. Overall, there are three factors which stop an AI from obtaining instrumental convergence: autopoiesis, embodiment, and environmental scope (4). Autopoiesis is when living systems general goals by themselves, AI systems do not necessarily have the capability to generate such goals unless programmed to do so. AI does not embody a structural presence, meaning it is detached from hardware and physical experience, more of an art rather than a moving and living being (4). Finally, living beings operate in expansive environments, ill-defined and ever moving. AI systems usually operate in well-defined computational spaces. All of these factors matching with the program aspects of the previous article's counterarguments explain why X-Risk is not possible, and imagining a scenario where unalignment would lead to mass destruction is unrealistic.

CONCLUSION

To conclude, the hypothetical situation presented by those who believe in Instrumental Convergence is unrealistic and unattainable for various reasons. It presents a theory that AI can possess human-like selfish traits like resource acquisition, hinting at anthropomorphism, which uncharacteristically overestimates AI's capabilities. Although AI alignment is a genuine concern, stating that an AI would extend to this level is impossible. Developing sub-goals may be possible; however, as demonstrated by the counterarguments discussed previously, these sub-goals may not be bad or lead to extreme consequences. Thus, the anthropomorphism fallacy distorts the public's knowledge of AI, thinking it can act for its own sake, leading experts to assign agency to tools that do not possess consciousness, desires, or self-awareness. Technically speaking, speculative conclusions are the only things that strengthen the possible risks of instrumental convergence. Most AI systems are too weak or pseudogenic to be associated with such levels of resource acquisition. This is why perfect alignment does not need to be achieved; with the weak agents we see today, "good enough" alignment can more than prevent danger. Instrumental convergence is built by science fiction rather than scientific likelihood. So, alignment can be better achieved and obtainable by correctly understanding the capabilities of Artificial Intelligence with a technically informed approach and acknowledging the weaknesses of the instrumental convergence risk. Currently, I suggest focusing on educating others about the true capabilities of AI in order to prevent the risk of anthropomorphism, emphasizing an agent's technical identity instead of its imaginary human one. What may change my mind about the fictionality of instrumental convergence would possibly be the revelation of programming AI to process complicated human values reaching closer to AI alignment, since such an

understanding can spark the human need for selfishness. Maybe then the topic of instrumental convergence would finally be up for debate.

FUNDING SOURCES

The author has no funding sources that supported the research and preparation of this article.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Placani A. "Anthropomorphism in AI: Hype and fallacy - ai and Ethics," SpringerLink, https://link.springer.com/article/10.1007/s43681-024-00419-4 (accessed 2025-06-16).
2. Bostrom N. *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.
3. Omohundro S. "The Basic AI Drives," in *Proceedings of the 2008 Conference on Artificial General Intelligence*, IOS Press. 2008; pp. 483–492.
4. Jaeger J. "Artificial Intelligence is Algorithmic Mimicry: Why Artificial 'Agents' are Not (and Won't Be) Proper Agents," *arXiv preprint*, https://arxiv.org/abs/2307.07515 (accessed 2025-08-15). https://doi.org/10.51628/001c.94404
5. Dennett DC. *The Intentional Stance*, MIT Press. 1987.
6. Russell S and Norvig P. *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2020.
7. "The Difference Between a Media Agency and a Tech Agency," *WellDesigned Blog*, https://www.welldesigned.se/blog/the-difference-between-a-media-agency-and-tech-agency#:~:text=Tech%20agency%20(Technical%20experts)%20… (accessed 2025-08-15).
8. Pyinya A. "ACI#9: What is Intelligence," LessWrong, https://www.lesswrong.com/posts/YQeAjNBACLBmSAzXr/aci-9-what-is-intelligence (accessed 2025-06-16)
9. "Ai value alignment: Aligning AI with human values," World Economic Forum, https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/ (accessed 2025-06-16).
10. U. S. G. A. Office, "Artificial Intelligence's use and rapid growth highlight its possibilities and perils," U.S. GAO, https://www.gao.gov/blog/artificial-intelligences-use-and-rapid-growth-highlight-its-possibilities-and-perils#:~:text=Despite%20improvements%2C%20inaccuracies%20and%20bias,technology%20violates%20individuals'%20personal%20privacy. (accessed 2025-06-16).
11. Why the arguments against AI are so confusing, https://arthurjohnston.com/arguments_against_ai/ (accessed 2025-06-16).
12. Katja Grace. "Counterarguments to the basic AI x-risk case," LessWrong, https://www.lesswrong.com/posts/LDRQ5Zfqwi8GjzPYG/counterarguments-to-the-basic-ai-x-risk-case (accessed 2025-06-16).