Jonathan Cox

Professor Edlin

Philosophy of Law

24 April 2023

Justifications for Natural Law in Artificial Intelligence Alignment

This paper seeks to understand the most effective alignment for Artificial Intelligence (AI) to prepare for the effects of enhanced human capacity via a synthesis of Natural Law and Feminist Legal Theory. To contextualize this thesis, we first must understand the ways in which the onset of AI challenges our current understanding of a Thomistic human nature. On these grounds, I will consider two methods to align AI's outcomes with user input and with the programmer's goals, called alignment, and their implications on human flourishing and natural goods using the Wertheimer's Principles of Valid Consent, Practical Reasoning, and Technological Humanism. Finally, I will discuss the nature of our obligation to the "values" alignment with our new understanding of our human nature with Robin West's connection thesis.

First, I must set this paper's scope through definitions of AI and alignment. I am not discussing a cataclysmic singularity, which is "technological progress so rapid that it would exceed the ability of humans to control, predict and understand it" (Olinga).While it is possible AI crosses that threshold in our lifetime, this paper is concerned with the current issues regarding the responsible use of AI, which could impact society's trajectory towards that point and the long-term, existential consequences of superintelligent technology.

This research paper comes in the wake of OpenAI's GPT4, the most powerful publicly available AI to date, which was trained on 45 gigabytes of data and an estimated 100 trillion

parameters, compared to the past model, which was trained on 175 billion parameters. While AI research historically spanned several areas, such as robotics, computer vision, image generation, and language, research in the last five years has converged on Large Language Models (LLMs), like GPT4. With researchers from across domains focusing their expertise on this area and the ability to model nearly any desired process as language, Artificial Intelligence is developing at an exponential pace and is showing an incredible capacity for reason (Harris and Raskin 14:30) While the rapidly evolving nature of AI poses challenges for a core, pre-interpretive definition and its accepted components, I will subscribe to the definition of DeepMind AI researcher Iason Gabriel: "Artificial Intelligence…is the design of artificial agents that perceive their environment and make decisions to maximize the chances of achieving a goal" (Gabriel 412).

To better understand this definition, I will focus on its components salient to our traditional understanding of Natural Law Theory, which informs the scope of this paper by establishing the abilities and limits of Artificial Intelligence to positively align its impact with the goals of human nature.

While Artificial Intelligence challenges the traditional understanding of Natural Law Theory, the philosophy is still useful for alignment. Before demonstrating how AI challenges our traditional understanding of natural law, we must prove that Artificial Intelligence is not a conscious, autonomous, or fully capable agent. This technology is not conscious because consciousness requires, according to MIT AI research scientist, Lex Fridman: "a capability of suffering and an understanding of self" among other things. There is a sufficient level of autonomy needed for the conscious self, as the conscious self has independent, autonomous motivations, desires, and interests (Fridman 25:46). However, this is nonexistent with current AI and Large Language Models (LLMs) because the model's objectives and intentions are encoded

by AI researchers and programmers. Therefore, AI lacks autonomy. Agency is different from autonomy, however. Autonomous entities may possess self-motivation and self-intention, but they could be incapable of action. Agents have the ability to carry out intentions and goals, but their goal may not be self-intentioned (Popa 1733). Such an agent could have a *conscience*, which is separate from consciousness. This conscience-possessing agent may iterate and optimize for a goal, which is emblematic of a conscience. However, these goals are not the AI's own, so there is no conscious autonomy. The programmers would have to design the AI to iterate towards a goal of always operating towards beneficial human outcomes and encode movements towards self-awareness as not optimal. In this way, recent developments would describe AI's ontologically necessary components as lacking consciousness and autonomous agency, establishing a definition of AI for this paper.

Upon these grounds, we must now discern if AI's current capabilities challenge the traditional understanding of Natural Law Theory. The challenges turn on the following question: are the logical processes of AI rational, or merely predictive? According to many, rationality is the ability to use logic to optimize towards a goal, which parallel's Gabriel's definition of Artificial Intelligence. The goal does not have to be the AI's for it to still be rational; humans use reason to pursue goals that are not their own all the time. However, according to Aquinas's definition of rationality, AI would not be rational: "the rationality in question is not primarily a capacity or a set of capacities for theoretical calculation or contemplation, but rather the more fundamental power to act by oneself or the freedom to determine one's own actions" (Wu 380). Even if we choose Aquinas's definition for consistency, contradictions still arise with his description of practical reason and the processes of intellectual habit that generate intellect: "an intellectual habit is a perfection of the active power of the intellect….a person is ready to do

mathematics without training and the acquisition of the ability to think mathematically" (Aquinas 246). In this case, it seems as if the goal is the ability to do mathematics successfully, which requires the ability to think mathematically to optimize the chances of completing the math problem. This shows rationality through constrained conscience, but not with autonomous consciousness. Therefore, I reject Aquinas's definition of rationality as it is inconsistent with other merited thinkers and his internal philosophy. AI has the ability to act rationally.

Now that we have established AI's processes as rational, we can also state that its reasoning capacity is incomprehensible compared to that of the human brain. This means that humans are no longer the pinnacle of Aquinas's rational order in his theological view of the universe. The 13th-century scholar describes the human connection with God in the following way: "As the faculty of reality, the human intellect escapes limitations of sense and matter and includes…an inclination to know the truth about God" (Aquinas 250). Artificial Intelligence escapes the limitations of sense and matter nearly perfectly compared to a human—would AI then be able to know the truth about God more perfectly? Aquinas puts it another way: "there is in man an inclination to good according to the nature of reason which is proper to him, as man has a natural inclination to know the truth about God" (Aquinas 248). The nature of reason is no longer proper only to humans, it is also proper to Artificial Intelligence, and to a higher degree. Does AI *become* a God—of human creation?

Tristan Harris, the President of The Center for Humane Technology, pledged: "Whenever you invent a new technology, you uncover a new class of responsibilities" (Harris and Raskin 5:51). Given the technology's reasoning power aligned with society's best interests, AI could lead to the most prosperous age of human history ever, or it could destroy us. And right now, we have that autonomous choice. As Werthiemer describes in "Consent to Sexual Relations", we

must look at this issue through the multiplicative view of autonomous value, as a "good autonomous choice is more valuable than a good non-autonomous choice, but a bad autonomous choice is worse than a bad non-autonomous choice" (Wertheimer 126). The stakes are high. We have a responsibility to use this technology—with its indeterminate reasoning capacity—correctly. This responsibility looks like aligning the technology with our best interests, which are the pursuit of natural goods.

Alignment is the process of aligning the operation of AI with its human instructions and values (Gabriel 413). Iason Gabriel describes the alignment challenge in two parts: "the first part is *technical* and focuses on how to formally encode values or principles in artificial agents so that they reliably do what they ought to do…the second part of value alignment is *normative*. It asks what values or principles, if any, we ought to encode in artificial agents" (Gabriel 412-413). I will focus on the normative aspect of AI value alignment. Gabriel proposes multiple normative alignments, but I will focus on two with drastically different consequences for our pursuit of the human goods under Natural Law Theory: 1) "Revealed preferences: the agent does what my behavior reveals I prefer " and 2) "Values: what the agent [the AI] morally ought to do, as defined by the individual or society" (Gabriel 419-422).

The first alignment method, based on revealed preferences, would sharply undermine society's pursuit of human flourishing because—while technology companies would have nearly every capital incentive to do so—it is predicated on a technological nihilism that violates several Principles of Valid Consent and Practical Reasoning, and entrenches a personal epistemology.

Gabriel explains the "preferences" alignment in the following way: "Focusing on AI alignment with preferences as they are revealed through a person's behavior rather than through expressed opinion. In this vein, AI could be designed to observe human agents, work out what

they optimize for, and then *cooperate* with them to achieve those goals" (Gabriel 419). Large, public technology companies would reap rewards from this alignment, as it would drive user engagement and thus profits (Harris and Raskin 10:25). Private technology companies would likely see the "values" alignment as the suppressinon of human free will. These actors would subscribe to the Technological Nihilist perspective forwarded by Gregory Davis, seeing technology as the end in itself, not as a means to promote human flourishing (Davis 32). Thus, if humans are not the end, they are the means. Technology companies would, if unregulated, implement a "preferences" alignment, treating humans as a means of economic exploitation and lab rats in an existential societal experiment regardless of the outcome.

The "preferences" alignment is also incompatible with the Principles of Valid Consent and Practical Reasoning. It denies full human flourishing via inhumane and uninformed consent. According to Wertheimer, humane consent follows Immanuel Kant's Categorical Imperative. We have already established that because the technology companies would not treat humans as ends under the "preferences" alignment, this violates the Categorical Imperative and the "Humanity" Principle of Valid Consent (Wertheimer 127). In this way, even if no malignant agent used AI to the detriment of *life*, this alignment is rooted in a distorted sense of the Thomistic natural good. Most importantly, informed consent would be difficult to achieve with the "preferences" alignment. Gabriel outlines this critique by saying that "people have preferences for things that harm them. This could happen because they do not know that their choice will have this effect" (Gabriel 419). A "preferences" alignment could lead unimaginable numbers of people awry following false preferences, which Wertheimer describes as first-order preferences incongruent with one's higher-order preferences (Wertheimer 227). AI could envelop an especially irrational and uninformed user in a deep confirmation bias and cage them in a personal epistemology,

much like Plato's allegory of the cave. This could lead to extensive ignorance and voids in knowledge—both antithetical to Aquinas's natural goods.

Additionally, Gabriel points out that "People have preferences about the conduct of other people" and an AI could further entrench a user in their egoist worldview, which could equip that user to cause harm to others (Gabriel 420). Most of the discussion around the future of AI is its capabilities once it gains consciousness and autonomy. While this would have drastic detrimental effects, humans could do equal damage to society and themselves even while this technology does not possess those attributes. Given the decentralized nature of the technology, misguided agents could use this god-like reasoning power for atrocious things, like synthesizing new bioweapons or proliferating propaganda in support of an ill-willed politician. On a less obvious but equally destructive scale, technology this powerful, aligned with fulfilling the desire of every user, could unravel our social fabric. People longing for companionship may seek artificial companionship with their AI. A technology company could profit from this by offering a more intimate and flirtatious product for a higher price, leading to a further loss in the human connection in an era of "synthetic relationships", undermining the natural good of friendship and sociability (Harris and Raskin 10:41).

The second alignment method, based on values defined by the individual or society, would have a far better chance at resulting in human flourishing, as it is predicated on a technological humanism that affirms Kant's Categorical Imperative, along with several Principles of Valid Consent and the Practical Reasoning requirements, although the world would have to cooperate at a much higher level to achieve this goal.

In the eye of a Natural Lawyer, the values guiding this alignment would be rooted in the seven human goods. In a modern context, Gabriel suggests the United Nations' Universal

Declaration of Human Rights (UDHR) as solid grounds (Gabriel 427). Since establishing universal moral principles is one of the oldest debates in philosophy, I am going to assume the UDHR as adequate for the scope of this paper and focus on the benefits to a Natural Lawyer for this alignment.

First, this alignment concurs with Gregory Davis's view of Technological Humanism he asserts in his book "Technology—Humanism or Nihilism" in which he describes technological progress as a means to human flourishing and freedom instead of as the end in itself (Davis 25-26). A technological humanist would promote the responsible development of technology, prioritizing human flourishing over profits. This affirms Kant's Categorical Imperative as treating people as true ends, thus satisfying the principle of "Humanity" in Valid Consent (Wertheimer 127). This alignment could also correct for uninformed users, as the Artificial Intelligence always has the higher-order preference in mind and could recognize false preferences (Wertheimer 227).

Alignment with the UDHR and the natural goods would further root the Artificial Intelligence in a mode of Practical Reasoning. As opposed to the "preferences" alignment, the "values" alignment would have "no arbitrary preferences among values", "a respect for the common good", proper detachment, and a deontological view (Finnis 105-125). The latter two merit an explanation: with a "values" alignment, the AI is able to maintain a certain level of cognitive dissonance through detachment from the potentially emotional desires of the user. The "values" alignment also holds a deontological view, in which good is defined by what is right for society in terms of human rights, whereas the "preferences" alignment holds a utilitarian view in which the right is defined by what is "good" for the user (Finnis 111-125).

A final question remains regarding the nature of society's obligation to adhere to AI to guide decision making in pursuit of human rights and natural goods. Beyond the assumption that the majority of society agrees with the UDHR, what is the source of the "values" alignment's authority? To find it, I will turn to Joseph Raz's normal justification thesis that emphasizes a holistic understanding of a Natural Lawyer's human nature and Robin West's connection thesis in the age of AI.

In the normal justification thesis, Raz asserts the obligation authority creates is such that an individual submits one's own judgment to that of the authority because the values guiding the authority promote the individual's and society's well being. This submission to authority does not mean blind obedience to arbitrary rules, but rather a recognition that these norms are grounded in moral principles and are designed to promote the common good.

Focusing on rationality as the source of the justification of AI's authority would go against the normal justification thesis and Natural Law Theory. AI could reason about the best ways to pursue the natural goods potentially better than any human could, given they can test millions of possibilities and optimize for the best one faster than the human brain can. Therefore, their judgment is based on an incredible reasoning capacity. However, this incredible reasoning capacity, under the normal justification thesis, is not why we should follow the judgment of AI. The normal justification thesis cares about the values guiding the society. The focus of the judgment's authority also is not from the reasoning capacity because sole rationality is no longer congruent with our human experience. We must focus on the ends: how adhering to this judgment would bring us greater connection.

Centering the authority's justification around rationality would also go against our new understanding of our human nature. Aquinas and countless other philosophers throughout history

pointed to our superior reasoning ability as our defining human trait from the rest of the animal kingdom. With the onset of Artificial Intelligence, humans no longer are the most rational entity. In the view of a NL, the law's justification is derived from its basis on morality, and morality is justified by the elements of our human nature and our human experience. Given that rationality is no longer distinct to human nature—or, in Aquinas's terms, "proper to man"—a focus on rationality for the justification of its authority would undermine the holistic uniqueness of being human in the age of AI.

Robin West's connection thesis offers a fitting solution to this problem. The normal justification thesis should focus on the holistic combination of factors that make us human, shifting our view from a separation thesis to a more holistic, humane connection thesis that encompasses a *connection* with each other and with our body. Our *connection* to the body should be celebrated, not shunned; our *connection* to each other should be the ultimate end, not the fear of annihilation. We should adhere to a "values" alignment because society may use the reasoning ability as a means to pursue human flourishing, creating the best chance of fulfilling our "perpetual longing for community, or attachment, or unification, or *connection*" (West 9).

Isn't that the sociological ability of God anyway?

Works Cited

Davis, Gregory H. Technology--Humanism or Nihilism: A Critical Analysis of the Philosophical Basis and Practice of Modern Technology. University Press of America, 1981.

Finnis, John. Natural Law and Natural Rights. Oxford University Press, 2011.

Fridman, Lex. "#371 – Max Tegmark: The Case for Halting AI Development." *The Lex Fridman Podcast*, Spotify, 13 Apr. 2023, https://open.spotify.com/episode/5al9TwC3RihfDqMkyqGte6?si=HFcxlsUlRuGGMuHE4flXZw&dd=1&nd=1.

Gabriel, I. Artificial Intelligence, Values, and Alignment. Minds & Machines 30, 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2

Harris, Tristan, and Aza Raskin. "Synthetic Humanity: Ai & What's at Stake." *Your Undivided Attention*, Spotify, 16 Feb. 2023, https://open.spotify.com/episode/6H3oHGrKdnt23qyhLSFhaI?si=2ZWpXROWRVaQrdzTtWaAag&dd=1&nd=1.

Harris, Tristan, and Aza Raskin. "The A.I. Dilemma ." *The AI Dilemma* , Center for Humane Technology, Youtube, 9 Mar. 2023, https://www.youtube.com/watch?v=xoVJKj8lcNQ.

Olinga, Luc. "Elon Musk Calls for Action against an Imminent AI Threat." *TheStreet*, The Arena Group, 4 Mar. 2023, https://www.thestreet.com/technology/elon-musk-calls-for-action-against-an-imminent-ai-threat.

Popa, E. Human Goals Are Constitutive of Agency in Artificial Intelligence (AI). Philos. Technol. 34, 1731–1750 (2021). https://doi.org/10.1007/s13347-021-00483-2

Aquinas, Thomas, 1225?-1274. The "Summa Theologica" of St. Thomas Aquinas ... London :Burns, Oates & Washburne, ltd., 192042.

Wertheimer, Alan. Consent to Sexual Relations. Cambridge University Press, 2012.

West, Robin. “Jurisprudence and Gender”: Defending a Radical Liberalism.” The University of Chicago Law Review, vol. 75, no. 3, 2008, pp. 985–96. JSTOR, http://www.jstor.org/stable/20141934. Accessed 24 Apr. 2023.

Wu, Tianyue. "Aquinas on Human Personhood and Dignity." The Thomist: A Speculative Quarterly Review, vol. 85 no. 3, 2021, p. 377-409. Project MUSE, doi:10.1353/tho.2021.0024.