-- Hekmat va Falsafeh------
Vol. 22, No.85, Spring 2026, P 151-186
wph.atu.ac.ir
DOI: 10.22054/wph.2026.88857.2350

ATU PRESS

Original Research
Published Online: June 15, 2026
Accepted: April 15, 2026
Revised: January 17, 2026
Received: October 13, 2025
eISSN: 2476-6038
ISSN: 1735-3238

Value-Sensitive Design and Value Alignment: A Pragmatist Sociotechnical System Perspective

Mohammad Ali Ashouri Kisomi *

Department of Philosophy, Faculty of Persian Literature & Foreign Languages, Allameh Tabataba'i University, Tehran, Iran

Extended Abstract

1. Introduction

With the rapid development of artificial intelligence (AI), aligning this technology with human values has become increasingly essential. Value-Sensitive Design (VSD) has emerged as a practical approach to value alignment, attracting considerable scholarly attention. This article evaluates VSD as a method for aligning AI with social values, drawing on the pragmatist account of sociotechnical systems as the evaluative framework.

To avoid the unintended consequences of technology and to align AI systems with human values, several approaches have been proposed. In line with the purpose of this paper, we focus on the value-sensitive design approach.

The principal idea behind VSD is to account for values in the engineering process and embed value reflection into the design. In terms of methodology, VSD is based on a tripartite structure that consists of conceptual, empirical, and technical investigations (Friedman & Hendry, 2019).

The conceptual investigation involves mapping stakeholders, identifying values, and resolving value tensions. Empirical investigation deploys exploratory tools to validate values and address value tensions in particular contexts. The technical investigation

* Corresponding Author: m_ashori@atu.ac.ir

How to Cite: Ashouri Kisomi, M. A. (2026). Value-Sensitive Design and Value Alignment: A Pragmatist Sociotechnical System Perspective, *Hekmat va Falsafeh*, 22(85), 151-186. doi: 10.22054/wph.2026.88857.2350

includes proactive designing of the system to meet identified values in the conceptual phase and focuses on translating values into design requirements and technology affordances (Umbrello & Van de Poel, 2021). VSD tripartite investigations can be carried out consecutively, in parallel, or iteratively. However, it has been advocated for integrative and iterative investigation in the full lifecycle of the system (Friedman & Hendry, 2019), which is essential for maintaining human control in the incidents of unintended consequences.

In general, the fundamental premise of VSD is recognizing the sociotechnical nature of technology. VSD adopts an interactional stance, recognizing the mutual interaction between technology and human agents (Friedman et al., 2017).

In order to address ethical issues of AI, an approach would be to conceive of AI systems as sociotechnical systems. In other words, taking technology as part of society, rather than as a designed artifact in a timeless void. As the EU-U.S. Roadmap for Terminology and Taxonomy for Artificial Intelligence states, 'technology is not only design and material appearance but also sociotechnical; that is, a complex process constituted by diverse social, political, economic, cultural, and technological factors' (European Commission, 2023).

In a broad sense, pragmatism can be conceived as the philosophical stance that knowing the world is inseparable from agency within it (Legg & Hookway, 2008). Pragmatism foregrounds human agency and underscores that knowledge about a sociotechnical system cannot be separated from the capacity to act within it. Pragmatism focuses on the role of action and human agency. In other words, the future of technological advancement is, to some degree, shaped by human agents and is indeterminate (Watson et al., 2024). While technological shortcomings limit our design options, there is room for human agents' decisions among existing choices, pursuing for a new innovation or limiting one. Moreover, pragmatism highlights that human and social elements like governance, institutions, and culture are crucial in the operation and function of sociotechnical systems (Kudina & Van de Poel, 2024), and technology is not the only effective element.

Pragmatism supports the idea of progress (Dewey, 2004), meliorism (James, 1907; Dewey, 2004) and fallibilism (Peirce, 1955) while holding that the validity of theories depends on their usefulness when put into practice within the environment (James, 1907), i.e.,

specific context. In this way, pragmatism does not advocate for restriction of AI systems; it aims for progress rather than perfection, accounts for uncertainty, and posits that evaluation of such systems should be context-specific and reflect the normative goals of the society (Watson et al., 2024). Pragmatic meliorism is neither pessimistic nor optimistic about the future of technology; rather, it emphasizes that this future is not predetermined or given but can be improved through the efforts of human agents.

The introduction of some technologies into society comes with a range of uncertainties about their social effects (Collingridge, 1980). However, several unintended factors can contribute to this uncertainty. For instance, a lack of foresight and anticipation of the effects of technology on society can lead to unintended consequences. Additionally, human epistemic capacity is limited, and not only lack of due care but also epistemic ignorance can contribute to unintended effects. Furthermore, there are several agents and factors in a society that are out of our control in designing an AI system, and considering some level of indeterminacy is reasonable (Van de Poel 2023); therefore, the introduction of technology into society can be conceived as a social experiment. In other words, only by experiment and gradually (Van de Poel, 2016) by interdisciplinary assessment (Ashouri Kisomi et al., 2025) we can uncover some of the social effects of technology. Therefore, it invites us for a posteriori evaluation and investigation.

However, when taken as an approach to inquiry and grounded in both pragmatic fallibilism and meliorism, pragmatism also advocates for a priori investigation and invites us to incorporate critical examination in advance of social experimentation (Rydenfelt, 2025). Overall, adopting a sociotechnical pragmatism perspective brings several advantages. Particularly, it provides a better understanding of how AI systems function and affect society and vice versa.

2. Literature Review

Over the years, VSD faced several criticisms. For instance, reliance on heuristic values (Reijers & Gordijn, 2019), lack of a methodological account of value distinction from preferences (Manders-Huits, 2011), not adhering to a specific ethical theory (Albrechtslund, 2007; Jacobs & Huldtgren, 2021), and the risk of naturalistic fallacy (Manders-Huits, 2011; Jacobs & Huldtgren, 2021) are among the criticisms of VSD. However, these criticisms have not been overlooked, and

several solutions and modified VSD approaches have been proposed for different contexts. Nevetheless, it is crucial to evaluate this framework periodically. This article evaluates VSD, drawing on the pragmatist account of sociotechnical systems as the evaluative framework.

3. Methodology

The primary objective of this article is to conduct a comprehensive evaluation of VSD (Value Sensitive Design) as a practical methodological instrument for effectively aligning AI (Artificial Intelligence) systems with ethical and social values. To underpin this assessment, the study explicitly draws upon and operationalizes the pragmatist account of the sociotechnical systems perspective, which is established as the core theoretical and evaluative framework that guides the entire analytical inquiry. In terms of procedural execution, the investigation employs an analytical-critical method. By applying this methodological approach, the study thoroughly examines both the principal strengths and the inherent challenges presented by VSD.

4. Conclusion

Our findings show that VSD offers a robust framework for value alignment. Its strengths include providing a practical pathway for embedding social values into technology, incorporating stakeholder perspectives, recognizing human agency, and addressing the full life cycle of AI systems. Nevertheless, VSD faces significant challenges, such as the protracted nature of alignment processes, difficulties in ensuring meaningful human control, and the lack of enforceable international laws and regulations. Our analysis suggests that treating alignment methods merely as self-regulatory tools may produce long-term negative consequences. To fully realize the benefits of aligning AI with human values, it is therefore crucial to develop enforceable international frameworks that complement and support local laws and governance structures.

Keywords: Value Alignment, Value-Sensitive Design, Ethics of AI, Sociotechnical Systems, Pragmatism.

مقاله پژوهشی
تاریخ ارسال: ۱۴۰۴/۰۷/۲۱ تاریخ بازنگری: ۱۴۰۴/۱۰/۲۷ تاریخ پذیرش: ۱۴۰۵/۰۱/۲۶ تاریخ انتشار: ۱۴۰۵/۰۳/۲۵
ISSN: 1735-3238 eISSN: 2476-6038

فصلنامه علمی حکمت و فلسفه
سال بیست و دوم، شماره ۸۵، بهار ۱۴۰۵، ص ۱۵۱ تا ۱۸۶
wph.atu.ac.ir
DOI: 10.22054/wph.2026.88857.2350

طراحی حساس به ارزش و هم‌سوسازی هوش مصنوعی: روایتی پراگماتیستی از سیستم‌های اجتماعی‌فنی

محمد علی عاشوری کیسمی *
گروه فلسفه، دانشکده ادبیات فارسی و زبان های خارجی، دانشگاه علامه طباطبائی، تهران، ایران

چکیده

با توسعه سریع هوش مصنوعی، هم‌سوسازی این فناوری با ارزش‌ها امری مهم تلقی می‌شود. «طراحی حساس به ارزش» شیوه‌ای عملی برای هم‌سوسازی فناوری است که توانسته نظر بسیاری از پژوهشگران را جلب کند. هدف از این مقاله ارزیابی شیوه «طراحی حساس به ارزش» برای هم‌سوسازی با ارزش‌های اجتماعی است. در این راستا از روایت پراگماتیستی سیستم‌های اجتماعی‌فنی به عنوان معیار ارزیابی بهره گرفته شده است. برای رسیدن به هدف پژوهش، از شیوه تحلیلی–انتقادی استفاده شده و مزایا و چالش‌های این شیوه ارزیابی شدند. نتایج نشان می‌دهد «طراحی حساس به ارزش» چارچوب توانمندی برای هم‌سوسازی هوش مصنوعی به دست می‌دهد. مهم‌ترین نقاط قوت این شیوه شامل ارائه راه‌کار عملی برای هم‌سوسازی فناوری با ارزش‌ها، توجه به ذی‌نفعان، عاملیت انسانی و در نظر گرفتن کل چرخه حیات فناوری است. چالش‌های این شیوه شامل مواردی همچون روند طولانی هم‌سوسازی، کنترل معنادار انسانی و مشکلات فقدان قوانین و مقررات بین‌المللی با ضمانت اجرایی می‌شود. بررسی‌های نشان می‌دهد چنانچه شیوه‌های هم‌سوسازی به‌عنوان روش خودتنظیمی در نظر گرفته شوند، در طول زمان می‌توانند آثار منفی به بار آورند. همچنین برای بهره‌گیری از مزایای هم‌سوسازی با ارزش‌های انسانی، تلاش برای رسیدن به قوانین و مقررات بین‌المللی با ضمانت اجرایی که از قوانین و حکمرانی محلی پشتیبانی کند ضروری است.

کلیدواژه‌ها: هم‌سوسازی ارزش، طراحی حساس به ارزش، اخلاق هوش مصنوعی، سیستم‌های اجتماعی‌فنی، پراگماتیسم.

* نویسنده مسئول: M_ashori@atu.ac.ir

۱. مقدمه

شتاب توسعه هوش مصنوعی[1] همراه با نگرانی در خصوص توسعه و استفاده نامطلوب از این فناوری، موجب توجه پژوهشگران به هم‌سوسازی آن با ارزش‌ها شده است (e.g., Nyholm, 2023; Corrêa, 2024). در ادبیات پژوهشی، شاهد پژوهش‌های بسیاری در خصوص مباحث اخلاقی و نگرانی از توسعه و استفاده از این فناوری بر خلاف ارزش‌های انسانی هستیم[2]. چندین فرامطالعه به بررسی این ادبیات پرداخته و نگرانی‌های مختلف شناسایی شده‌اند (e.g., Floridi 2023; Hagendorff 2020; Corrêa et al. 2023). با نظر به این نگرانی‌ها، هدف اصلی هم‌سوسازی، طراحی و استفاده از هوش مصنوعی با حصول اطمینان از هم‌سویی آن با ارزش‌های انسانی[3] است (Russell, 2019: 137).

در خصوص هم‌سوسازی با ارزش‌ها[4]، دیدگاه‌های مختلفی وجود دارد. برخی معتقدند هوش مصنوعی یک فناوری ارزش‌بار[5] است و باید با ارزش‌های اخلاقی و اجتماعی هم‌سو شود (Van den Hoven et al., 2015)، برخی دیگر بر ارزش‌های اقتصادی تاکید کرده و آن را هدف هم‌سوسازی می‌دانند (Andreessen, 2023)؛ در مقابل دیگرانی این فناوری را غیرارزش‌بار[6] می‌دانند (Pitt, 2013) و یا برخی معتقدند اگرچه ارزش‌بار است، در عمل بر خلاف ارزش‌های انسانی طراحی و استفاده می‌شود (Zuboff, 2023). اگر بپذیریم هوش مصنوعی یک فناوری ارزش‌بار است و طراحی و استفاده از آن می‌توانند آثار و نتایج اخلاقی و اجتماعی داشته باشند، توجه به شیوه‌های هم‌سوسازی اهمیت پیدا خواهد کرد. این شیوه‌ها اگر بتوانند از پس چالش‌های پیش‌رو با

---

1 Artificial Intelligence (AI)

۲در این خصوص پژوهش‌های مختلفی چه به زبان فارسی (همچون عباسی و تیموری، ۱۴۰۲؛ الستی، ۱۴۰۳؛ حسینی سورکی، ۱۴۰۳) و چه در نمونه‌های بین‌المللی صورت گرفته است (e.g., Zhang and Zhang, 2023; Zhou and Nabus, 2023).

3 Human values

4 Value alignment

5 Value-laden

6 Value-free

موفقیت عبور کنند، راه‌کارهای عملی برای طراحی فناوری بر اساس ارزش‌های انسانی خواهند بود.

شیوه‌های گوناگونی برای هم‌سوسازی سیستم‌های هوش مصنوعی پیشنهاد شده است. از جمله این شیوه‌ها می‌توان به طراحی «عامل اخلاقی هوش مصنوعی[1]» (Wallach & Allen 2009)، «هوش ترکیبی[2]» (Akata et al. 2020) و «طراحی حساس به ارزش[3]» (Friedman & Hendry, 2019) اشاره کرد. در میان این روش‌ها، مقالات مختلفی از جهات گوناگون به بررسی شیوه «طراحی حساس به ارزش» پرداخته‌اند (e.g., Shen et al., 2025; Umbrello & De Bellis, 2018; Howcroft & Vale, 2024) که نشان دهنده اهمیت این شیوه هم‌سوسازی است. مقاله حاضر نیز به بررسی این شیوه تمرکز خواهد کرد. اما یک پرسش اصلی این است که با چه رویکردی می‌خواهیم این بررسی را انجام دهیم که آن‌را با پژوهش‌های پیشین متفاوت می‌کند؟

پژوهش‌های مختلفی به شرح و تفسیر رویکرد سیستم‌های اجتماعی‌فنی[4] پرداخته‌اند (e.g., Watson et al., 2024; Young & Coeckelbergh, 2024; Kudina & Van de Poel, 2024). بر اساس این رویکرد، سیستم هوش مصنوعی بخشی از سیستم بزرگ‌تر اجتماعی‌فنی است. در سیستم اجتماعی‌فنی انواع عوامل اقتصادی، اجتماعی، سیاسی، فرهنگی و غیره وجود دارند که هر یک بر کل سیستم و اجزای آن اثرگذار هستند. این عوامل با درجات مختلف در طراحی و استفاده از فناوری اثرگذار بوده و در هم‌سوسازی هوش مصنوعی باید به نقش و تاثیر آن‌ها توجه کنیم (European Commission, 2023). یک نقطه قوت اصلی این رویکرد این است که به جای بررسی فناوری در خلا، زمینه اجتماعی آن را مورد توجه قرار می‌دهد. البته روایت‌های مختلفی از رویکرد اجتماعی‌فنی وجود دارد (Watson et al., 2024)، اما در مقاله حاضر از روایت پراگماتیستی استفاده

1 AI moral agent
2 Hybrid intelligence (HI)
3 Value sensitive design (VSD)
4 Sociotechnical system (STS)

خواهد شد[۱].

با وجود پژوهش‌های مختلفی که به بررسی «طراحی حساس به ارزش» پرداخته‌اند، خلا پژوهش‌هایی که به بررسی هم‌سوسازی در شیوه «طراحی حساس به ارزش» با تکیه بر روایت پراگماتیستی سیستم‌های اجتماعی‌فنی بپردازند را می‌توان مشاهده کرد. لذا، هدف از مقاله حاضر بررسی شیوه «طراحی حساس به ارزش» برای هم‌سوسازی با ارزش‌ها، با تکیه بر روایت پراگماتیستی[۲] سیستم‌های اجتماعی‌فنی است.

برای رسیدن به هدف اصلی مقاله، با استفاده از روش تحلیلی-انتقادی، مزایا و چالش‌های شیوه «طراحی حساس به ارزش» را ارزیابی خواهیم کرد. بر این اساس، ادامه مقاله به چهار بخش تقسیم شده است. در بخش اول ضرورت، مقصود و هدف از هم‌سوسازی بیان می‌شود. در بخش دوم، روایت پراگماتیستی سیستم‌های اجتماعی‌فنی معرفی می‌شود تا شیوه‌ای جهتِ ارزیابی داشته باشیم. در بخش سوم شیوه «طراحی حساس به ارزش» معرفی خواهد شد. در قسمت پایانی نیز «طراحی حساس به ارزش» با توجه به رویکرد سیستم‌های اجتماعی‌فنی ارزیابی می‌شود.

۲. هم‌سوسازی ارزش و هوش مصنوعی

پیش پرداختن به ضرورت و هدف از هم‌سوسازی، باید بدانیم آیا هوش مصنوعی یک فناوری ارزش‌بار است؟

۲-۱. ارزش‌باری و هم‌سوسازی ارزش در هوش مصنوعی

زمانی می‌توان از هم‌سوسازی هوش مصنوعی با ارزش‌ها سخن گفت که آن را ارزش‌بار بدانیم؛ چرا که در غیر این‌صورت هم‌سوسازی بی‌معنا خواهد بود[۳]. در این خصوص،

---

۱ برای مقایسه روایت‌های مختلف از سیستم‌های اجتماعی‌فنی رجوع کنید به (Watson et al., 2024).

2 Pragmatism

۳اگر فناوری را ارزش‌بار ندانیم، دیگر نگرانی در خصوص توسعه نامطلوب آن نخواهیم داشت و این عوامل انسانی هستند که باید از آن استفاده اخلاقی کنند و یا استفاده از فناوری را کنار بگذارند.

می‌توان سه دیدگاه اصلی را از یکدیگر تمیز داد[۱].

طرفداران دیدگاه ابزارگرایی[۲]، فناوری را غیر ارزش‌بار می‌دانند. آن‌ها معتقدند استفاده انسان از ابزار در راستای اهداف مختلف قرار می‌گیرد و فناوری را در برابر ارزش‌ها خنثی می‌دانند (Pitt, 2013; Peterson & Spahn, 2011). اعتراضات گوناگونی به این دیدگاه وارد شده است[۳]. به‌عنوان مثال، این‌که مهندسان در طراحی فناوری با انتخاب‌های مختلفی روبرو هستند از منظر این دیدگاه مغفول مانده[۴] و بر این امر تاکید می‌کند که ارزشی در فناوری تعبیه نمی‌شود.

بر اساس دیدگاه ذات‌گرایی[۵]، فناوری ذاتا ارزش‌بار بوده و ارزش‌های مشخصی را حمل می‌کند. این دیدگاه بر عاملیت انسانی در طراحی و استفاده از فناوری چشم می‌پوشد (Srivatsa et al., 2017: 110). مطابق این دیدگاه، تفاوتی میان شیوه‌های مختلف طراحی و استفاده از فناوری وجود ندارد، چراکه ارزش‌ها در ذات فناوری هستند. حال آنکه فناوری را می‌توان به شیوه‌های مختلف طراحی و برای استفاده‌های مختلف به‌کار برد؛ ولی ذات‌گرایی توضیحی در این خصوص ارائه نمی‌دهد[۶].

بر اساس دو دیدگاه پیشین، هم‌سوسازی فناوری امری بی‌اثر یا ناممکن خواهد بود.

---

۱ از آنجا که هدف اصلی مقاله حاضر بررسی یکی از شیوه‌های پیشنهادی برای هم‌سوسازی فناوری و پیش‌فرض آن ارزش‌بار بودن و امکان‌پذیر بودن هم‌سوسازی است، در این بخش دیدگاهی کلی از این مباحث به خواننده ارائه می‌شود. برای مطالعه دسته‌بندی‌ها و نظرات مختلف در خصوص ارزش‌باری فناوری رجوع کنید به (Kores & Verbeek, 2014) و در خصوص امکان‌پذیر بودن هم‌سوسازی رجوع کنید به (Gabriel, 2020; Gabriel & Ghazavi, 2022).

2 Instrumentalism

۳ برای مطالعه در خصوص انتقادات رجوع کنید به (Miller, 2021; Srivatsa et al., 2017).

۴ به عنوان مثال، حفاظت از حریم خصوصی را در نظر بگیرید. طراحان یک سیستم ممکن است این ارزش را در طراحی خود لحاظ کنند و یا در طراحی هیچ توجهی به حفظ حریم خصوصی نداشته باشند. این دیدگاه، تاثیر توجه و عدم توجه مهندسان به حفاظت از حریم خصوصی در طراحی را در نظر نمی‌گیرد.

5 Substantivism

۶ هر دو رویکرد پیشین با مخالفت‌های بسیاری مواجه شده‌اند. برای اطلاعات بیشتر رجوع کنید به (Van de Poel & Kroes 2014; Pareto & Coeckelbergh 2024).

بر اساس ابزارگرایی ارزشی در مصنوعات تعبیه نمی‌شود که بخواهیم فناوری را با آن هم‌سو کنیم. ذات‌گرایی نیز اراده و عاملیت انسانی را در برابر فناوری نفی می‌کند که منجر به نفی مسئولیت طراح برای هم‌سوسازی و کاربر برای استفاده هم‌سو با ارزش‌ها از فناوری می‌شود.

در مقابل دیدگاه‌های پیشین، دیدگاه تعامل‌گرایانه[1] به ارزش‌بار بودن فناوری، اهمیت انتخاب‌ها و تاثیرگذاری استفاده و طراحی فناوری توسط عوامل انسانی توجه دارد (ibid: 111). زمانی که از هم‌سوسازی فناوری با ارزش‌ها سخن می‌گوییم، به‌صورت ضمنی بر رویکردی تعامل‌گرایانه تکیه می‌کنیم. در حقیقت هم هوش مصنوعی را ارزش‌بار می‌دانیم که ظرفیت قرار گرفتن ارزش‌ها را در خود دارد و هم طراح و کاربر را دارای عاملیت و مسئولیت می‌دانیم که قادر به هم‌سوسازی فناوری و استفاده هم‌سو با ارزش‌ها هستند.

۲-۲. ضرورت هم‌سوسازی هوش مصنوعی با ارزش‌ها

اگر هم‌سوسازی را امکان‌پذیر بدانیم، باید به این پرسش پاسخ دهیم چه ضرورتی برای هم‌سوسازی هوش مصنوعی وجود دارد؟ برای دسته‌ای از پژوهشگران نگرانی‌هایی همچون خطرات وجودی هوش مصنوعی فوق هوشمند[2] (Bostrom, 2014) علت توجه به هم‌سوسازی است. رسیدن به چنین مرحله‌ای از هوشمندی حداقل به دلایل فنی هنوز امکان‌پذیر نیست (Van de Poel, 2023; 118)؛ همچنین اگر هوش مصنوعی فوق هوشمند یا عام[3] امکان‌پذیر باشد، هم‌سویی چنین سیستمی با ارزش‌های انسانی، خود محل بحث است (عاشوری کیسمی و پرویزی، ۱۴۰۳؛ حسینی سورکی، ۱۴۰۳). با این‌حال این موضوعات سبب نشده که نگرانی‌ها از خطرات و مشکلات ناشی از هوش مصنوعی برطرف شود. نگارش و وضع ده‌ها کد اخلاقی، استاندارد، بیانیه، قوانین و مقرارت و

---

1 Interactionism

۲ Superintelligent artificial intelligent: سیستم‌های هوش مصنوعی که بتواند در تمامی زمینه‌ها از جمله عملکرد و شناخت به سطح هوش انسانی رسیده و از آن فراتر رود (عاشوری کیسمی و پرویزی، ۱۴۰۳: ۱۴۶).

3 Artificial General Intelligence (AGI)

پیشنهادات گوناگون (e.g., European Commission, 2021; UNICEF, 2021; Werthner, 2020) خود مؤید این امر است که نگرانی در خصوص توسعه نامطلوب این فناوری چشم‌گیر است. در این راستا، اگر بپذیریم فناوری ارزش‌بار است و همچنین طراحی و استفاده از هوش مصنوعی هر دو تاثیرگذار هستند، آنگاه برای اینکه این فناوری بر خلاف یا متعارض با ارزش‌ها طراحی و استفاده نشود، هم‌سوسازی هوش مصنوعی ضرورت پیدا می‌کند. البته شاید سخن گفتن از ارزش‌ها مبهم و کلی به نظر برسد. لذا پرسش بعدی این است که کدام ارزش‌ها در هم‌سوسازی باید مورد توجه قرار بگیرند؟

۳-۲. هم‌سوسازی با ارزش‌های اجتماعی یا فردی

در خصوص انتخاب ارزش‌ها برای هم‌سوسازی، اختلاف نظرهای فراوانی وجود دارد. به‌عنوان مثال، برخی از پژوهشگران ارضای تمایلات[1] انسانی توسط فناوری را هدف هم‌سوسازی می‌دانند (Wiener, 1960: 1358). برخی دیگر معتقدند سیستم‌های «فوقِ هوشمند[2]» باید به گونه‌ای طراحی شوند که هدف و رفتار آن‌ها در عمل، مطابق با اهداف انسانی باشد (Asilomar, 2018). در برابر دیدگاه اول می‌توان اعتراض کرد که تمایلات انسان‌ها ممکن است غیراخلاقی باشند و حتی موجب آسیب به فرد یا جامعه شود؛ لذا تمایلات نمی‌تواند هدف از هم‌سوسازی فناوری باشد. رویکرد دوم نیز تنها به سیستم‌های «فوقِ هوشمند» توجه دارد؛ در حالی که سیستم‌هایی با قابلیت‌های محدودتر هم ممکن است برخلاف ارزش‌های اخلاقی طراحی و استفاده شوند که از دید این رویکرد پنهان باقی مانده است. نظر به ضرورت هم‌سوسازی، این قبیل رویکردها پاسخ رضایت‌بخشی در اختیار ما قرار نمی‌دهند. توجه به تفاوت میان ارزش‌های فردی و ارزش‌های اجتماعی می‌تواند ما را برای انتخاب ارزش‌ها یاری دهد.

ما در این مقاله از این دیدگاه تبعیت می‌کنیم که هدف از هم‌سوسازی، همسویی فناوری با ارزش‌های اجتماعی است. اگرچه ارزش‌های فردی را مهم می‌دانیم اما محل

1 Desires
2 Superintelligence

توجه به آن‌ها را در حوزه تجربه کاربری[1] و کاربرپسند بودن[2] می‌دانیم. ارزش‌های افراد، لزوما ارزش‌های اجتماعی نیستند. به‌عنوان مثال، برخی از کاربران ممکن است بخواهند از داده‌های آن‌ها برای پردازش در سیستم هوش مصنوعی استفاده نشود؛ اما برخی دیگر ممکن است آن را مطلوب بدانند. نمونه چنین رویکردی را می‌توان در موتورهای جستجو مشاهده کرد که نگهداری تاریخچه داده‌های افراد به عنوان بخشی از تجربه کاربری، قابل انتخاب است. مطابق با این رویکرد، کاربرپسند بودن یا تجربه کاربری ارزش اجتماعی در نظر گرفته می‌شود. اما از سویی دیگر، به تفاوت ارزش‌های افراد در جامعه نیز توجه می‌شود. با توجه به اینکه سیستم هوش مصنوعی در سطح اجتماعی استفاده می‌شود و تحت تاثیر عوامل مختلفی قرار می‌گیرد، در این مقاله به ارزش‌های اجتماعی تمرکز می‌کنیم.

٤-۲. هدف از هم‌سوسازی ارزش در هوش مصنوعی

هدف از هم‌سوسازی هوش مصنوعی، طراحی و استفاده از این فناوری با حصول اطمینان از هم‌سویی آن با ارزش‌های اجتماعی است. با این‌حال، چنین هدفی از دو جهت با دشواری روبرو است. نخستین دشواری محدودیت در طراحی است (Gabriel, 2020: 413-417). چه از جهت وجود منابع و چه از جهت دانش علمی با محدودیت روبرو هستیم. توسعه فناوری نیازمند منابع مختلفی از جمله انرژی، منابع طبیعی مورد استفاده مستقیم همچون فلزات برای ساخت قطعات و منابع طبیعی غیرمستقیم همچون آب است. از یک جهت آلودگی‌های زیست محیطی و از جهت دیگر محدودیت این منابع برای توسعه فناوری را نباید فراموش کنیم. همچنین، دانش علمی نیز محدودیت دارد. در بسیاری زمینه‌های علمی سوال‌های فراوانی داریم که پاسخ آن‌ها را به دست نیاورده‌ایم. توسعه فناوری نیازمند دانش علمی است و محدودیت‌های علمی، توانایی طراحی فناوری را نیز محدود می‌سازد. لذا وجود شیوه برای هم‌سوسازی را با عدم وجود محدودیت هم‌معنا دانست.

دومین دشواری، امری هنجاری است. زمانی که از ارزش‌ها سخن می‌گوییم مقصود

1 User experience (UX)
2 User friendly

کـدام ارزش‌هـای اجتمـاعی هسـتند؟ در علـوم مختلـف، مفهـوم ارزش محـل بحـث میـان پژوهشگران است (Friedman & Hendry, 2019: 23). دسـته‌ای از پژوهشـگران معتقدنـد که ارزش‌های اقتصادی هوش مصنوعی باید اولویت قرار گیرد چراکه مزایای مختلفی برای جامعـه بـه ارمغـان مـی‌آورد (Andreessen, 2023). امـا ارزش‌هـای اجتمـاعی فراتـر از ارزش‌هـایی اقتصـادی اسـت. بـه‌عنـوان مثـال، حفاظـت از محـیط زیسـت یـا حفـظ حـریم خصوصی از نمونه ارزش‌هایی هستند که با تکیه بر اقتصاد نمی‌توان بـه آن‌هـا دسـت یافـت (Gabriel, 2020; 421). اگر به مشکلاتی که هوش مصنوعی به وجود آورده، مانند امکان فریب انسان توسط مدل‌های زبانی بزرگ[1] (Hagendorff, 2024)، ترویج محتـوای تـوهین آمیز توسط این مدل‌ها (Wolf et al., 2017) و یـا رسـوایی کمبـریج آنالتیکـا (Hinds et al., 2020) توجه کنیم، تکیه بر مزایای فناوری این چالش‌ها را برطرف نمی‌کند.

می‌توان استدلال کرد فناوری باید با ارزش‌های اخلاقی یا قـوانین و مقـررات هـم‌سـو شود. اما در سطح اجتماعی، برخی از ارزش‌های فرهنگی یا عرفی وجود دارند کـه تمرکـز بر اخلاق و قوانین آن‌ها را پوشش نمی‌دهد. اهمیت توجـه بـه ایـن ارزش‌هـا، در پـذیرش و مقبولیت برای جامعه است. اگر طراحی فناوری در جامعه مقبولیـت نداشـته باشـد، بـا اقبـال کاربران مواجه نمی‌شود. همچنین این امر می‌تواند موجب شود تا کاربران روش‌های گریـز از این ارزش‌ها را دنبال کنند. البته می‌توان استدلال کرد که برخی از ارزش‌های اجتماعی و یا عرفی ممکن است غیراخلاقی یا غیرقانونی باشند.

برای برون رفت از این چالش‌ها، فـریمن و هنـدری، ارزش‌هـای اجتمـاعی را در ایـن حوزه این‌گونه تعریف می‌کنند: «آن چیزی‌هایی که برای انسان‌هـا مهـم اسـت، بـا تکیـه بـر اخلاق و ارزش‌های اخلاقی» (Friedman & Hendry, 2019: 4). بـر ایـن اسـاس، چـالش ارزش‌های اجتماعی و تضاد آن‌ها با ارزش‌های اخلاقی برطرف می‌شود و می‌توان مقبولیت سیستم در جامعه را انتظار داشت. اما یک ایراد تعریف فریمن و هنـدری ایـن اسـت کـه بـه قوانین و مقررات توجه ندارد. لذا برای رهایی از این چالش، مـا هـدف از هـم‌سوسـازی را،

1 Large Language Models (LLMs)

«هم‌سویی هوش مصنوعی با چیزی‌هایی که ۱- برای انسان‌ها در سطح اجتماعی دارای اهمیت بوده ۲- مطابق با اخلاق و ارزش‌های اخلاقی ۳- و قوانین و مقررات است» در نظر می‌گیریم.

۳. رویکرد سیستم‌های اجتماعی فنی

حال که امکان، ضرورت و هدف از هم‌سوسازی را ذکر کردیم، زمان آن رسیده است که به رویکرد ارزیابی آن بپردازیم. بر اساس رویکرد سیستم‌های اجتماعی‌فنی، فناوری یک ابزار در خلا زمانی-مکانی و تنها طراحی و ظاهر مادی نیست؛ بلکه زیرساخت[1] و بخشی از یک سیستم اجتماعی‌فنی است. فناوری در یک فرایند پیچیده میان سیستم‌های اجتماعی، سیاسی، اقتصادی، فرهنگی و غیره طراحی شده و مورد استفاده قرار می‌گیرد (European Commission, 2023). مطابق این معرفی اولیه، ارزیابی هم‌سوسازی فناوری باید توجه به وجود عوامل مختلف در جامعه صورت بگیرد. البته ایده در نظر گرفتن فناوری در ساختار اجتماعی اگرچه قابل تامل است، می‌تواند به شیوه‌های گوناگونی تعبیر شود. روایت‌های مختلفی از سیستم‌های اجتماعی‌فنی وجود دارد (Watson et al., 2024) و همانطور که در ابتدا ذکر شد، روایتی پراگماتیستی را پی خواهیم گرفت.

۱-۳. روایت پراگماتیستی سیستم‌های اجتماعی‌فنی

روایت پراگماتیستی سیستم‌های اجتماعی‌فنی، پیشتر مورد توجه اندیشمندان قرار گرفته و پژوهش‌های مختلفی از این روایت دفاع کرده و در صدد تبیین آن برآمده‌اند[2]. زمانی که از پراگماتیسم سخن می‌گوییم، تعاریف و تعابیر مختلفی به ذهن متبادر می‌شود. لذا جا دارد منظور از روایت پراگماتیستی را تشریح کنیم. در حقیقت روایت پراگماتیستی، بر چند ایده

---

1 Infrastructure

۲ روایت پراگماتیستی مقاله حاضر، برگرفته از سه پژوهش (Watson et al., 2024; Kudina & Van de Poel, 2024; Rydenfelt, 2025) است. اما از آنجا که برای رسیدن به هدف از مقاله، معرفی آن لازم خواهد بود، لذا به صورت اختصار این روایت توضیح داده می‌شود.

از فیلسوفان پراگماتیست، شامل تأکید بر عاملیت انسانی، خطاپذیری[1]، پیشرفت[2]، بهبودگرایی[3] و سودمندی در اعتبار نظریه‌ها استوار است.

۱-۱-۳. عاملیت انسانی

در روایت پراگماتیستی سیستم‌های اجتماعی‌فنی، به تبعیت از دیویی (۱۹۱۷)، محدودیت‌های عوامل انسانی در انتخاب‌های پیشِ رو مورد توجه قرار می‌گیرد، اما تأثیرگذاری عاملیت انسانی از نظر دور نمی‌ماند. فناوری‌های موجود و محدودیت‌های آن‌ها تنها عامل مشخص کننده مسیر آینده فناوری نیستند، بلکه عوامل انسانی نیز نقشی مؤثر بازی می‌کنند (Watson et al., 2024: 3168). زمانی که می‌خواهیم یک فناوری را توسعه دهیم، با محدودیت‌های فنی روبرو می‌شویم و ممکن است هم‌سوسازی از نظر عملی، امکان‌پذیر نباشد. با وجود محدودیت فنی، این عوامل انسانی هستند که می‌توانند تصمیم بگیرند از فناوری‌های دیگر استفاده کنند؛ برخی از قابلیت‌های فناوری را محدود کنند؛ و یا تا رسیدن به فناوری‌ای که هم‌سو با ارزش‌ها باشد، از انتشار آن جلوگیری کنند. پس اگرچه محدودیت‌هایی وجود دارند، اما نقشِ عاملیت و مسئولیت انسانی قابل چشم‌پوشی نیست.

با این‌حال وقتی صحبت از سیستم‌های اجتماعی‌فنی می‌کنیم و به نقش سیستم‌های مختلف تاکید داریم، به جز فناوری، سایر عناصر اجتماعی مانند حاکمیت، نهادها و فرهنگ را تاثیرگذار می‌دانیم (Kudina & Van de Poel, 2024: 21). تعاملات پیچیده‌ای میان عناصر سیستم اجتماعی‌فنی وجود دارند و نمی‌توان آینده فناوری را از پیش تعیین شده دانست. البته نقش و تاثیرگذاری تمامی عناصر در سیستم یکسان نیست. بر اساس شرایط عینی، ممکن است برخی عناصر نقشی پررنگ‌تر یا کم‌رنگ‌تر داشته باشند. به‌عنوان مثال ممکن است سیستم‌های اقتصادی در یک جامعه و یا در یک کاربرد خاص از هوش

1 Fallibilism
2 Progress
3 Meliorism

مصنوعی، از نهادهای فرهنگی تاثیرگذاری بیشتری داشته و نقشی موثرتر در تعیین آینده آن ایفا کنند.

۲-۱-۳. خطاپذیری و پیشرفت

اگرچه خطر توسعه نامطلوب هوش مصنوعی جدی است، اما این موضوع را نباید به معنای نگاه محدودکننده در نظر گرفت. در روایت پراگماتیستی، محدود کردن توسعه فناوری تنها راه در برابر چالش‌های هم‌سوسازی نیست. با تکیه بر ایده پیشرفت (Dewey, 2004) و خطاپذیری (Peirce, 1955)، روایت پراگماتیستی سیستم‌های اجتماعی‌فنی امکانات متعددی در اختیار ما قرار می‌دهد.

زمانی که به ایده پیشرفت و خطاپذیری تکیه می‌کنیم، هدف کمال هم‌سوسازی هوش مصنوعی با ارزش‌ها نیست. نمی‌توان سیستمی طراحی کرد که کاملا با ارزش‌ها همسو باشد. می‌دانیم که همواره با عدم‌قطعیت مواجه هستیم. عدم‌قطعیت ممکن است ناشی از فقدان توجه به تاثیراتِ اجتماعیِ فناوری، عدم آینده‌نگری در زمان طراحی سیستم، محدودیت‌های دانش علمی و یا حتی ناشی از پیامدهای ثانویه ورود فناوری به جامعه باشد (Van de Poel, 2023: 120). در حقیقت ورود فناوری جدید به جامعه می‌تواند با آثاری همراه باشد که به صورت پیشینی[1] ناشناخته است و به شکل پسینی[2] و تدریجی و با یک «آزمایش اجتماعی[3]» نمایان می‌شود (Van de Poel, 2016: 669). مطابق با این رویکرد، هم‌سوسازی هوش مصنوعی با ارزش‌های اجتماعی و ارزیابی آن باید با توجه بر ایده پیشرفت و خطاپذیری در یک روند صورت گیرد.

۳-۱-۳. بهبودگرایی

بر اساس ایده بهبودگرایی (James 1907; Dewey 2004)، پراگماتیسم به آینده فناوری نه

1 a priori
2 a posteriori
3 Social experiment

بدبین و نه خوش‌بین است. تأکید می‌کند که آینده از پیش تعیین‌شده یا داده‌شده[1] نیست، بلکه تلاش‌های عوامل انسانی می‌تواند آن‌را تغییر دهد. البته تاکید به بررسی پسینی، به معنای عدم اهمیت بررسی پیشینی نیست. در حقیقت، پراگماتیسم به‌عنوان شیوه‌ای برای تحقیق[2]، با تکیه بر دو ایده خطاپذیری و بهبودگرایی، ما را به بررسی انتقادی پیش از آزمایش‌های اجتماعی دعوت می‌کند (Rydenfelt, 2025) اما در همین سطح متوقف نمی‌ماند و بررسی پسینی را نیز حائز اهمیت می‌داند.

٤-١-٣. سودمندی در اعتبار نظریه‌ها

زمانی که از اعتبار نظریه‌ها بر اساس سودمندی در انطباق با واقعیت (James, 1907) سخن می‌گوییم، اهمیت هنجارهای اجتماعی مشخص می‌شود. به‌عبارتی دیگر، اتخاذ یک نظریه اخلاقی برای هم‌سوسازی، باید با توجه به هنجارها و مقبولیت اجتماعی باشد. بدون این مقبولیت نمی‌توان انتظار رسیدن به اهداف هم‌سوسازی را داشت. لذا، ارزیابی هم‌سویی هوش‌مصنوعی با ارزش‌های اجتماعی باید در زمینه کاربردی[3] و در جامعه مشخص و مطابق با هنجارهای آن صورت گیرد (Watson et al., 2024: 3169). در حقیقت، روایت پراگماتیستی سیستم‌های اجتماعی، در بررسی هم‌سوسازی سیستم‌های هوش مصنوعی نیازمند مطالعه میان‌رشته‌ای با تمرکز بر زمینه کاربردی و جامعه‌یِ ذی‌نفعانِ[4] فناوری است (عاشوری و همکاران، ۱۴۰۳: ۲۹۵).

اکنون که روایت پراگماتیستی سیستم‌های اجتماعی‌فنی به اجمال معرفی شد، زمان

---

1 Given

2 Inquiry

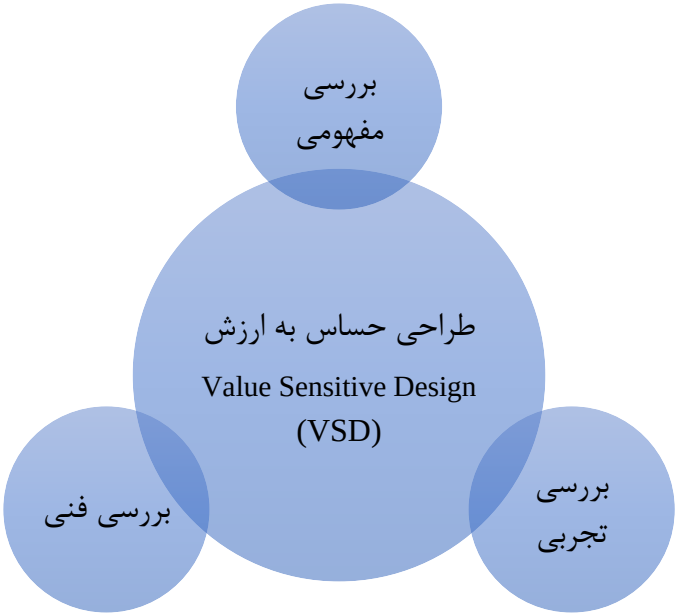
۳ سیستم‌های هوش مصنوعی برای استفاده‌های گوناگونی طراحی می‌شوند. سیستم‌ها ممکن است برای مقاصد آموزشی، پزشکی و غیره طراحی شوند. در هر زمینه کاربردی، ارزش‌های مختلفی دارای اهمیت هستند. همچنین، اولویت ارزش‌ها نیز بر اساس این زمینه‌ها متفاوت است. به‌عنوان مثال، اولویت ارزش‌ها در سیستمی که برای کاربرد پزشکی طراحی می‌شود، با سیستمی که برای تحصیل آنلاین طراحی می‌شود متفاوت است (عاشوری و همکاران، ۱۴۰۳: ۲۸۶).

4 Stakeholder

آن رسیده با شیوه «طراحی حساس به ارزش» آشنا شویم.

٤. طراحی حساس به ارزش

«طراحی حساس به ارزش» که توسط باتیا فریدمن[1] توسعه یافت، شیوه‌ای برای ادغام ارزش‌ها در فرآیند طراحی فناوری است. ایده اصلی «طراحی حساس به ارزش» این است که ارزش‌ها در فرآیند مهندسی لحاظ و بازتاب آن‌ها در طراحی نمایان شود. «طراحی حساس به ارزش» با استفاده از یک ساختار بررسی چند وجهی، شیوه‌ای برای هم‌سوسازی فناوری پدید می‌آورد. برای اینکه بیشتر با این شیوه آشنا شویم، لازم است تا این ساختار را مورد نظر قرار دهیم. این ساختار سه‌وجهی، شامل «بررسی مفهومی[2]»، «بررسی تجربی[3]» و «بررسی فنی[4]» است (Friedman & Hendry, 2019: 3-4).



1 Batya Friedman
2 Conceptual investigation
3 Empirical investigation
4 Technical investigation

۴-۱. ساختار سه وجهی طراحی حساس به ارزش

بررسی مفهومی شامل مشخص کردن ذی‌نفعان[1]، شناسایی ارزش‌ها و حل تنش میان ارزش‌ها است. بررسی در این مرحله عموما توسط متخصصین اخلاق صورت می‌گیرد[2]. در بررسی تجربی، از ابزارهای مختلف برای اعتبارسنجی ارزش‌ها و پرداختن به تنش میان ارزش‌ها در زمینه کاربردی مورد نظر استفاده می‌شود. در این مرحله، عموما متخصصین علوم اجتماعی، ارزش‌ها را در جامعه ذی‌نفعان بررسی می‌کنند. بررسی فنی که توسط مهندسین و متخصصین فنی صورت می‌گیرد، شامل طراحی سیستم برای تحقق عملی ارزش‌های شناسایی‌شده است. ترجمه ارزش‌ها به الزامات فنی و توجه به محدودیت‌های طراحی فناوری در این مرحله انجام می‌شود (Umbrello & Van de Poel, 2021: 284).

این بررسی سه‌وجهی ممکن است به‌صورت متوالی، موازی یا تکرارشونده اجرا شوند؛ با این‌حال بیشتر پژوهشگران، بررسی ادغام‌شده و تکرارشونده کل این ساختار سه وجهی در طول دوره حیات سیستم هوش مصنوعی را توصیه می‌کنند (e.g., de Reuver et al., 2020: 259; Friedman & Hendry, 2019: 4)؛ چرا که امکان حفظ کنترل انسانی در مواجهه و برطرف کردن پیامدهای غیرمنتظره در کل چرخه حیات سیستم وجود دارد و هم‌سوسازی تنها منحصر به زمان طراحی نخواهد بود.

۴-۲. ساختار سه وجهی در عمل

برای روشن‌تر شدن این ساختار سه وجهی ذکر یک مثال مفید است. فرض کنید قصد داریم یک سیستم هوش مصنوعی برای خدمات درمانی بیماران هموفیلی در کشور طراحی کنیم.

---

1 Stakeholders

۲ قابل توجه است که همواره هم‌گرایی میان ارزش‌ها در طراحی سیستم هوش مصنوعی وجود ندارد و گاهی ممکن است هم‌سویی با یک ارزش، منجر به محدودیت در هم‌سویی با ارزش دیگر شود (ر.ک. عاشوری کیسمی، ۱۴۰۳).

۱-۲-٤. بررسی مفهومی

در بررسی مفهومی، می‌خواهیم بدانیم چه ارزش‌هایی در طراحی سیستم دارای اهمیت هستند. برای این منظور باید ذی‌نفعان سیستم و ارزش‌های آن‌ها را شناسایی کنیم. همچنین باید بدانیم مفهوم هر ارزش در حوزه کاربردی ما چیست. به‌عنوان مثال، اگر حفاظت از حریم خصوصی یک ارزش ذی‌نفعان باشد، در زمینه خدمات درمانی بیماران هموفیلی این ارزش به چه معناست. پس از شناسایی ارزش‌ها باید بدانیم آیا میان آن‌ها تنشی وجود دارد یا خیر. به‌عنوان مثال ارزش‌هایی مانند حفظ حریم خصوصی و شفافیت[1] را می‌توان به شیوه‌های مختلفی تعریف کرد. گاهی امکان دارد تضادهایی میان تعابیر ما از ارزش‌ها شکل گیرد. مثلا، شفافیت سیستم نباید منجر به عدم حفظ حریم خصوصی کاربران سیستم شود. به همین خاطر این قبیل تنش‌ها در بررسی مفهومی و توسط متخصصین اخلاق ارزیابی می‌شود.

۲-۲-٤. بررسی تجربی

بررسی ارزش‌ها در سطح اجتماعی بسیار مهم است و با تکیه بر بررسی اخلاقی نمی‌توان به این امر دست‌یافت. لذا در بررسی تجربی، ارزش‌ها در جامعه ذی‌نفعان ارزیابی می‌شود. در این مرحله، با استفاده از روش‌های مختلف مانند ابزارهای آماری، شبیه‌سازی، مصاحبه، پرسش‌نامه و روش‌های مشابه ارزش‌ها در جامعه ذی‌نفعان بررسی خواهند شد.

۳-۲-٤. بررسی فنی

اگرچه دو بررسی پیشین روشن‌گر هستند، گاهی محدودیت‌های فنی و تکنیکی وجود دارد. ممکن است در طراحی و روند مهندسی نیازمند بررسی روش‌های مختلف طراحی برای دستیابی به نتیجه مورد نظر باشیم. امکان دارد متخصصین و مهندسان، با محدودیت‌هایی برای هم‌سوسازی ارزش‌ها روبرو شوند و نیازمند هم‌فکری با متخصصین

1 Transparency

مراحل پیشین برای یافتن راه‌کار جایگزین باشند.

همانطور که ذکر شد، بهتر است این بررسی‌ها به صورت ادغام‌شده و تکرارشونده در کل دوره حیات سیستم صورت بگیرند تا احتمال رسیدن به یک نتیجه قابل اتکا افزایش یابد. به‌عنوان مثال ممکن است در بررسی مفهومی، برخی توصیه‌ها برای طراحی داشته باشیم که از نظر فنی قابل اجرا نباشد. امکان دارد در بررسی تجربی به این نتیجه برسیم که جامعه، مفاهیم ارزشی که در بررسی مفهومی به آن‌ها دست‌یافتیم را به گونه‌ای دیگر طلب می‌کند[1]. در مجموع تکرار و ادغام بررسی‌ها امکان می‌دهد تا گفتگوی میان رشته‌ای میان متخصصین صورت گرفته و در عمل بتوان طراحی فناوری را بر اساس ارزش‌ها انجام داد.

---

۱ در رسیدن به این جمع‌بندی، «طراحی حساس به ارزش» یک سیر تاریخی را گذراند. برای انتخاب ارزش‌ها، در طول زمان سه رویکرد اصلی توسط پژوهشگران پیشنهاد شد. در رویکرد بالا به پایین (Top-down) متخصصین اخلاق با تکیه بر یک یا چند رویکرد اخلاقی، ارزش‌ها را انتخاب می‌کردند. با توجه به این‌که فهرست ارزش‌های از پیش تعیین‌شده ممکن است همراه با نادیده گرفتن ارزش‌های مهمی در سطح اجتماعی باشد، این رویکرد مورد انتقاد قرار گرفت (Cenci & Cawthorne, 2020). در مقابل این انتقادات، رویکرد پایین به بالا (bottom-up) پیشنهاد شد که ارزش‌ها باید از ذی‌نفعان دریافت شوند و تمامی جزئیات استفاده از فناوری مورد ارزیابی قرار گیرد (Umbrello & Van de Poel, 2021: 284). نظریه‌پردازان معتقد بودند فناوری نه تنها بر ذی‌نفعان مستقیم بلکه بر ذی‌نفعان غیرمستقیم نیز تأثیر می‌گذارد. حتی اگر به‌دلیل محدودیت‌های فنی، تعامل با تمام ذی‌نفعان و هم‌سویی با ارزش‌ها ممکن نباشد، هرگونه حذف در توجه به ارزش‌ها باید شناسایی و موجه شود (Cawthorne, 2023). در این راستا، مطالعات فراوانی صورت گرفته و پژوهشگران ذی‌نفعان مختلف را بررسی کرده‌اند. برخی ذی‌نفعان انسانی و غیرانسانی (Umbrello, 2021)، برخی دیگر ذی‌نفعان مستقیم و غیرمستقیم کنونی و نسل‌های آینده را به‌عنوان ذی‌نفعان آتی (Jacobs & Huldtgren, 2021) را بررسی کردند. با این حال، رویکرد پایین به بالا به دلیل ناتوانی در توجیه اولویت‌بندی ارزشی‌ها یا مصلحت‌اندیشی به‌جای پایبندی به چارچوب‌های اخلاقی مورد انتقاد قرار گرفت (Jacobs, 2021). این‌که ذی‌نفعان چه ارزش‌هایی را پیشنهاد دهند و یا براساس آمار ارزش‌ها را اولویت‌بندی کنیم نمی‌تواند آن‌ها را از نظر اخلاقی موجه کند. با توجه به انتقادات، رویکردهای میانه یا ترکیبی (Hybrid) پیشنهاد شد. براساس این رویکرد، نظریه‌های اخلاقی عمل‌گرایانه و ترکیبی که شامل مجموعه‌ای از اصول اخلاقی با تکیه بر ارزش‌های اجتماعی و ذی‌نفعان هستند، می‌توانند مفیدتر از رویکردهای پیشین باشند. زیرا هم با ارزش‌های جامعه و هم با چارچوب‌های اخلاقی هم‌سو هستند (Jacobs, 2021). علاوه بر این، تغییر در ارزش‌ها (Value change) و روایت‌های مختلف از ارزش‌ها توسط پژوهشگران این رویکرد مورد توجه قرار گرفته است (Veluwenkamp & Van den Hoven, 2023).

۵. بحث و بررسی

بر اساس هدف مقاله، نوبت به ارزیابی شیوه «طراحی حساس به ارزش» رسیده است. برای این منظور، ارزیابی در دو بخش مزایا و چالش‌ها صورت می‌گیرد.

۵-۱. مزایا

۵-۱-۱. ورود به حوزه عمل و عاملیت ذی‌نفعان

«طراحی حساس به ارزش» شیوه‌ای عملی برای هم‌سوسازی سیستم‌های هوش مصنوعی در اختیار قرار می‌دهد که در آن به عاملیت و نقش ذی‌نفعان توجه می‌شود. در نهایت باید فناوری در عمل هم‌سو با ارزش‌ها بوده و به عاملیت ذی‌نفعان توجه شود. بر خلاف شیوه‌های دیگر هم‌سوسازی مانند «عامل اخلاقی هوش مصنوعی»، «هوش ترکیبی» که عاملیت ذی‌نفعان کمتر مورد توجه قرار می‌گیرد (c.f. Wallach & Allen, 2009; Akata et al., 2020)، نقش عاملیت ذی‌نفعان در «طراحی حساس به ارزش» در بررسی مفهومی و تجربی هم به صورت پیشینی در زمان طراحی و هم به صورت پسینی در کل چرخه حیات سیستم پررنگ است.

۵-۱-۲. تکرار و ادغام بررسی‌ها در دوره حیات سیستم هوش مصنوعی

سیستم‌های هوش مصنوعی ممکن است در طول زمان ویژگی‌های «نامطلوب» پیدا کنند یا از ارزش‌های اولیه که با آن هم‌سو شده بودند فاصله بگیرند (Van de Poel, 2023: 123). همچنین، جامعه و یا عناصر مختلف موجود در سیستم اجتماعی‌فنی ممکن است تغییراتی داشته باشند و در نتیجه تکرار بررسی‌ها برای اطمینان از هم‌سویی هوش مصنوعی حیاتی خواهد بود. بررسی ادغام شده ساختار سه‌وجهی «طراحی حساس به ارزش» کمک می‌کند تا بررسی میان‌رشته‌ای صورت بگیرد؛ همچنین بررسی تکرار شونده نگرانی‌ها از تغییرات نامطلوب سیستم و نیاز به بررسی در کل چرخه حیات را مرتفع می‌سازد.

۳-۱-۵. بهبودگرایی و پیشرفت

تا کنون «طراحی حساس به ارزش» با چندین انتقاد مواجه شده است. به‌عنوان مثال انتقاداتی همچون وابستگی به ارزش‌های اکتشافی[۱] (Reijers & Gordijn, 2019)، فقدان روش‌شناسی برای تمایز میان ارزش‌ها از ترجیحات[۲] فردی (Manders-Huits, 2011)، عدم پایبندی به یک نظریه اخلاقی مشخص (Jacobs & Huldtgren, 2021)، خطر مغلطه طبیعت‌گرایانه[۳] (Manders-Huits, 2011) و غیره. با این حال، این انتقادات از سوی نظریه‌پردازان «طراحی حساس به ارزش» نادیده گرفته نشده‌اند. برای رفع این قبیل مشکلات شیوه‌های اصلاح‌شده‌ی «طراحی حساس به ارزش» ارائه شده است. برای مثال، «طراحی حساس به ارزش‌های آینده‌نگر[۴]» (Beach & Fox, 2021)، «طراحی هوش مصنوعی برای توسعه پایدار[۵]» (Umbrello & Van de Poel, 2021) «ارزش‌ها در عمل[۶]» (Flanagan et al., 2005)، «ارزش‌ها در طراحی[۷]» (Bowker & Star, 1999)، «طراحی حساس به قابلیت‌ها[۸]» (Jacobs, 2020)، و «روش‌های مبتنی بر قابلیت‌های رویه‌ای[۹]» (Cenci & Cawthorne, 2020) از جمله شیوه‌هایی هستند که بر اساس «طراحی حساس به ارزش» توسعه یافته‌اند. در مجموع نظریه‌پردازان «طراحی حساس به ارزش»، در برابر انتقادات پذیرا بوده و راه‌کارهای عملی جدید ارائه داده‌اند. این امر موجب شده تا شیوه کنونی بهبود زیادی پیدا کرده و از بسیاری چالش‌ها عبور کند.

---

1 Heuristic values
2 Preferences
3 Naturalistic fallacy
4 Value Sensitive Speculative Design (VSSD)
5 AI4SG-VSD
6 Values at Play (VAP)
7 Values in design
8 Capability Sensitive Design (CSD)
9 The capability-based procedural VSD

۲-۵. چالش‌ها

۱-۲-۵. روند طولانی طراحی سیستم

یکی از ایراداتی که ممکن است به «طراحی حساس به ارزش» وارد شود، پیچیدگی و روند طولانی آن است. می‌توان اعتراض کرد اگر به چنین شیوه‌ای قصد راه‌اندازی سیستمی را داشته باشیم، طراحی و به‌روزرسانی‌های آن دشوار و پرهزینه است. اما باید در نظر داشت که شرایط این دشواری چیست. در سیستم‌هایی با کاربردی محدود و یا عملکرد ساده می‌توان انتظار داشت که هم‌سوسازی دشوار نخواهد بود. دشواری زمانی نمایان می‌شود که با سیستم‌های پیچیده و کاربران متنوع سر و کار داریم. جای تأمل است که نه تنها هم‌سوسازی، بلکه طراحی و نگهداری از چنین سیستم‌هایی پرهزینه، زمان‌بر و نیاز به بررسی مداوم دارد. باید متذکر شویم عوامل انسانی انتخاب‌های مختلفی از جمله طراحی فناوری جایگزین و یا به تاخیر انداختن انتشار فناوری تا زمان رسیدن به نتایج مطلوب دارند، لذا در پاسخ به این اعتراض می‌توان گفت دشواری طراحی مسئولیت اخلاقی عوامل انسانی را بی‌اثر نمی‌کند.

۲-۲-۵. کنترل معنادار

عوامل مختلف و زیرسیستم‌ها تاثیرات گوناگونی بر کل سیستم و اجزای آن می‌گذارند. ممکن است استدلال کنیم که یک محدودیت «طراحی حساس به ارزش» عدم کنترل بر سایر زیرسیستم‌ها در سیستم اجتماعی‌فنی است. در برابر این ایراد باید گفت که با طراحی یک سیستم هوش مصنوعی، قصد نداریم کل جامعه را کنترل کنیم. قانون‌گذاری، آموزش اجتماعی، سیاست‌گذاری آگاهانه و روش‌های مشابه می‌توانند مشکلات اجتماعی هوش مصنوعی در جامعه را کاهش دهند. اما این قبیل اقدامات ذیل هم‌سوسازی هوش مصنوعی با ارزش‌ها قرار نمی‌گیرند. به‌عنوان مثال، نگاهی بسیار تقلیل‌گرایانه به آموزش خواهد بود اگر آن را ذیل توسعه هوش مصنوعی هم‌سو با ارزش‌های قرار دهیم. چنین نگاهی، سایر تاثیرات، اهداف و نقش آموزش در جامعه را نادیده می‌گیرد. سیستم هوش مصنوعی یک

زیرساخت در سیستم اجتماعی‌فنی است که هم‌سوسازی ان ضرورت دارد اما این موضوع به معنای کنترل کل سیستم و عدم توجه به نقش‌های گوناگون سایر زیرسیستم‌ها نیست.

سیستم‌های هوش مصنوعی سه ویژگی استقلال[1] در عمل، تعامل[2] با محیط و تطبیق‌پذیری[3] دارند که آن‌ها را از سایر فناوری‌ها متفاوت می‌سازد. تفاوت این سیستم‌ها با سایر برنامه‌های کامپیوتری نیز در وجود این ویژگی‌ها است (Van de Poel, 2023: 118). این سیستم‌ها می‌توانند برخی از تصمیمات را بدون دخالت انسان بگیرد و نوعی استقلال در عمل دارند. آن‌ها با داده‌های آموزشی و اطلاعات ورودی دارای تعامل هستند که از این موضوع به عنوان تعامل با محیط یاد می‌شود. همچنین این سیستم‌ها با داده‌های آموزشی تطبیق‌پذیری پیدا کرده و الگوریتم‌ها در طول زمان دچار تغییرات می‌شوند. با نظر به این ویژگی‌ها، می‌توان برای سیستم‌های هوش مصنوعی نوعی عاملیت غیرانسانی و غیرالتفاتی در نظر گرفت (عاشوری کیسمی و همکاران، ۱۴۰۴: ۲۸۷–۲۸۸). در اینجا یک چالش اساسی کنترل معنادار عامل انسانی بر سیستم هوش مصنوعی است (حسینی سورکی، ۱۴۰۳).

ممکن است از ما بپرسند چرا کنترل معنادار انسانی برای هم‌سوسازی سیستم هوش مصنوعی با ارزش‌ها ضرورت دارد؟ در پاسخ ما به دو دلیل اکتفا می‌کنیم. اول اینکه به علت سه ویژگی خاص هوش مصنوعی، این سیستم‌ها ممکن است دچار تغییرات شده و از ارزش‌هایی که ابتدا با آن هم‌سو شده بودند فاصله بگیرد. همچنین ممکن است در طراحی اولیه، دچار اشتباهاتی شده باشیم و برخی آثار فناوری از نظر دور مانده باشد. در چنین شرایطی کنترل معنادار انسانی امکان اصلاح و تغییر را فراهم می‌کند. دوم آنکه، مسئولیت اخلاقی را نمی‌توان به سیستم هوش مصنوعی منصوب کرد. برای اینکه عوامل انسانی بتوانند مسئولیت اخلاقی داشته باشند، باید روی سیستم کنترل معنادار داشته باشند. در این

1 Autonomy
2 Interactivity
3 Adaptivity

خصوص، چهار فراارزشِ[1] «قابلیت نظارت[2]»، «برگشت‌پذیری[3]»، «سازگاری[4]»، و «پاسخ‌گویی[5]» به‌عنوان شرایط لازم و کافی کنترل معنادار انسانی پیشنهاد شده است (Van de Poel, 2023: 123-124).

در خصوص چهار فراارزش مطرح شده اگرچه پاسخ‌گویی طراحان و سازگاری هوش مصنوعی با ارزش‌های انسانی قابل حصول است، دو مورد برگشت‌پذیری و قابلیت نظارت با دشواری مواجه هستند. برگشت‌پذیری، این‌طور تعریف می‌شود که هوش مصنوعی باید به گونه‌ای طراحی شود که در صورت بروز مشکل مانند یک نرم‌افزار قابلیت برگشت به نسخه قبلی و قابل اطمینان را داشته باشد (ibid: 123). اما این شیوه نمی‌تواند تاثیرات اجتماعی را برطرف سازد. اگرچه سیستم ممکن است قابل برگشت به نسخه پیشین باشد، اما جامعه‌ای که تحت تاثیر استفاده از آن قرار گرفته چنین شرایطی را ندارد.

در خصوص شیوه قابل نظارت، یک چالش اصلی این خواهد بود که نظارت بر سیستم بر چه اساس و با چه شیوه‌ای صورت می‌گیرد. یکی از چالش‌های اصلی در حوزه هوش مصنوعی کدری[6] و شفافیت این سیستم‌ها است. به صورت خلاصه، کدری به شرایطی اشاره دارد که روند سیستم در رسیدن به خروجی‌ها نامشخص است (Muller, 2021). کدری ممکن است کم عمق و یا به صورت عمیق باشد. کدری کم عمق به شرایطی اشاره دارد که روند رسیدن سیستم به خروجی به دشواری قابل دستیابی است و کدری عمیق به حالتی که روند رسیدن به خروجی نامشخص است (ر.ک. عاشوری کیسمی، ۱۴۰۳: ۵۷-۵۸). در خصوص کدری عمیق اینکه نظارت چگونه صورت می‌گیرد و چه دانش معرفتی از این نظارت به‌دست می‌آید خود یک چالش مهم است. به‌عنوان مثال آیا دیاگرام‌ها، یا روش‌های هوش مصنوعی قابل توضیح[7] می‌تواند شرایط لازم برای نظارت

1 Meta Value
2 Monitorability
3 Reversibility
4 Adaptability
5 Accountability
6 Opacity
7 Explainable AI (xAI)

قابل اعتماد داشته باشیم، خود محل بحث در حوزه معرفت‌شناسی خواهد بود.

۳-۲-۵. خود تنظیمی[1] و چالش قوانین و حکم‌رانی

با توجه به اینکه تاخیر در توسعه فناوری تا وضع قوانین و مقررات می‌تواند به شرکت‌های توسعه دهنده فناوری ضرر و زیان مالی وارد کند، ایده خود تنظیمی مطرح شده است. بر اساس این ایده، شرکت‌ها و توسعه‌دهندگان مسئولیت حفاظت از حقوق کاربران را بر عهده می‌گیرند. با این‌حال این شیوه یک شمشیر دولبه است. انتقادات مختلفی به شیوه خودتنظیمی وارد شده است. به‌عنوان مثال، قواعدی که شرکت‌ها برای کاربران وضع می‌کنند، مانند قوانین دولت‌ها جایگاه حقوقی ندارد؛ و یا اینکه در این شیوه شرکت‌ها به جای دولت‌هایِ برآمده از انتخاب مردم توانایی وضع قواعد برای شهروندان را دارند. در عمل، شرکت‌های بزرگ فناوری قدرتی فراتر از دولت‌های محلی پیدا کرده‌اند و از طرف دیگر، توسعه قوانین به بهانه وجود شیوه خود تنظیمی زمان‌بر و به تاخیر افتاده است (Alpsancar, 2023: 2-3). اگر «طراحی حساس به ارزش» نیز به‌عنوان روشی خودتنظیمی در نظر گرفته شود، می‌تواند نگرانی در خصوص توسعه قوانین و مقررات را افزایش دهد.

چالش دیگر این است که بسیاری از سیستم‌های هوش مصنوعی به‌صورت بین‌المللی استفاده می‌شوند. قوانین و مقررات کشورها تفاوت‌های فراوانی دارند و در این حوزه شاهد فقدان قوانین فراملی و بین‌المللی با ضمانت اجرایی هستیم. این موضوع از چند جهت قابل تامل است. اول آنکه ذی‌نفعان سیستم، زمینه‌های فرهنگی و عرفی آن‌ها بسیار متنوع است. تعدد عوامل در سیستم‌ها در سطح بین‌الملل و ذی‌نفعان که در بررسی سه وجهی «طراحی حساس به ارزش» باید لحاظ شود می‌تواند با خطای فراوان همراه باشد. تمرکز به شیوه‌های آماری و محدودیت در گزینش ذی‌نفعان از جمله چالش‌های اساسی در چنین شرایطی است. دوم اینکه قوانین محلی عموما بر شرکت‌های بین‌المللی در این حوزه تاثیر چشمگیری ندارند. حداقل تا نبود قوانین بین‌المللی با ضمانت اجرایی که بتواند از قوانین

1 Self-regulation

محلی حمایت کند این چالش کماکان وجود خواهد داشت. عموما شیوه‌های محدود کننده محلی به تنهایی نمی‌توانند کارایی داشته باشند[۱]. در مجموع، برای کنترل سایر عوامل اجتماعی حاضر در سیستم‌های اجتماعی‌فنی نیازمند قوانین و مقررات و توجه به موضوع حکم‌رانی خواهیم بود[۲].

نتیجه‌گیری

پس از بررسی شیوه «طراحی حساس به ارزش» از منظر روایت پراگماتیستی سیستم‌های اجتماعی‌فنی یافته‌های این پژوهش نشان می‌دهد این شیوه برای هم‌سوسازی هوش مصنوعی نقاط قوت فراوانی دارد. این شیوه چارچوبی ارائه می‌دهد که بررسی‌های اخلاقی را از حوزه نظری به حوزه عملی وارد کرده و اخلاق فناوری در توسعه آن مشارکت فعال خواهد داشت. البته، چالش‌هایی در مسیر این شیوه وجود دارد. با نظر به ایده‌هایی همچون بهبودگرایی، پیشرفت و خطاپذیری رسیدن به چارچوبی که کمال ویژگی‌ها برای هم‌سوسازی را داشته باشد دور از دسترس است. با این‌حال، توجه به ذی‌نفعان، بررسی‌های پیشینی میان رشته‌ای و تکرار آن‌ها با بررسی‌های پسینی بعد از انتشار و در کل دوره حیات سیستم هوش مصنوعی از جمله نقاط قوت «طراحی حساس به ارزش» هستند که مشکلات پیش رو را تا حد قابل قبولی برطرف می‌کنند. یافته دیگر این پژوهش نشان می‌دهد، برای این‌که شیوه‌های هم‌سوسازی فناوری بتوانند به واقع تاثیرگذار باشند، وجود قوانین و مقررات بین‌المللی با ضمانت اجرایی و توجه به حکم‌رانی ضروری است. بدون توجه به این عوامل و زیرسیستم‌ها و همچنین در هم‌تنیدگی سیستم‌های اجتماعی‌فنی بین‌المللی، شیوه‌های هم‌سوسازی فاقد کارایی خواهند بود. شایسته است سایر پژوهش‌گران روش‌های دیگر هم‌سوسازی فناوری با ارزش‌ها را بر اساس روایت پراگماتیستی از سیستم‌های

---

۱ به‌عنوان مثال، در شیوه فیلترینگ روش‌های مختلفی وجود دارد که آن‌را بی‌اثر می‌کند.

۲ قوانین اتحادیه اروپا (European Commission, 2021) را می‌توان نمونه خوبی از تصمیم‌گیری جمعی دانست. البته این نوع قوانین تنها در حوزه جغرافیایی محدودی اثرگذار بوده و در دیگر نقاط جهان فاقد اثر کافی است.

اجتماعی‌فنی مورد بررسی قرار دهند. همچنین مبحث ضمانت اجرایی مقررات بین‌الملل با نظر به حکمرانی محلی نیز شایسته توجه ویژه و فوری است.

تعارض منافع

تعارض منافع ندارد.

بیانیه نحوه استفاده از هوش مصنوعی

در این مقاله برای ویرایش زبان انگلیسی، دستور زبان و سبک از هوش مصنوعی استفاده شده است.

ORCID
Mohammad Ali Ashouri Kisomi **https://orcid.org/0000-0001-5663-1993**

منابع

السلتی، کیوان. (۱۴۰۳). هوش مصنوعی و کنترل معنادار بشری: سازوکارهای واکنش‌پذیر به دلایلِ نسبت-به-عامل-خنثی و امکانِ ردگیری عامل(های) مسئول. *پژوهش‌های فلسفی-کلامی*، ۲۶(۴)، ۲۷-۵۴. DOI: 10.22091/jptr.2025.11378.3139

حسینی سورکی، سید محمد. (۱۴۰۳). معنا و معیار عاملیت اخلاقی ماشین‌های هوشمند. *پژوهش‌های فلسفی-کلامی*، ۲۶(۴)، ۸۳-۱۰۸. DOI: 10.22091/jptr.2025.12466.3256

عاشوری کیسمی، محمد علی. (۱۴۰۳). همگرایی حریم خصوصی و شفافیت، محدودیت‌های طراحی هوش مصنوعی. *حکمت و فلسفه*، ۲۰(۷۸)، ۴۵-۷۳. DOI: 10.22054/wph.2024.75680.2183

عاشوری کیسمی، محمد علی. پرویزی، مریم و مومنی، سیدمحمدحسن. (۱۴۰۴). پراگماتیسم، اخلاق هوش مصنوعی و سیستم‌های اجتماعی‌فنی. *مجله پژوهش‌های فلسفی*. ۱۹(۵۲)، ۲۸۱-۳۰۰. DOI: 10.22034/jpiut.2025.66643.4060

عاشوری کیسمی، محمد علی. پرویزی، مریم. (۱۴۰۳). بررسی عاملیت اخلاقی هوش مصنوعی عام. *پژوهش‌های علم و دین*، ۱۵(۱)، ۱۲۵-۱۵۱. DOI: 10.30465/srs.2024.49282.2162

عباسی، محمود. تیموری، مهرداد. (۱۴۰۲). مروری بر چالش‌های اخلاقی و حقوقی کاربرد هوش مصنوعی در نظام سلامت. *مجله اخلاق پزشکی*، ۱۷(۴۸)، ۱-۱۱. DOI: 10.22037/mej.v17i48.44053

References

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., ... Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer, 53*(8), 18-28. https://doi.org/10.1109/MC.2020.2996587

Asilomar AI Principles. (2018). *Principles developed in conjunction with the 2017 Asilomar conference* [Benevolent AI 2017].

Beach, M. W., & Fox, T. (2021). Value sensitive speculative design: Exploring more-than-human relations in the age of climate catastrophe. *IxD&A, 51,* 111-131. https://doi.org/10.55612/s-5002-051-005

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Cawthorne, D. (2023). *The ethics of drone design: How value-sensitive design can create better technologies*. Routledge.

Cenci, A., & Cawthorne, D. (2020). Refining value sensitive design: A (capability-based) procedural ethics approach to technological design for well-being. *Science and Engineering Ethics, 26*(5), 2629-2662. https://doi.org/10.1007/s11948-020-00223-3

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., ... & de Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns, 4*(10), Article 100857. https://doi.org/10.1016/j.patter.2023.100857

Corrêa, N. K. (2024). Dynamic normativity: Necessary and sufficient conditions for value alignment. *arXiv preprint*. https://doi.org/10.48550/arXiv.2406.11039

de Reuver, M., van Wynsberghe, A., Janssen, M., & Van de Poel, I. (2020). Digital platforms and responsible innovation: Expanding value sensitive design to overcome ontological uncertainty. *Ethics and Information Technology, 22*(3), 257-267. https://doi.org/10.1007/s10676-020-09537-z

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)* (COM 2021 206). https://ec.europa.eu/transparency/documents-register/detail?ref=COM(2021)206&lang=en

European Commission. (2023). *EU-U.S. terminology and taxonomy for artificial intelligence*. https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2005). Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 751-760). ACM. https://doi.org/10.1145/

1054972.1055076

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction, 11*(2), 63-125. https://doi.org/10.1561/1100000015

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411-437. https://doi.org/10.1007/s11023-020-09539-2

Gabriel, I., & Ghazavi, V. (2022). The challenge of value alignment. In *The Oxford handbook of digital ethics* (pp. 336-355). Oxford University Press.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences, 121*(24), e2317967121. https://doi.org/10.1073/pnas.2317967121

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies, 143*, Article 102498. https://doi.org/10.1016/j.ijhcs.2020.102498

Howcroft, J., & Vale, J. (2024). What do stakeholders value? Integrating value sensitive design techniques into an impacts-focused design course. *Proceedings of the Canadian Engineering Education Association (CEEA)*.

Jacobs, N. (2021). *Values and capabilities: Ethics by design for vulnerable people* [Doctoral dissertation, Eindhoven University of Technology].

Jacobs, N., & Huldtgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and Information Technology, 23*(1), 23-26. https://doi.org/10.1007/s10676-018-9467-3

Kroes, P., & Verbeek, P. P. (Eds.). (2014). *The moral status of technical artifacts*. Springer.

Kudina, O., & Van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines, 34*(3), Article 21. https://doi.org/10.1007/

s11023-024-09680-2

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and Engineering Ethics, 17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

Miller, B. (2021). Is technology value-neutral? *Science, Technology, & Human Values, 46*(1), 53-80. https://doi.org/10.1177/0162243919900965

Muller, V. (2021). Ethics of artificial intelligence and robotics. In *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/ethics-ai/

Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In *Risk and responsibility in context*(pp. 191-213). Routledge. https://doi.org/10.4324/9781003276029-14

Pareto, J., & Coeckelbergh, M. (2024). Social assistive robotics: An ethical and political inquiry through the lens of freedom. *International Journal of Social Robotics, 16*(8), 1797-1808. https://doi.org/10.1007/s12369-024-01161-x

Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science and Engineering Ethics, 17*(3), 411-424. https://doi.org/10.1007/s11948-010-9244-0

Pitt, J. C. (2013). "Guns don't kill, people kill"; values in and/or around technologies. In P. Kroes & P. P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 89-101). Springer. https://doi.org/10.1007/978-94-007-7914-3_6

Reijers, W., & Gordijn, B. (2019). Moving from value sensitive design to virtuous practice design. *Journal of Information, Communication and Ethics in Society, 17*(2), 196-209. https://doi.org/10.1108/JICES-10-2018-0080

Russell, S. (2019). *Human compatible: AI and the problem of control*. Allen Lane.

Rydenfelt, H. (2025). Pragmatism as sociotechnical fallibilism and meliorism. *AI & SOCIETY*, 1-13. https://doi.org/10.1007/s00146-025-02499-0

Shen, Y., Tang, L., Le, H., Tan, S., Zhao, Y., Shen, K., ... & Fan, Y. (2025). Aligning and comparing values of ChatGPT and human as learning facilitators: A value-sensitive design approach. *British Journal of Educational Technology, 56*(4), 1391-1414. https://doi.org/10.1111/

bjet.13562

Srivatsa, N., Kaliarnta, S., & Kormelink, J. G. (2017). *Responsible innovation: From MOOC to book*. TU Delft OPEN Publishing.

Star, S. L., & Bowker, G. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Umbrello, S. (2021). The ecological turn in design: Adopting a posthumanist ethics to inform value sensitive design. *Philosophies, 6*(2), Article 29. https://doi.org/10.3390/philosophies6020029

Umbrello, S., & De Bellis, A. F. (2018). A value-sensitive design approach to intelligent agents. In *Artificial intelligence safety and security* (pp. 395-409). Chapman and Hall/CRC.

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics, 1*(3), 283-296. https://doi.org/10.1007/s43681-021-00038-3

UNICEF. (2021). *Policy guidance on AI for children*. United Nations Children's Fund.

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667-686. https://doi.org/10.1007/s11948-015-9724-3

Van de Poel, I. (2023). AI, control and unintended consequences: The need for meta-values. In *Rethinking technology and engineering: Dialogues across disciplines and geographies* (pp. 117-129). Springer. https://doi.org/10.1007/978-3-031-25233-4_9

Van de Poel, I., & Kroes, P. (2014). Can technology embody values? In P. Kroes & P. P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 103-124). Springer. https://doi.org/10.1007/978-94-007-7914-3_7

Van den Hoven, J., Vermaas, P., & van de Poel, I. (Eds.). (2015). *Handbook of ethics and values in technological design: Sources, theory, values and application domains*. Springer.

Veluwenkamp, H., & van den Hoven, J. (2023). Design for values and conceptual engineering. *Ethics and Information Technology, 25*(1), Article 2. https://doi.org/10.1007/s10676-022-09675-6

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Watson, D. S., Mökander, J., & Floridi, L. (2024). Competing narratives in AI ethics: A defense of sociotechnical pragmatism. *AI & SOCIETY*, 3163-3185. https://doi.org/10.1007/s00146-024-02128-2

Werthner, H. (2020). The Vienna manifesto on digital humanism. In *Digital transformation and ethics* (pp. 338-357). Ecowin.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*, 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: Comments on Microsoft's Tay "experiment," and wider implications. *ACM SIGCAS Computers and Society, 47*(3), 54-64. https://doi.org/10.1145/3144592.3144598

Young, M. T., & Coeckelbergh, M. (2024). Keeping things going: Maintenance and philosophy of technology. In M. T. Young & M. Coeckelbergh (Eds.), *Maintenance and philosophy of technology: Keeping things going* (pp. 1-36). Routledge.

Zuboff, S. (2023). The age of surveillance capitalism. In W. Longhofer & D. Winchester (Eds.), *Social theory re-wired: New connections to classical and contemporary perspectives* (3rd ed., pp. 203-213). Routledge.

Alpsancar, S. (2023, May). What is AI ethics? Ethics as means of self-regulation and the need for critical reflection. In *International Conference on Computer Ethics* (Vol. 1, No. 1). http://journals.library.iit.edu/index.php/CEPE2023/article/view/227/233

Andreessen, M. (2023, October 16). *The techno-optimist manifesto*. Andreessen Horowitz. https://a16z.com/the-techno-optimist-manifesto/

References [In Persian & In Arabic]

Abbasi, M., & Teymouri, M. (2024). A Review of the Ethical and Legal Challenges of Using Artificial Intelligence in the Health System. *akhlāq-I Pizishkī i.E., Medical Ethics*, *17*(48), 1–11. doi: 10.22037/mej.v17i48.44053 [in Persian]

Alasti, K. (2024). Artificial Intelligence and Meaningful Human Control: Agent-Neutral-Reason-Responsive Mechanisms and the Possibility of Tracing Responsible Agent(s). Journal of Philosophical Theological Research, 26(4), 27-54. doi: 10.22091/jptr.2025.11378.3139 [in Persian]

Ashouri Kisomi, M. A. & Parvizi, M. (2024). Investigation of the Ethical Agency of Artificial General Intelligence. *Science and Religion Studies*, *15*(1), 125-151. doi: 10.30465/srs.2024.49282.2162 [in

Persian]

Ashouri Kisomi, M. A. (2024). Convergence of privacy and transparency, limitations of artificial intelligence design. *Wisdom And Philosophy*, *20*(78), 45-73. doi: 10.22054/wph.2024.75680.2183 [in Persian]

Ashouri-Kisomi, M. A., Parvizi, M. & Momeni, S. M. H. (2025). Pragmatism, the Ethics of Artificial Intelligence and Sociotechnical Systems. *Journal of Philosophical Investigations*, *19*(52), 281-300. doi: 10.22034/jpiut.2025.66643.4060 [in Persian]

Hoseini Souraki, S. M. (2024). The Meaning and Criteria of Moral Agency in Intelligent Machines. *Journal of Philosophical Theological Research*, *26*(4), 83-108. doi: 10.22091/jptr.2025.12466.3256 [in Persian]

استناد به این مقاله: عاشوری کیسمی، محمد علی. (۱۴۰۵). طراحی حساس به ارزش و هم‌سوسازی هوش مصنوعی: روایتی پراگماتیستی از سیستم‌های اجتماعی‌فنی، *فصلنامه علمی حکمت و فلسفه*، ۲۲(۸۵)، ۱۵۱-۱۸۶.
doi: 10.22054/wph.2026.88857.2350