Scientific and Social Research, 2026, Volume 8, Issue 10
https://bbwpublisher.com/journal/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Aligning Artificial Intelligence with Human Values: Implications for Society, Democracy, and Human Flourishing

Zixuan Da*

China University of Political Science and Law, Beijing 100091, China

Author to whom correspondence should be addressed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: The growing use of artificial intelligence in consequential decision-making raises questions about how computational systems can reflect human values and support democratic societies. Through a conceptual discussion of AI alignment and an examination of COMPAS in criminal justice and the Colorado Benefits Management System (CBMS) in welfare administration, this article explores the difficulties of translating ethical principles and legal rules into automated decisions. These cases illustrate how algorithmic bias, opaque decision-making, and rigid interpretations of policy can reinforce social inequalities and undermine individual rights. The article argues that alignment requires attention to the intentions, social contexts, and ethical considerations underlying human instructions. It examines the limitations of relying on expressed intentions, revealed preferences, and informed preferences to guide AI behavior, and emphasizes the need to translate abstract values into operational standards. It further advocates governance frameworks, ongoing oversight, and participation by technical experts, policymakers, and affected communities. Aligning AI with human values is both a technical and an ethical undertaking, essential to protecting fairness, supporting democratic principles, and promoting human flourishing.

Keywords: AI alignment; Human values; Automated decision-making; Algorithmic fairness; AI governance

Online publication: November 25, 2026

1. Introduction

Today, AI technology has increasingly permeated all aspects of life. The development and growth of AI raises important new questions for technicians, humans, and sensory life more generally. Foremost among these is the question of what or whose values AI systems should follow [1]. That is called "AI alignment."

As mentioned in Brian Christian's book "The Alignment Problem: How Can Machines Learn Human Values?," machine learning and AI systems are becoming more and more commonplace and powerful, but it is like a "sorcerer's apprenticeship" [2]. People design AI systems to perform tasks, give them commands

like a “wizard” summoning his powers, and the AI system follows the commands to perform the task obediently [2]. But if one’s commands are vague or incomplete, the AI system may execute them precisely in a way that people cannot anticipate and that may have negative consequences. Therefore, it is important to ensure that AI is aligned with people’s values and behaves in a way that upholds people’s principles and ideals in accordance with their wishes. Many organizations have issued policy documents that attempt to clarify the ethical principles that should be adhered to when designing AI systems [3]. The EU has identified four principles for the ethical use of AI: non-maliciousness, respect for autonomy, algorithmic fairness, and transparency [4]. Based on a review of 84 policy documents, Jobin et al. argue that there is an impending global convergence around the following five principles of ethics in the use of AI: transparency, justice and fairness, non-maleficence, responsibility, and privacy [5].

This paper will explore the importance of bringing AI in line with human values for democracy and social prosperity, showing that the potential of AI should be used to compensate for human vulnerabilities, not to create more problems for humans. The paper uses two examples (COMPAS and CBMS) to illustrate the value alignment challenge and the importance of AI alignment. Additionally, the paper will discuss methods for addressing these challenges, thereby ensuring that AI acts as an empowering tool for humanity, steering people towards a more equitable future.

2. AI alignment and human values

2.1. The importance of AI alignment

In exploring the intersection of AI and human values, people need to start with the understanding that machine learning is the driving force behind many AI systems, which is categorized into three different types of learning. Unsupervised learning, where AI detects patterns in a vast trove of data without guidance; supervised learning, where it learns from examples with known outcomes to make future predictions; and reinforcement learning, where AI makes decisions based on a system of rewards and penalties to achieve its goals [2].

In these three areas, more and more places are being managed to some extent by these mathematical and computational models. For example, there is a growing demand in the legal field for the use of risk assessment software to determine whether to grant bail or parole. Self-driving vehicles are becoming more common on the motorways and streets. In many cases, people’s mortgage applications, CVs, or medical test results no longer need to be manually reviewed to reach a straightforward conclusion. These phenomena are indicative of mankind’s gradual process of placing the world in automation [2].

This growing dependence on AI brings to light the potential issues inherent in entrusting machine learning with tasks that carry significant ethical weight. If a facial recognition system is grossly inaccurate in identifying people of a certain race or gender, or if someone is denied bail because of a statistical model that even judges, lawyers, and defendants do not understand how it operates, this is a problem. It is not just a technical issue, but also a social and ethical one that requires experts from different fields to come together in dialogue and decision-making to ensure that the development of technology can be reconciled with human values and social justice [2].

2.2. Examples of AI alignment in practice

2.2.1. COMPAS

An example of AI alignment is COMPAS. In practice, an increasing number of judges in the judiciary rely

on “risk assessment” tools, which are used by judges to determine whether a defendant should be detained or released. One of the most popular risk assessment tools is COMPAS, developed by Northpointe. It is an automated risk and needs assessment tool (an alternative to the sanctioned Correctional Offender Management Profile) for use in the criminal justice system, designed to inform decision-making, offender management, sentencing, parole, and reintegration programs. COMPAS is rated on a scale of 1 to 10, but its problem is that everyone in the justice system does not know exactly how the COMPAS model works [2].

Although COMPAS was initially lauded and widely used for its technological advancements, it continues to face significant controversy. The ProPublica team conducted an investigation in 2016 that exposed the bias issues with COMPAS. In this investigation, the ProPublica team looked at a large number of defendants’ criminal records and risk scores, and they found that the COMPAS risk assessment was racially biased against African-American offenders, who were more likely to be judged by COMPAS as being at a higher risk of recidivism compared to white offenders [3]. For example, two defendants were both charged with drug offenses, but one of them, Dylan Fugett, who is white, had a risk score of 3, while Bernard Packer, a black defendant, had a risk score of 10. However, their criminal records later showed that Fugett, who had a lower risk of recidivism, was subsequently convicted of drug-related offenses, while Packer, who had an extremely high-risk score, did not reoffend. In addition, there were two defendants who were also charged with petit larceny; the “low-risk” defendant subsequently committed aggravated burglary, while the “high-risk” defendant had no criminal record. Unsurprisingly, the “low-risk” rating remained white, while the “high-risk” rating was black [2].

The ProPublica team’s research has sparked a broader debate about the fairness and transparency of using algorithmic decision-making in the criminal justice system, not just limited to COMPAS, but provoking thought about “how principles in law (e.g., fairness, justice) can be defined in computational terms” [2]. The principles, rights, and concepts of law are often abstract and subjective and not easily quantifiable. Statistical and computational methods, on the other hand, rely on concrete, quantifiable data to operate. In addition, human intentions and needs are also complex and ambiguous, and are not easily and completely expressed in simple instructions or code. For example, people may want AI to be both efficient and fair in delivering services, but the concept of “fairness” may have different understandings and standards in different cultures and contexts. How to encode this abstract and multidimensional concept in AI systems is a technical and ethical challenge.

2.2.2. CBMS

Another example of a failed automated system is CBMS. In September 2004, the State of Colorado implemented a large-scale information system called the Colorado Benefits Management System (CBMS). The purpose was to replace six aging systems with a new, identical system to support the various state-administered benefit programs [6]. It can automate eligibility determinations for Medicaid, food stamps, and other welfare programs [7]. The original intent of the CBMS system was to provide better service to people and to ensure that the state’s benefit programs were properly administered [6].

But in practice, CBMS makes a lot of bad decisions. For example, in Colorado, a food stamp recipient is experiencing homelessness due to the loss of his flat. She explained her situation to the eligibility worker that she had initially stayed at her sister’s house for a fortnight, but was now confined to the streets. Under federal law, people who are found to be homeless may be eligible for certain accommodations and may also receive

increased food stamp subsidies on that basis. Even if they are temporarily housed somewhere, they can still be considered "homeless" as long as they have not been homeless for more than 90 days [8].

However, when the worker entered the homeless person's information into CBMS, the system required the worker to select her residential status from the options given and required the worker to enter whether or not she was a "panhandler." Based on the staff member's input, the system ultimately determined that she could not be considered "homeless" and determined her food stamp amount accordingly. The problem is that CBMS is a "black box," meaning that the internal workings of the system are completely opaque [7]. The denial of the determination may have been due to CBMS's misinterpretation of her living with her sister. Due to a coding error, the system excludes this type of transient situation from "homelessness," but this is a violation of established policy.

It is also possible that the reason for the refusal is that the system, based on the staff's answer to the question of whether they are "panhandlers," has created its own new rule: it must first determine whether the individual receiving the food stamps is a "panhandler." However, the definition of "homelessness" in federal law does not require a judgment as to whether an individual is a beggar [8]. Whatever the reason, food stamp recipients are deprived of their property rights, and CBMS's automated decision-making fails to accurately reflect the definition of "homelessness," leaving individuals who should be receiving additional benefits without the help they deserve. This demonstrates the problematic nature of AI in understanding and executing instructions on complex human values and social policies.

Both examples, COMPAS and CBMS, illustrate the complexities and pitfalls of automated decision-making systems when they are not aligned with human values and social policies. These examples illustrate that AI must be careful to align with human values and that people need to give AI a deeper understanding of the spirit behind the rules. Reducing AI bias ensures that the technology serves society in a way that is consistent with human values and avoids exacerbating social inequality. The potential of AI should be used to protect society and contribute to the flourishing of democracy, not to undermine the foundations of democracy and society.

3. Strategies for achieving AI alignment with human values

In the early days of AI research, Norbert Wiener emphasized the importance of ensuring that the machine's purpose is aligned with people's [1]. More recently, the Asilomar Principles of Artificial Intelligence argue that highly autonomous AI systems should be designed to ensure that their goals and behaviors are consistent with human values throughout their operation [9]. However, it remains a challenge to articulate these values in a way that is operable by the AI. This requires distinguishing between merely following instructions and conforming to the intentions and ethical standards of those instructions.

A significant risk in the development of AI is the possibility that the system will carry out commands too literally. In this case, the AI carries out the commands precisely, but the results are contrary to the intended outcome. To mitigate this risk, AI systems must be designed to understand not only the literal commands, but also the larger context in which they are issued. This requires the development of AI with the ability to process and evaluate the intent behind human commands so that it can act in ways that truly reflect human goals. Complicating the adjustment process, however, is the distinction between expressed intent and revealed or informed preference. Ideally, an AI system should understand not only the explicit instructions given by the user, but also the implicit preferences and intentions that guide human behavior. This can be

achieved either by observing human interactions and behavior (revealed preferences) or by modeling human desires under conditions of full information and rationality (informed preferences) [1]. However, there are challenges with each approach. Revealed preferences may not always accurately reflect true intentions due to misinterpretation of behavior, while informed preferences require the AI to make judgments about what humans would want under conditions of full information and complete rationality, which involves complex ethical and philosophical considerations [1].

To effectively operationalize these concepts, AI must be programmed to interpret complex human values, translating these abstractions into concrete computational terms. Developing algorithms that understand and realize these values is essential to prevent bias and ensure that AI systems operate in an ethical and fair manner.

Finally, robust governance and oversight mechanisms are essential to ensure that AI systems are consistent with human values in the long term. Establish a regulatory framework that sets clear guidelines and standards for the development of AI. It is also crucial to involve different perspectives in the development and management of AI. This should include not only technicians and ethicists, but also policymakers, community representatives, and the wider public to ensure that AI systems are developed in a way that is responsive to the needs and values of the societies they serve.

4. Conclusion

In conclusion, as AI systems increasingly influence various aspects of people's lives and societal functions, the imperative of aligning these systems with human values becomes more pronounced. The examples of COMPAS and CBMS reveal the complexities and potential pitfalls of automated decision-making when it is not sufficiently attuned to ethical standards and social policies. These cases demonstrate that AI, if not properly aligned, can perpetuate biases and make decisions that are contrary to established human rights and societal norms. Hence, the pursuit of AI alignment is not merely a technical challenge but a profound ethical obligation. People must ensure that AI systems do not simply execute tasks based on superficial instructions but engage with the deeper ethical implications of their actions. As people continue to integrate AI into more critical areas of human interaction and governance, developing strategies that promote understanding and implementation of human values in AI will be crucial. This alignment will safeguard against the erosion of democratic principles and contribute positively to societal prosperity, thereby making AI a truly empowering tool for humanity.

Disclosure statement

The author declares no conflict of interest.

References

[1] Gabriel I, 2020, Artificial Intelligence, Values, and Alignment. Minds & Machines, 2020(30): 411–437.

[2] Christian B, 2021, The Alignment Problem: How Can Machines Learn Human Values? Atlantic Books, London.

[3] Peterson M, Gärdenfors P, 2024, How to Measure Value Alignment in AI. AI and Ethics, 4(4): 1493–1506.

[4] European Commission, 2019, High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI.

[5] Jobin A, Ienca M, Vayena E, 2019, The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 2019(1): 389–399.

[6] Fukami C, McCubbrey DJ, 2011, Colorado Benefits Management System (C): Seven Years of Failure. Communications of the Association for Information Systems, 29(1): 97–102.

[7] Citron DK, 2007, Technological Due Process. Washington University Law Review, 2007(85): 1249.

[8] United States Code, 2012, 7 USC 2012: Definitions. https://uscode.house.gov/view.xhtml?req=(title:7%20section:2012%20edition:prelim)

[9] Future of Life Institute, 2024, Asilomar AI Principles. https://futureoflife.org/open-letter/ai-principles/

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.