Socio-Cultural Strategy
Center for Strategic Research
https://rahbordfarhangi.csr.ir
E-ISSN: 2588-655X
Vol. 15, Issue. 2, No. 60, 2026
Received: 2026/02/12 **Revised:** 2026/07/28 **Accepted:** 2026/07/28
Research paper

A Value-Based Artificial Intelligence Governance Model in Iran Based on the National AI Document

Mohammad Hossein Zarifian Yeganeh
Assistant Professor, Department of Land Management, Faculty of Management, Farabi Campus, University of Tehran, Tehran, Iran.
Email: mhzarifian@ut.ac.ir
0009-0006-6080-4563

Abstract

The rapid evolution of artificial intelligence and its growing penetration into governance domains have doubled the necessity of redesigning policy-making and regulatory models. In Iran, the "National Artificial Intelligence Document" is considered the first upstream document for aligning technological development with Islamic-Iranian values, social justice, and data sovereignty. Nevertheless, the gap between value principles and executive mechanisms necessitates the presentation of a coherent and value-based model. Drawing on the theory of "Value-Sensitive Design" and thematic analysis of documents and interviews with nine experts, the present study elucidates a value-based AI governance model for Iran. Data were collected through document analysis and semi-structured interviews and coded using thematic analysis. Findings reveal that AI governance in Iran faces challenges such as "the lack of an independent governing institution," "technical complexity and the transparency crisis of models," and "the multiplicity of parallel institutions." Based on the analyses, the proposed model consists of three layers—strategic, operational, and supervisory—which respectively focus on the "core of ethics, awareness, and human dignity," "dynamic and participatory regulation," and "algorithmic transparency and accountability." By translating values such as justice, human dignity, and data sovereignty into executive mechanisms, this model lays the groundwork for responsible and trust-building AI development. As stated in the article: "The National Document seeks an indigenous, justice-oriented, and human-centered AI" and "Every AI project must first obtain ethical clearance." The proposed model can serve as a basis for transitioning from fragmented approaches toward integrated, learning-oriented, and value-based governance in Iran.

Keywords: Artificial intelligence governance, value-based governance, National Artificial Intelligence Document, data sovereignty and algorithmic transparency, media management.

Socio-Cultural Strategy Center for Strategic Research

E-ISSN: 2588-655X https://rahbordfarhangi.csr.ir

Vol. 15, Issue. 2, No. 60, 2026

The Socio-cultural Strategy Journal of Rahbord

doi 10.22034/scs. 2026.575326.1794

Mohammad Hossein Zarifian Yeganeh

Extended abstract

Purpose

Artificial intelligence, as a general-purpose and multi-faceted technology, has today transcended its role as a mere computational tool and has become a driving force for transformation in the strategic layers of power, economy, health, judiciary, and security. The rapid penetration of deep learning algorithms and autonomous decision-making systems has confronted the nature of modern governance with an unprecedented challenge, as the pace of technological evolution far exceeds the adaptive capacity of traditional laws and institutions. In such a context, the "black box" phenomenon in complex neural models, the reproduction of systematic biases, and large-scale privacy violations threaten the stability of contemporary societies, rendering the redefinition of governance a vital imperative. In the Islamic Republic of Iran, the approval of the "National Artificial Intelligence Document" in 2024, as the highest policy document of the country, constitutes a strategic starting point for steering technological development based on Islamic-Iranian values, social justice, data sovereignty, and human dignity. This ten-year document, emphasizing goals such as scientific authority, self-reliance, and the creation of an innovative ecosystem, outlines a value-oriented roadmap. However, the profound gap between the value principles enshrined in this document and concrete operational mechanisms represents the most significant challenge to achieving its objectives. The lack of an independent and cross-sectoral governing body, the inherent complexity of unexplainable models, the multiplicity of parallel institutions, and the predominance of purely technical developmental approaches over ethical and social considerations are among the structural and cognitive barriers that disrupt national policy coherence and heighten the risk of a social acceptability crisis for technology. The principal purpose of this study is therefore to design and formulate a coherent, indigenous, and operationalized model for value-based artificial intelligence governance in Iran. To this end, the paper pursues three questions: (1) what are the key challenges facing AI governance in Iran; (2) how can the abstract values of the National AI Document be translated into concrete executive mechanisms; and (3) what operational model can bridge the gap between value principles and governance mechanisms?

Design/Methodology/Approach

This research is developmental in purpose and qualitative in nature. Data were collected from two main sources: first, library research and content

analysis of upstream documents, particularly the National AI Document; and second, semi-structured interviews with nine experts in the fields of AI, governance, and media management, who were selected through purposive sampling based on the criteria of having at least three years of relevant experience and expertise. The interview process continued until theoretical saturation was achieved. The data were analyzed using thematic analysis with a theory-driven coding approach (informed by the theoretical framework of "Value-Sensitive Design" and the provisions of the National Document), conducted manually. The validity of the research was ensured through triangulation of data sources (documents and interviews), providing feedback on the results to some participants, and offering a rich description of the analysis process. Its reliability was established by presenting the findings to an independent expert in the relevant field.

Findings

Data analysis led to the identification of 37 initial codes, 12 basic themes, and ultimately 3 organizing themes, which were organized into the following three-layer model:

- First Layer (Strategic Core) – "Core of Ethics, Awareness, and Human Dignity": This layer emphasizes ethical institutionalization (including the establishment of a Supreme Council for Technology Ethics), the development of an Ethical Impact Assessment (EIA) framework, and the requirement to obtain an ethical clearance as a prerequisite for budget allocation and market entry.
- Second Layer (Operational Layer) – "Dynamic and Participatory Regulation": This layer proposes the establishment of regulatory sandboxes (in areas such as health and finance) and multi-stakeholder participation (government, university, industry, and civil society) to reduce compliance costs and develop mature regulations based on real-world experience.
- Third Layer (Supervisory Layer) – "Algorithmic Transparency and Accountability": This layer focuses on the requirement for explainability and auditability of high-risk systems, the establishment of a national algorithmic monitoring and auditing system (using automated supervisory dashboards), and the creation of AI incident reporting databases and compensation mechanisms.

These three layers are interconnected through a continuous feedback loop of monitoring data and operational experiences, enabling learning-oriented and dynamic governance. The findings also reveal that AI governance in Iran faces challenges such as "the lack of an independent governing institution," "technical complexity and the transparency crisis of models," and "the multiplicity of parallel institutions." By translating values such as justice, human dignity, and data sovereignty into executive mechanisms, this model lays the groundwork for responsible and trust-building AI development.

Research Limitations/Implications

This research has several limitations that should be acknowledged. First, the reliance on a limited number of expert interviews (nine participants), while sufficient for theoretical saturation in qualitative research, may not capture the full diversity of perspectives across all relevant stakeholder groups. Second, the study's focus on the Iranian context and the National AI Document means that findings may not be directly transferable to other countries with different institutional, cultural, and technological environments. Third, the rapidly evolving nature of AI technology means that the proposed model may require periodic revision as new technological developments emerge. Fourth, the study does not empirically test the proposed model through implementation, which limits the ability to assess its practical feasibility and effectiveness. Accordingly, it is suggested that future research conduct pilot studies to test the proposed model in specific sectors (such as health or finance), expand the sample to include a wider range of stakeholders (including technology developers, civil society representatives, and international experts), conduct comparative studies examining AI governance models in other countries with similar value orientations, and develop quantitative indicators for measuring the effectiveness of the proposed governance mechanisms.

Practical Implications

The findings of this research are of significant importance for AI policymakers, technology regulators, and institutional leaders. The proposed three-layer model provides a practical framework for implementing the National AI Document by translating abstract values into concrete executive mechanisms. At the strategic level, the establishment of a Supreme Council for Technology Ethics and the requirement for ethical clearance as a prerequisite for budget allocation and market entry provide tangible institutional mechanisms for value-based governance. At the operational level, the establishment of regulatory sandboxes enables experimentation and learning while maintaining appropriate oversight. At the supervisory level, the requirement for explainability and auditability of high-risk systems, the establishment of a national algorithmic monitoring and auditing system, and the creation of AI incident reporting databases provide concrete mechanisms for transparency and accountability. Realizing this model requires the consolidation of a governance institution with a supra-governmental and cross-sectoral status, the development of technological diplomacy and active participation in international standardization, the enhancement of digital governance literacy at managerial levels, and targeted support for interdisciplinary research centered on ethics and technology.

Social Implications

From a social perspective, this research reveals that AI governance in Iran must move beyond purely technical approaches toward integrating ethical

and value-based considerations into the design, development, and deployment of AI systems. The proposed model, by emphasizing human dignity, social justice, transparency, and accountability, aims to build public trust in AI technologies and prevent the social acceptability crisis that has emerged in many countries due to the unchecked deployment of AI systems. By creating a dynamic balance between the "speed of innovation" and the "stability of law," and transforming the governance approach from a purely controlling stance to a guiding and responsibly empowering one, this model provides a platform for trust-building and justice-oriented AI development. At the level of overall policy, the central recommendation is to redefine AI governance as a living and learning ecosystem that, by integrating ethical rationality and artificial intelligence, takes steps toward the realization of an indigenous digital civilization and the transformation of technology into a servant of human dignity and a driver of sustainable national development.

Originality/Value

The primary innovation of this research lies in providing an operationalized and indigenous model for AI governance in Iran that translates the abstract values of the National AI Document into concrete components and tangible executive mechanisms (such as ethical clearance, regulatory sandboxes, algorithmic auditing, and legal accountability). In contrast to prior studies, which have either addressed AI governance in general terms without considering the Iranian context, or have focused on technical aspects without integrating value-based considerations, this research provides a systematic and culturally grounded framework based on the theory of "Value-Sensitive Design" and the specific provisions of the National AI Document. The identification of three interconnected governance layers—strategic, operational, and supervisory—and the proposal of specific mechanisms for each layer offer a novel contribution to the literature. The proposed model provides a basis for transitioning from fragmented approaches toward integrated, learning-oriented, and value-based governance in Iran. Its value accrues to AI policymakers, technology regulators, institutional leaders, researchers interested in AI governance and value-sensitive design, and civil society actors concerned with the ethical and social implications of AI technologies.

E-ISSN: 2588-655X / Center for Strategic Research / The Socio-cultural Strategy Journal of Rahbord
doi 10.22034/scs. 2026.575326.1794

راهبرد اجتماعی فرهنگی

سال پانزدهم، شماره دوم (پیاپی ۶۰)، تابستان ۱۴۰۵، صص. ۱۸۱-۲۱۴
تاریخ دریافت: ۱۴۰۴/۱۱/۲۳ تاریخ بازنگری: ۱۴۰۵/۰۵/۰۶ تاریخ پذیرش: ۱۴۰۵/۰۵/۰۶

مقاله پژوهشی

الگوی حکمرانی ارزش مبنای هوش مصنوعی در ایران مبتنی بر سند ملّی هوش مصنوعی

محمّد حسین ظریفیان یگانه
استادیار گروه مدیریت سرزمین دانشکده مدیریت دانشکدگان فارابی دانشگاه تهران، تهران، ایران
Email: mhzarifian@ut.ac.ir 0009-0006-6080-4563

چکیده

تحول شتابان هوش مصنوعی و نفوذ فزاینده آن در عرصه‌های حکمرانی، ضرورت بازطراحی الگوهای سیاست‌گذاری و تنظیم‌گری را دوچندان کرده است. در ایران، «سند ملّی هوش مصنوعی» نخستین سند بالادستی برای همسوسازی توسعه فناوری با ارزش‌های اسلامی-ایرانی، عدالت اجتماعی و حاکمیت داده به شمار می‌رود. بااین‌حال، شکاف میان اصول ارزشی و سازوکارهای اجرایی، نیازمند ارائه الگویی منسجم و ارزش‌مبنا است. پژوهش حاضر با اتکا به نظریه «طراحی حساس به ارزش‌ها» و تحلیل مضمون اسناد و مصاحبه با ۹ نفر از خبرگان، به تبیین الگوی حکمرانی ارزش مبنای هوش مصنوعی در ایران می‌پردازد. داده‌ها از طریق تحلیل اسناد و مصاحبه‌های نیمه‌ساختاریافته گردآوری و با روش تحلیل مضمون کدگذاری شده‌اند. یافته‌ها نشان می‌دهد که حکمرانی هوش مصنوعی در ایران با چالش‌هایی نظیر «فقدان نهاد حکمران مستقل»، «پیچیدگی فنی و بحران شفافیت مدل‌ها» و «تعدد نهادهای موازی» مواجه است. براساس تحلیل‌ها، الگوی پیشنهادی شامل سه لایه راهبردی، عملیاتی و نظارتی است که به ترتیب بر «هسته اخلاق، آگاهی و کرامت انسانی»، «تنظیم‌گری پویا و مشارکت‌جو» و «شفافیت و مسئولیت‌پذیری الگوریتمیک» تمرکز دارد. این الگو با تبدیل ارزش‌هایی همچون عدالت، کرامت انسانی و استقلال داده‌ای به سازوکارهای اجرایی، زمینه توسعه مسئولانه و اعتمادآفرین هوش مصنوعی را فراهم می‌کند. همان‌گونه که در متن مقاله آمده است: «سند ملّی به دنبال هوش مصنوعی بومی، عدالت‌محور و انسان‌مدار است» و «هر پروژه هوش مصنوعی ابتدا باید مجوز اخلاقی بگیرد». الگوی پیشنهادی می‌تواند مبنایی برای گذار از رویکردهای پراکنده به حکمرانی یکپارچه، یادگیرنده و ارزش‌محور در ایران باشد.

کلیدواژه‌ها: حکمرانی هوش مصنوعی، حکمرانی ارزش‌مبنا، سند ملّی هوش مصنوعی، حاکمیت داده و شفافیت الگوریتمی، مدیریت رسانه.

شاپای الکترونیک: ۲۵۸۸-۶۵۵X/ پژوهشکده تحقیقات راهبردی/ فصلنامه علمی راهبرد اجتماعی- فرهنگی

doi 10.22034/scs. 2026.575326.1794

۱. مقدمه و بیان مسئله

امروز، هوش مصنوعی از یک ابزار صرفاً محاسباتی به یک «فناوری تحول‌آفرین» تغییر ماهیت داده که تمام ارکان زیست‌جهان انسانی را تحت تأثیر قرار داده است (Batool et al, 2025, p.23) و نفوذ الگوریتم‌های یادگیری ماشین و سیستم‌های تصمیم‌ساز هوشمند در لایه‌های راهبردی قدرت، از مدیریت کلان اقتصادی تا فرایندهای قضایی و امنیتی، ضرورت بازتعریف مفهوم حکمرانی را بیش از هر زمان دیگری آشکار ساخته است (Tallberg et al, 2023, p.10).

مسئله بنیادین در این میان، ظهور یک «تناقض حکمرانی» است؛ از یک‌سو، سرعت نمایی تکامل تکنولوژی، چهارچوب‌های حقوقی و نظارتی سنتی را که دارای ماهیتی ایستا و خطی هستند، به چالش کشیده و از سوی دیگر، فقدان ریل‌گذاری صحیح می‌تواند منجر به بروز مخاطرات جبران‌ناپذیری در حوزه‌های اخلاقی، امنیتی و حقوق شهروندی گردد (ترابی، ۱۴۰۴: ۴۱-۴۰).

عدم شفافیت در فرایندهای استنتاجی مدل‌های پیچیده که از آن به‌عنوان پدیده «جعبه سیاه» یاد می‌شود، در کنار پتانسیل بازتولید سوگیری‌های سیستماتیک و نقض حریم خصوصی در ابعاد کلان، ازجمله چالش‌های اساسی است که ثبات جوامع مدرن را تهدید می‌کند (حسینی، ۱۴۰۳: ۳۶۶). بنابراین، ضرورت این پژوهش از آنجا ناشی می‌شود که حکمرانی کارآمد هوش مصنوعی دیگر تنها یک انتخاب سیاستی نیست، بلکه یک ضرورت گریزناپذیر برای صیانت از نظم عمومی و عدالت اجتماعی محسوب می‌شود (Atwood, 2025, p.23).

۲. چهارچوب مفهومی

۲-۱. حکمرانی ارزش‌مبنا

حکمرانی ارزش‌مبنا رویکردی است که در آن ارزش‌های بنیادین جامعه، شالوده طراحی سازوکارهای حکمرانی قرار می‌گیرد و مشروعیت حکمرانی را فراتر از ملاحظات فنی و کارایی‌محور، به دستگاه ارزشی حاکم پیوند می‌زند. این مفهوم پاسخی به کاستی‌های رویکردهای رویه‌ای است که در آن ارزش‌ها به سازوکارهای عملیاتی ترجمه نمی‌شوند. در حوزه هوش مصنوعی، با توجه به نفوذ فزاینده آن در عرصه‌های عمومی، حکمرانی ارزش‌مبنا مستلزم به‌کارگیری ارزش‌هایی چون عدالت، شفافیت، مسئولیت‌پذیری و کرامت انسانی به‌عنوان ستون‌های اصلی طراحی نظام حکمرانی

است (ویتمن و ماینهارت، ۲۰۲۵). هم‌راستایی ارزش‌های عمومی با نظام‌های هوش مصنوعی به تقویت اعتماد عمومی و استحکام پایه‌های دموکراتیک حکمرانی می‌انجامد (Wittmann & Meynhardt, 2025, p.2)؛ اما چالش اساسی، فقدان تعاریف عملیاتی از ارزش‌ها و نبود سازوکارهای سنجش تحقق آن‌ها است (Trik et al, 2025, p.8).

حکمرانی ارزش‌مبنا در نظام جمهوری اسلامی ایران به معنای اتکا و استناد به ارزش‌های اسلامی، انقلابی و ملّی در تدوین، اجرا و ارزیابی سیاست‌ها و برنامه‌های توسعه است. این رویکرد در اسناد بالادستی نظام به‌وضوح مشاهده می‌شود.

در قانون اساسی جمهوری اسلامی ایران (منصور، ۱۴۰۴: ۳)، بر حاکمیت ارزش‌های اسلامی (اصل دوم: ایمان به خدا، معاد، عدل، امامت)، اصالت دادن به کرامت انسانی و ارزش‌های اخلاقی تأکید و در سند چشم‌انداز ۱۴۰۴ (مجموعه سیاست‌های کلّی نظام، ۱۴۰۲: ۱۲) نیز به «تکیه بر اصول ارزشی اسلام» به‌عنوان یکی از ارکان اصلی، «معنویت، اخلاق و عزت» به‌عنوان ویژگی‌های جامعه ایرانی در افق چشم‌انداز و هدف‌گذاری برای الهام‌بخشی به جهان اسلام توجه شده است. در سیاست‌های کلّی نظام نیز، تأکید بر ارزش‌های اسلامی در حوزه‌های مختلف (اقتصادی، اجتماعی، فرهنگی، سیاسی) مشهود است. همچنین در سیاست‌های کلّی اقتصاد مقاومتی (مجموعه سیاست‌های کلّی نظام، ۱۴۰۲: ۱۲۷) بر عدالت، پیشگیری از اسراف و مصرف‌گرایی و در نقشه جامع علمی کشور (شورای عالی انقلاب فرهنگی، ۱۳۹۰) بر تولید علم بر پایه ارزش‌های اسلامی، هدف‌گذاری برای تمدن‌سازی اسلامی-ایرانی و توجه به تعالی اخلاقی و معنوی در کنار پیشرفت علمی تأکید شده است.

۲-۲. هوش مصنوعی

هوش مصنوعی به نظام‌های محاسباتی و الگوریتمی اطلاق می‌شود که قادر به انجام وظایف نیازمند هوش انسانی مانند یادگیری، استدلال و تصمیم‌گیری هستند و به‌عنوان یک فناوری عمومی و چندمنظوره، فرایندهای اجتماعی، اقتصادی و سیاسی را بازتعریف می‌کنند (اکبری و جلائی‌نیا، ۱۴۰۴). گسترش شتابان این فناوری با دغدغه‌های بنیادین اخلاقی، حقوقی و اجتماعی همراه بوده است (Agarwal & Nene, 2025). در ایران، با وجود تدوین اسناد بالادستی، چالش‌هایی نظیر ضعف زیرساخت‌های قانونی، کمبود نیروی متخصص و نبود چهارچوب‌های مدون برای ارزیابی پیامدهای اخلاقی و اجتماعی، از موانع حکمرانی مؤثر این فناوری به شمار می‌روند.

۲-۳. حکمرانی هوش مصنوعی

حکمرانی هوش مصنوعی به مجموعه سیاست‌ها، مقررات، چهارچوب‌های اخلاقی و سازوکارهای نهادی برای مدیریت و نظارت بر توسعه و به‌کارگیری این فناوری به شیوه‌ای اخلاقی، شفاف و پاسخ‌گو اطلاق می‌شود (صیادی، ۱۴۰۲). این مفهوم ابعاد چندگانه‌ای ازجمله تنظیم‌گری فنی، حکمرانی داده و حکمرانی اخلاقی را در برمی‌گیرد (آقایاری، ۱۴۰۳). در سطح جهانی، الگوهای متنوعی مانند مدل ریسک‌محور اتحادیه اروپا و رویکرد بخشی‌محور ایالات‌متحده شکل گرفته است و پژوهش‌ها نشان می‌دهند الگوی واحدی برای حکمرانی وجود ندارد (Prasad et al, 2025, p.18). چالش اساسی، شکاف میان اصول کلان و سازوکارهای عملی است که چهارچوب‌های چندلایه در پی پُرکردن آن هستند (Agarwal & Nene, 2025). در ایران، این حوزه با چالش‌هایی مانند فقدان قوانین اختصاصی و تعدد مراجع تصمیم‌گیری مواجه است.

۲-۴. سند ملّی هوش مصنوعی

حکمرانی هوش مصنوعی در ایران تحت تأثیر «سند ملّی هوش مصنوعی جمهوری اسلامی ایران» (شورای عالی انقلاب فرهنگی، ۱۴۰۳، ص ۱.) از رویکردی غایت‌گرایانه و اخلاق‌مدار پیروی می‌کند. براساس این سند، هوش مصنوعی نباید صرفاً برای رشد اقتصادی، بلکه باید در جهت «کرامت انسانی» و «عدالت اجتماعی» به‌کار گرفته شود. این رویکرد با نظریه «خیر عمومی» در حکمرانی مطابقت دارد که براساس آن، دولت‌ها موظف‌اند تکنولوژی را به نفع تمامی اقشار جامعه جهت‌دِهی کنند.

این سند، صرفاً یک سند فنی نیست، بلکه یک نقشه راه ارزش‌محور است که تلاش می‌کند توسعه فناوری را با مبانی فکری و فرهنگی کشور هم‌سو کند. این سند بر چندین ارزش کلیدی تأکید دارد که به‌صورت دقیق و مصداقی عبارت‌اند از: «کرامت انسانی»[1] و اصالت اراده، «عدالت اجتماعی و نفی تبعیض»[2]، «استقلال و حاکمیت ملّی»[3]، «رعایت اخلاق اسلامی»[4]، «حریم خصوصی و امنیت داده»[5]، «شفافیت و پاسخ‌گویی»[6]، «بهره‌وری و رشد اقتصادی»[7].

1. Human Dignity
2. Justice & Non-discrimination
3. National Sovereignty
4. Islamic Ethics
5. Privacy & Security
6. Accountability
7. Economic Productivity

سند ملّی هوش مصنوعی جمهوری اسلامی ایران که در سال ۱۴۰۳ به‌عنوان بالاترین سند سیاستی کشور در حوزه هوش مصنوعی تصویب شد، نقشه راه ده‌ساله توسعه و کاربست این فناوری را با هدف دستیابی به مرجعیت علمی و فناورانه ترسیم می‌کند. محورهای اصلی آن شامل ایجاد زیرساخت‌های داده‌ای و محاسباتی، توسعه منابع انسانی، تقویت زیست‌بوم نوآوری و تدوین چهارچوب‌های حقوقی و اخلاقی است. مبانی ارزشی سند بر اخلاق اسلامی، عدالت، کرامت انسانی و خوداتکایی تأکید دارد (رئیسی، ۱۴۰۴). بااین‌حال، نقدهایی چون فقدان اولویت‌بندی در اهداف، نبود برنامه برای حوزه‌های پُرریسک و عدم تصریح سازوکارهای عملیاتی برای پیاده‌سازی اصول ارزشی بر آن وارد است (اکبری و جلائی‌نیا، ۲۰۲۵). این سند نقطه عزیمتی راهبردی برای طراحی الگوی بومی حکمرانی هوش مصنوعی در ایران محسوب می‌شود.

مفاهیم چهارگانه در رابطه‌ای طولی قرار دارند: «حکمرانی ارزش‌مبنا» شالوده نظری را فراهم می‌آورد و در بستر «هوش مصنوعی» به «حکمرانی هوش مصنوعی» شکل می‌دهد. «سند ملّی هوش مصنوعی» به‌عنوان حلقه واسط، اصول ارزشی و اولویت‌های ملی را صورت‌بندی می‌کند. حلقه مفقوده این زنجیره، نبود الگویی برای گذار از اصول ارزشی سند به سازوکارهای عملی حکمرانی است. پرسش اصلی مقاله، چگونگی طراحی الگویی منسجم، بومی و عملیاتی از حکمرانی هوش مصنوعی مبتنی بر سند ملّی است که از انسجام نظری و قابلیت انطباق پویا برخوردار باشد.

۳. مبانی نظری پژوهش

چهارچوب نظری این پژوهش، «نظریه طراحی حساس به ارزش»[1] است؛ رویکردی نظام‌مند که نخستین بار توسط بتیا فریدمن و پیتر کان در اواخر دهه ۱۹۸۰ و اوایل دهه ۱۹۹۰ میلادی در دانشگاه واشینگتن مطرح شد (فریدمن و کان، ۲۰۰۳) و به‌تدریج به یکی از جامع‌ترین چهارچوب‌ها برای لحاظ‌سازی ارزش‌های انسانی در طراحی فناوری تبدیل گردید (Friedman & Hendry, 2019, p. 33). هدف بنیادین این نظریه، ورود فعالانه و اصولی ارزش‌های اخلاقی و اجتماعی به فرایند طراحی فناوری است، به‌گونه‌ای که ارزش‌ها نه به‌عنوان امری حاشیه‌ای، بلکه از همان مراحل آغازین و در سراسر چرخه طراحی مورد توجه قرار گیرند (Friedman, 1996, p.27). سه رکن این

1. Value Sensitive Design

نظریه عبارت‌اند از: «شفافیت»[1] نفی تبعیض الگوریتمی، «پاسخ‌گویی»[2] (Dignum .(Dignumm, 2019, p.112

در پژوهش حاضر، این چهارچوب نظری به‌طور خاص برای تحلیل و طراحی «الگوی حکمرانی ارزش مبنای هوش مصنوعی در ایران» به‌کار گرفته می‌شود و حلقه واسطی میان «سند ملّی هوش مصنوعی جمهوری اسلامی ایران» و سازوکارهای حکمرانی فناورانه است. سند ملّی هوش مصنوعی ایران، چشم‌انداز راهبردی خود را بر پایه ارزش‌ها و آرمان‌های برآمده از جهان‌بینی اسلامی تعریف کرده و با تأکید بر توحیدگرایی، استقرار عدالت و امنیت فراگیر، جایگاه ایران را در جمع ۱۰ کشور پیشرو در افق ۱۰ ساله ترسیم می‌کند (شورای عالی فضای مجازی، ۱۴۰۳). بدین‌ترتیب، این چهارچوب نظری به پژوهشگر امکان می‌دهد تا با رویکردی نظام‌مند، فرایند طراحی حکمرانی هوش مصنوعی ایران را از مرحله شناسایی ارزش‌ها تا مرحله تجسّم و پیاده‌سازی آن‌ها در ساختارهای حکمرانی تحلیل و بازطراحی کند.

۴. ادبیات پژوهش

تحقق حکمرانی کارآمد صرفاً در گرو تدوین اسناد نیست و در حکمرانی هوش مصنوعی نیازمند درک دینامیک و پیوند ارگانیک میان مؤلفه‌های پنج‌گانه ذیل هستیم:

۴-۱. «تنظیم‌گری انعطاف‌پذیر»[3]

تنظیم‌گری در حوزه هوش مصنوعی با چالش سرعت تکامل مواجه است. تحلیل علمی نشان می‌دهد که مدل‌های سنتی قانون‌گذاری به‌دلیل ماهیت ایستا، در برابر خصلت «خودآموز»[4] الگوریتم‌ها ناکارآمد هستند. بنابراین حکمرانی کارآمد باید به سمت تنظیم‌گری «مبتنی بر ریسک»[5] حرکت کند. در این مدل، قوانین به‌جای تمرکز بر محدودسازی کدنویسی، بر خروجی‌ها و اثرات اجتماعی تمرکزیافته و از طریق صندوق‌های شبیه‌ساز قانونی، اجازه نوآوری تحت نظارت را صادر می‌کنند تا از خفگی تکنولوژیک جلوگیری شود (Arrieta, 2020, p 100-101).

1. Transparency
2. Accountability
3. Adaptive Regulation
4. Self-Learning
5. Risk-Based

۴-۲. «مهندسی ارزش‌ها»[1]

در تحلیل عمیق‌تر، اخلاق در حکمرانی از حالت توصیفی به حالت «اخلاق در طراحی» تغییر یافته است. مسئله اصلی، تبدیل مفاهیم کیفی نظیر «عدالت» به پارامترهای کمی و ریاضیاتی (فرمولیزه کردن عدالت) در لایه‌های کدنویسی است. حکمرانی موظف است با تدوین استانداردهای ملّی، از بروز «بی‌طرفی کاذب» جلوگیری کند؛ چراکه الگوریتم‌ها به‌ندرت بی‌طرف هستند و اغلب سوگیری‌های پنهان داده‌های آموزشی را بازتولید می‌کنند. این رکن، ضامن بقای کرامت انسانی در مواجهه با عقلانیت ابزاری ماشین است (Kaminski, 2020, p.109).

۴-۳. پارادایم نوین مسئولیت‌پذیری

از منظر حقوقی، هوش مصنوعی چالش «شکاف مسئولیت»[2] را پدید آورده است. در سیستم‌های خودمختار، به‌دلیل پیچیدگی شبکه عصبی، تعیین دقیق منشأ خطا دشوار است. تحلیل علمی در این بخش پیشنهاد می‌دهد که حکمرانی باید از مدل‌های سنتی به سمت مدل‌های «مسئولیت مشترک» یا «بیمه اجباری تکنولوژی» حرکت کند. این رویکرد، علاوه بر حمایت از حقوق مصرف‌کننده، از توقف پروژه‌های بزرگ به‌دلیل ترس از دعاوی حقوقی پیچیده پیشگیری می‌نماید (Mittelstadt et al, 2016, p. 19).

۴-۴. حکمرانی چند ذینفعی

هوش مصنوعی یک «فناوری با کاربرد دوگانه» است، بنابراین حکمرانی آن نمی‌تواند در انحصار دولت‌ها باشد. تحلیل پژوهشی نشان می‌دهد که بهترین مدل، ایجاد «زیست‌بوم نظارتی» است که در آن دانشگاه‌ها (به‌عنوان دیدبان علمی)، جامعه مدنی (به‌عنوان پاسدار حقوق شهروندی) و بخش خصوصی (به‌عنوان بازوی اجرایی) در تدوین استانداردها مشارکت کنند. این تکثرگرایی باعث می‌شود که قوانین در برابر تغییرات سریع بازار و نیازهای واقعی جامعه، از قدرت تطبیق بالایی برخوردار باشند (Shneiderman, 2022, p.77).

1. Value Engineering
2. Responsibility Gap

۴-۵. شفافیت الگوریتمیک

بزرگ‌ترین چالش فنی حکمرانی، پدیده «جعبه سیاه» در مدل‌های یادگیری عمیق است. تحلیل علمی در این حوزه بر ضرورت توسعه هوش مصنوعی تفسیرپذیر تأکید دارد. شفافیت در اینجا به معنای افشای کدهای منبع (که محرمانه است) نیست، بلکه به معنای ارائه منطقِ حاکم بر تصمیم‌گیری ماشین به زبانی قابل فهم برای انسان است. این رکن، زیربنای نظارت‌پذیری سامانه است؛ چراکه بدون درک چراییِ یک تصمیم، نه می‌توان اخلاق را رعایت کرد و نه مسئولیت را تعیین نمود (Shneiderman, 2022, p.79).

۵. پیشینهٔ پژوهش

۵-۱. پیشینه فارسی

مؤسسه پژوهش و برنامه‌ریزی آموزش عالی (۱۴۰۱). حکمرانی هوش مصنوعی: چهارچوب‌های بین‌المللی و اقتضائات داخلی. تهران: انتشارات مؤسسه.

اگرچه این اثر که به تحلیل تطبیقی چهارچوب‌های حکمرانی هوش مصنوعی در جهان و ارائه ملاحظاتی برای بومی‌سازی آن در ایران با توجه به ارزش‌های فرهنگی و دینی پرداخته است، ارزش‌های بومی را مدنظر دارد؛ اما عمدتاً تحلیلی مقایسه‌ای است و مدل عملیاتی مشخصی برای اجرای ارزش‌های اسلامی در حکمرانی هوش مصنوعی ارائه نمی‌دهد.

قمی، مهدی؛ میرعمادی، سید علی. (۱۴۰۱). اخلاق هوش مصنوعی در اسلام: مبانی و کاربردها. قم: انتشارات دانشگاه مفید.

این کتاب، به بررسی مبانی اخلاق اسلامی حاکم بر توسعه و استفاده از هوش مصنوعی و ارائه دستورالعمل‌های کاربردی می‌پردازد. البته از بعد "حکمت اسلامی" که ناظر به جهان‌بینی است، فاصله گرفته و بیشتر به احکام و بایدونبایدهای اخلاقی پرداخته است.

زارعی، محمد. (۱۳۹۸). حقوق و اخلاق هوش مصنوعی: رویکردی اسلامی-ایرانی. تهران: انتشارات دانشگاه تهران.

در این کتاب، تلاش برای استخراج اصول حقوقی و اخلاقی حاکم بر هوش مصنوعی از متون اسلامی و تطبیق آن با چالش‌های نوظهور مشهود است و تمرکز اصلی آن، بر اخلاق و حقوق است و ابعاد گسترده‌تر حکمرانی (مانند اقتصاد، امنیت، نهادسازی) را پوشش نمی‌دهد.

پژوهشکده سیاست‌گذاری علم، فناوری و صنعت دانشگاه صنعتی شریف. (۱۴۰۱). تحلیل سیاستی سند ملّی هوش مصنوعی ایران. تهران: انتشارات دانشگاه صنعتی شریف.

در این اثر، ارزیابی نقاط قوت، ضعف، فرصت‌ها و تهدیدهای سند ملّی هوش مصنوعی ایران و مقایسه آن با اسناد مشابه جهانی انجام شده است. اگرچه تحلیل سیاستی ارزشمندی ارائه می‌دهد، اما ممکن است نقدهای فنی و اجرایی را برجسته کند و کم‌تر به تحلیل «ارزش‌مبنا» بودن آن از منظر اسلامی بپردازد.

۵-۲. پیشینه انگلیسی

- Dignum, V. (2019). Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer.

در این اثر، چهارچوبی برای مسئولیت‌پذیری در توسعه و استفاده از هوش مصنوعی، با تأکید بر اخلاق، قابلیت توضیح و حکمرانی ارائه شده است. البته چهارچوب ارائه شده عمدتاً مبتنی بر ارزش‌های لیبرال-دموکراتیک غربی است و قابل تعمیم مستقیم به بافتار ارزشی ایران نیست.

- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

در این مقاله، بیش از ۸۰ سند اخلاقی هوش مصنوعی در سراسر جهان مرور سیستماتیک و تحلیل محتوا شده و نقاط مشترک و افتراق آن‌ها شناسایی شده است. این پژوهش، نشان می‌دهد که اغلب اسناد جهانی بر ارزش‌های مشابهی تأکید دارند، اما به نحوه عملّیاتی‌سازی این ارزش‌ها در بافتارهای فرهنگی-دینی خاص نمی‌پردازد.

- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.

این تحقیق، نقدی بر رهنمودهای اخلاقی هوش مصنوعی است که نشان می‌دهد بسیاری از آن‌ها فاقد مکانیسم‌های اجرایی و نظارتی کافی هستند. نقد ارائه شده می‌تواند به سند ملّی ایران نیز تعمیم یابد؛ اما این مقاله به‌صورت خاص به ارزش‌های غیرغربی یا دینی نمی‌پردازد.

- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review,. (1)1

در این اثر، چهارچوبی متحد شامل پنج اصل: بهروزی، فراگیری، احتیاط، توضیح‌پذیری و عدالت برای حکمرانی هوش مصنوعی. هرچند این چهارچوب جهانی

تلاش می‌کند جامع باشد؛ اما تفسیر و اولویت‌بندی این اصول (مثلاً عدالت) می‌تواند در بافتار اسلامی متفاوت باشد.

- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501-507.

این پژوهش، استدلال می‌کند که صرف تدوین اصول اخلاقی برای تضمین اخلاقمندی هوش مصنوعی کافی نیست و به نهادهای نظارتی، ارزیابی تأثیر و مکانیسم‌های اجرایی نیاز است. بهره‌ای که از این مقاله می‌توان برای تحلیل سند ملّی ایران برد این است که بررسی ما باید فراتر از اصول، به سازوکارهای اجرایی، نظارتی و ارزشیابی آن توجه کند.

۶. روش‌شناسی پژوهش

پژوهش حاضر را به لحاظ نوع، می‌توان در شمار مطالعات توسعه‌ای قلمداد کرد. روش پژوهش در این اثر، کیفی و بنابراین، فاقد فرضیه است. جمع‌آوری داده‌ها نیز از طریق مطالعات کتابخانه‌ای و مصاحبه صورت گرفته است. روش تجزیه‌وتحلیل داده‌ها، تحلیل مضمون (تم) بوده و ازآنجاکه در تحلیل مضمون، تمرکز و خلاقیت پژوهشگر اهمیت دارد. در مرحله اول، مطالعه ادبیات و مستندات در قلمرو موضوع پژوهش با مراجعه به منابع فارسی و انگلیسی انجام گرفت و سپس به‌موازات تدوین ادبیات پژوهش، با ۹ نفر از صاحب‌نظران و کارشناسان قلمرو هوش مصنوعی و مدیریت رسانه و حکمرانی، مصاحبه‌های نیمه‌ساختاریافته تا رسیدن به حدّ اشباع نظری انجام شد. معیار انتخاب مشارکت‌کنندگان، دانش و حداقل سه سال تجربه و مهارت آنان در موضوع پژوهش بود. در مرحله بعد، فهرستی اولیه از ایده‌های موجود در داده‌ها و اهم نکات آن‌ها تهیه شد. این مرحله، به تولید ۳۷ کد اولیه برای تقسیم داده‌های متنی، به عبارت‌های قابل استفاده، منجر شد. شایان ذکر است که کدگذاری به روش تئوری‌محور، با امعان نظر به نظریه «طراحی حسّاس به ارزش‌ها» و با الهام از اصول و مفاد سند ملّی هوش مصنوعی و به شکل دستی انجام گرفت. در گام بعد، کدهای اولیه تجزیه‌وتحلیل و ۱۲ مضمون پایه به دست آمد. سپس، از مضامین پایه، ۳ مضمون سازمان‌دهنده یعنی لایه‌های سه‌گانه که عبارت‌اند از: لایه اول (هسته مرکزی): لایه راهبردی؛ هسته اخلاق، آگاهی و کرامت انسانی، لایه دوم (لایه میانی): لایه عملیاتی؛ تنظیم‌گری پویا و مشارکت‌جو و لایه سوم (لایه بیرونی): لایه نظارتی؛ شفافیت و مسئولیت‌پذیری الگوریتمیک استخراج شدند. این لایه‌ها مبتنی بر مفهوم اصلی یعنی «الگوی حکمرانی

ارزش مبنای هوش مصنوعی در ایران مبتنی بر سند ملّی هوش مصنوعی» صورت‌بندی شدند که با جزییات در شکل (۱) به تصویر کشیده شده است. به‌منظور شفاف‌سازی فرایند گذار از داده‌های خام به مضامین نهایی و افزایش قابلیت ردیابی نتایج، جزئیات کدگذاری در جدول (۱) ارائه شده است. کدگذاری در این پژوهش بر پایه دو منبع مکمّل انجام گرفت: نخست، سند ملّی هوش مصنوعی جمهوری اسلامی ایران به‌عنوان چهارچوب مرجع نظری و ارزشی که کدگذاری تئوری‌محور بر مبنای مواد و بندهای آن انجام شد؛ و دوم، مصاحبه‌های نیمه‌ساختاریافته با ۹ خبره حوزه هوش مصنوعی، حکمرانی و مدیریت رسانه که کدهای برخاسته از تجربه زیسته و دانش تخصصی ایشان را در اختیار پژوهش قرار دادند. در جدول (۱)، برای هر مضمون پایه، هم مبنای استناد در سند ملّی به‌تفکیک ماده و بند مشخص شده و هم خبرگانی که در جریان مصاحبه بر آن مضمون تأکید داشته‌اند، شناسایی گردیده‌اند. این ساختار، هم‌گرایی میان دو منبع داده را به‌عنوان نشانه‌ای از مثلث‌سازی روش‌شناختی و اعتبار درونی الگو نشان می‌دهد. جزئیات کامل فرایند کدگذاری، شامل ۳۷ کد اولیه، ۱۲ مضمون پایه و انطباق آن‌ها با مواد سند ملّی و تأیید خبرگان، در جدول (۱) بخش یافته‌های پژوهش ارائه شده است.

و این فرایند طی‌شده، به مدل پیشنهادی اجرایی که مبتنی بر سه لایه یاد شده و هر یک مشتمل بر هدف، راهکار متناظر، پیامد و دستاورد است منجر شد. این مدل در شکل (۲) ترسیم شده است.

روایی این پژوهش، علاوه بر این‌که بر مضامین فراگیر، مضامین سازمان‌دهنده و مضامین پایه مبتنی بر مبانی نظری تحقیق استوار است، بر استفاده از دیدگاه‌ها و رهنمودهای تعداد قابل‌توجهی از خبرگان تکیه دارد. هم‌سو با تقویت اعتبار پژوهش، نتایج به برخی از مصاحبه‌شوندگان، ارائه و بازخورد گرفته شد. همچنین براساس اصول روش تحلیل مضمون، تلاش شد تا در متن پژوهش، «توصیف غنی» انجام گیرد تا نحوه دستیابی این پژوهش از داده‌ها به نتایج مشخص شود. چنان‌که گفته شد تا مرحله اشباع نظری فرایند کدگذاری به شکل دستی انجام گرفته که اعتبار نتایج را افزایش می‌دهد و پایایی آن نیز با ارجاع پژوهش، فرایند طی‌شده و نتایج به‌دست‌آمده به «خبره بی‌طرف مرتبط» با موضوع، حاصل شده است.

۷. یافته‌های پژوهش

براساس فرایند تحلیل مضمون که در بخش روش پژوهش تشریح شد، داده‌های گردآوری‌شده از دو منبع سند ملّی هوش مصنوعی جمهوری اسلامی ایران و مصاحبه‌های نیمه‌ساختاریافته با ۹ خبره، به شناسایی ۳۷ کد اولیه، ۱۲ مضمون پایه و ۳ مضمون سازمان‌دهنده منجر شد که ذیل مضمون فراگیر «الگوی حکمرانی ارزش مبنای هوش مصنوعی در ایران مبتنی بر سند ملّی هوش مصنوعی» سازمان‌دِهی گردید. جدول (۱) ساختار کامل مضامین استخراج‌شده و انطباق آن‌ها با دو منبع مرجع پژوهش را نمایش می‌دهد.

مضمون سازمان‌دهنده (لایه)	مضمون پایه	کدهای اولیه	مبنای استناد در سند ملّی هوش مصنوعی	تأیید در مصاحبه خبرگان
لایه اول (هسته مرکزی): لایه راهبردی؛ هسته اخلاق، آگاهی و کرامت انسانی	۱. مبنای ارزشی حکمت اسلامی و قانون اساسی	اتکا به جهان‌بینی توحیدی؛ توجه به ابعاد وجودی انسان بر پایه فلسفه اسلامی؛ تمدن‌سازی نوین اسلامی؛ پیوند با حیات طیبه	مقدمه سند؛ ماده ۲، بندهای ۱ و ۲	خبرگان ۱، ۳، ۵
	۲. نهادسازی اخلاقی (شورای عالی اخلاق فناوری)	ضرورت نهاد مرجع اخلاقی؛ ترکیب بین‌رشته‌ای (فقه، فلسفه اخلاق، فنی، اجتماعی)؛ استقلال نهاد از دستگاه‌های اجرایی	ماده ۱ (تعریف اخلاق هوش مصنوعی)؛ مقدمه سند	خبرگان ۱، ۵، ۷
	۳. چهارچوب ارزیابی تأثیر اخلاقی	سنجش تأثیر بر کرامت و اراده انسانی؛ سنجش تبعیض در داده‌های آموزشی؛ سنجش ریسک سلطه فناوری بر انسان؛ سنجش تأثیر بر	ماده ۲، بندهای ۵، ۱۰ و ۱۱	خبرگان ۲، ۴، ۶

مضمون سازمان‌دهنده (لایه)	مضمون پایه	کدهای اولیه	مبنای استناد در سند ملّی هوش مصنوعی	تأیید در مصاحبه خبرگان
		انسجام اجتماعی		
	۴. مجوز اخلاقی به‌عنوان پیش‌شرط تخصیص منابع	الزام مجوز اخلاقی برای پروژه‌های عمومی؛ الزام مجوز برای ورود به بازار؛ پیوند مجوز با نظام تخصیص بودجه	ماده ۱ (تعریف اخلاق هوش مصنوعی)	خبرگان ۲، ۶، ۸
لایه دوم (لایه میانی) :لایه عملیاتی؛ تنظیم‌گری پویا و مشارکت‌جو	۵. حوزه‌های آزمایشی مقرراتی (سندباکس)	ایجاد محیط‌های ایزوله برای نوآوری امن؛ معافیت موقت از مقررات سخت در ازای شفافیت؛ یادگیری تنظیم‌گر پیش از مقیاس‌دهی	مقدمه سند (تسهیل فضای کسب‌وکار)؛ ماده ۳، بند ۴ اهداف	خبرگان ۲، ۹
	۶. نهاد متولی تنظیم‌گری پویا	استقرار سازمان ملی هوش مصنوعی؛ رفع موازی‌کاری میان نهادها؛ جایگاه فرابخشی نهاد حکمران	ماده ۱ (تعریف سازمان)؛ ماده واحده تأسیس سازمان	خبرگان ۱، ۳، ۷
	۷. مشارکت چند ذینفعی در تنظیم‌گری	مشارکت دولت، دانشگاه، صنعت و جامعه مدنی؛ تشکیل کنسرسیوم ذی‌نفعان؛ توجه به همه بازیگران زیست‌بوم	ماده ۱ (تعریف زیست‌بوم هوش مصنوعی)	خبرگان ۴، ۵
	۸. تدوین مقررات بالغ مبتنی بر تجربه واقعی	کاهش هزینه انطباق برای دانش‌بنیان‌ها؛ تدوین قواعد بخشی	ماده ۳، بندهای ۴ و ۵ اهداف	خبرگان ۶، ۹

مضمون سازمان‌دهنده (لایه)	مضمون پایه	کدهای اولیه	مبنای استناد در سند ملّی هوش مصنوعی	تأیید در مصاحبه خبرگان
		(سلامت، مالی، قضایی)		
لایه سوم (لایه بیرونی) :لایه نظارتی؛ شفافیت و مسئولیت‌پذیری الگوریتمیک	۹. الزام به تفسیرپذیری الگوریتمی	ارائه منطق تصمیم‌گیری به زبان قابل فهم؛ تفکیک شفافیت از افشای کد منبع؛ الزام قابلیت توضیح در سیستم‌های پُرخطر	ماده ۱ (توضیح‌پذیری و شفافیت در تعریف اخلاق)	خبرگان ۲، ۳
	۱۰. سامانه ملّی نظارت و حسابرسی الگوریتمی	پنل‌های خودکار نظارتی؛ سنجش شاخص‌های سوگیری، دقت و انصاف؛ رویکرد «جعبه شیشه‌ای«	ماده ۱ (انصاف و عدم تبعیض)؛ ماده ۲، بند ۸	خبرگان ۴، ۵، ۸
	۱۱. سازوکار مسئولیت‌پذیری و جبران خسارت	دفاتر ثبت سانحه هوش مصنوعی؛ مکث اجباری بر سیستم‌های خطاآفرین؛ الزام سازنده به جبران خسارت	ماده ۱ (پاسخ‌گویی و مسئولیت‌پذیری)؛ ماده ۲، بندهای ۵ و ۱۰	خبرگان ۱، ۷
	۱۲. حلقه بازخورد و حکمرانی یادگیرنده	بازگشت داده‌های نظارتی به لایه راهبردی؛ انتقال تجربیات سندباکس به قانون‌گذاری؛ چرخه پویا میان سه لایه	ماده ۲، بند ۷ (زیست‌بوم آینده‌نگر)	خبرگان ۲، ۶، ۹

هدف این پژوهش دستیابی به «الگوی حکمرانی ارزش مبنای هوش مصنوعی در جمهوری اسلامی ایران» بود که مبتنی بر نظریه «طراحی حسّاس به ارزش‌ها» و با الهام از اصول و مفاد سند ملّی هوش مصنوعی طراحی شد. این الگو، مشتمل بر سه لایه است:

۷-۱. لایه اول (هسته مرکزی): لایه راهبردی؛ هسته اخلاق، آگاهی و کرامت انسانی

این لایه، فلسفه وجودی و جهت‌گیری کلان حکمرانی را تعیین می‌کند. مفهوم کلیدی در آن مبنای ارزشی ناظر به استخراج اصول اخلاقی از «حکمت اسلامی» (با تأکید بر عدالت، عقلانیت توحیدی، مسئولیت‌پذیری در برابر خدا و خلق) و «قانون اساسی» (اصول کرامت انسانی، عدالت، استقلال) است. پیشنهاد می‌شود، نهاد متولی تحقق این لایه، ستاد راهبری اجرای سند ملّی هوش مصنوعی، با مشاوره «شورای عالی اخلاق فناوری» متشکل از فقها، فیلسوفان اخلاق، جامعه‌شناسان و متخصصان فنی و ابزار اجرایی این لایه، چهارچوب ارزیابی تأثیر اخلاقی (EIA) هوش مصنوعی است. چه این‌که، هر پروژه یا سامانه هوش مصنوعی پیش از اجرا، موظف است پرسش‌نامه‌ای عمیق را پاسخ دهد: آیا این سامانه به استقلال فرد لطمه می‌زند؟ آیا تبعیض نظام‌مند ایجاد می‌کند؟ آیا کرامت ذاتی انسان را نقض می‌نماید؟ و آیا به هم‌بستگی اجتماعی آسیب می‌رساند؟

خروجی این لایه، صدور «مجوز اخلاقی» به‌عنوان شرط لازم برای تخصیص بودجه دولتی، مجوز بازار یا استفاده در بخش عمومی است.

۷-۲. لایه دوم (لایه میانی): لایه عملیاتی؛ تنظیم‌گری پویا و مشارکت‌جو

این لایه، فضای آزمایشی و تطبیقی برای نوآوری امن و هم‌سو با ارزش‌ها فراهم می‌آورد.

مفهوم کلیدی در آن، ایجاد «حوزه‌های آزمایشی مقرراتی»[1] یا «صندوق‌های شبیه‌ساز قانونی» است. این حوزه‌ها، محیط‌های ایزوله و زمانمند برای آزمایش محصولات نوآورانه هوش مصنوعی تحت نظارت نهاد ناظر هستند. نهاد متولی این لایه می‌تواند مرکز نوآوری و تنظیم‌گری پویای هوش مصنوعی زیرمجموعه معاونت علمی و فناوری ریاست جمهوری باشد. ابزار اجرایی می‌توانند شرکت‌ها و پژوهشگاه‌ها باشند

1. Regulatory Sandboxes

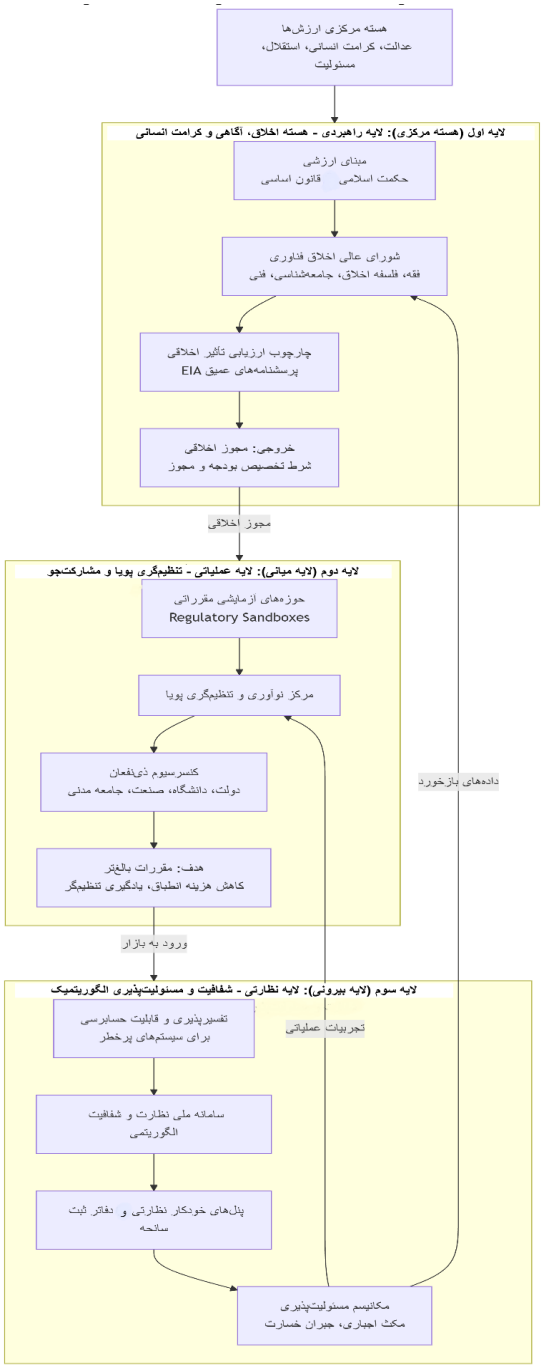
و در ازای معافیت موقت از برخی مقررات سخت، موظف به ثبت دقیق داده‌ها، گزارش‌دهی شفاف و پذیرش نظارت مستمر توسط یک کنسرسیوم از ذی‌نفعان (نماینده دولت، دانشگاه، صنعت و نهادهای مدنی مانند شوراهای شهر) می‌شوند.

پیامد اجرای سازوکار لایه تنظیم‌گر که پویا و مشارکت‌جو باشد، کاهش هزینه‌های انطباق برای استارتاپ‌ها، یادگیری تنظیم‌گر از پیچیدگی‌های فناوری و تدوین مقررات بالغ‌تر و عملّی‌تر براساس داده‌های واقعی، پیش از ورود به مقیاس کلان خواهد بود.

۷-۳. لایه سوم (لایه بیرونی): لایه نظارتی؛ شفافیت و مسئولیت‌پذیری الگوریتمیک

این لایه، سامانه پاسخ‌گویی و ممیزی مستمر پس از ورود فناوری به بازار را تضمین می‌کند. مفهوم کلیدی این لایه، الزام به «تفسیرپذیری»[1] و «قابلیت حسابرسی»[2] برای سیستم‌های هوش مصنوعی پُرخطر (مانند سیستم‌های قضایی، مالی، سلامت) است. پیشنهاد می‌شود، سامانه ملّی نظارت و شفافیت الگوریتمی زیر نظر سازمان تنظیم مقررات و ارتباطات رادیویی، عهده‌دار اجرای این لایه شود. ابزار اجرایی این لایه، استقرار «پنل‌های خودکار نظارتی»[3] است. این پنل‌ها به‌طور پیوسته عملکرد الگوریتم‌ها را رصد می‌کنند و شاخص‌هایی مانند «سوگیری»، «دقت»، «انصاف» و «تأثیر اجتماعی» را اندازه‌گیری می‌نمایند. بخشی از کد یا منطق تصمیم‌گیری سیستم‌های حیاتی، باید به‌صورت «جعبه شیشه‌ای» در اختیار نهاد ناظر قرار گیرد. «دفاتر ثبت سانحه هوش مصنوعی» مشابه سامانه‌های گزارش خطا در پزشکی می‌تواند سازوکار مسئولیت‌پذیری را شکل دهد تا هر خطا، آسیب یا تصمیم بحث‌برانگیز ثبت و علت‌یابی شود. در صورت لزوم، «مکث اجباری» بر سامانه اعمال شده و سازنده موظف به ارائه توضیح و جبران خسارت می‌شود.

1. Explainability
2. Auditability
3. Automated Dashboard



شکل ۱: الگوی حکمرانی ارزش مبنای هوش مصنوعی در ایران

۷-۴. نقشه ارتباطی و یکپارچه الگو

۱. جریان از داخل به خارج: هر پروژه هوش مصنوعی ابتدا باید مجوز اخلاقی (لایه ۱) بگیرد. سپس می‌تواند برای توسعه در صندوق شبیه‌ساز (لایه ۲) اقدام کند. پس از خروج موفق از صندوق و ورود به بازار، تحت نظارت مستمر شفافیت‌محور (لایه ۳) قرار می‌گیرد؛
۲. حلقه بازخورد: داده‌های جمع‌آوری شده از لایه نظارتی (۳) و تجربیات لایه عملیاتی (۲)، به‌طور مستمر به‌روزرسانی و تقویت چهارچوب ارزیابی اخلاقی (۱) را سبب می‌شوند. این یک حکمرانی یادگیرنده و پویا ایجاد می‌کند؛
۳. تجلی ارزش‌ها: این الگو عملّیاتی‌سازی می‌کند:
   - ❖ عدالت: از طریق شناسایی و کاهش سوگیری در لایه‌های (۱) و (۳)؛
   - ❖ کرامت انسانی: با قرار دادن آن به‌عنوان معیار نهایی در ارزیابی اخلاقی لایه (۱)؛
   - ❖ استقلال و پیشرفت علمی: از طریق حمایت از نوآوری مسئولانه در لایه (۲)؛
   - ❖ نظم و مسئولیت‌پذیری: از طریق شفافیت و مکانیسم‌های پاسخ‌گویی در لایه (۳).

این الگو، چهارچوبی جامع ارائه می‌دهد که هم از پیچیدگی غیرضروری قوانین جزئی‌نگر جلوگیری می‌کند و هم از آسیب‌های اخلاقی و اجتماعی فناوری رها شده ممانعت به عمل می‌آورد. این مدل، حکمرانی هوش مصنوعی در ایران را از یک رویکرد صرفاً «کنترلی» به یک رویکرد «هدایتی و توانمندساز مسئولانه» تبدیل می‌نماید.

۷-۵. راهبردهای بهبود حکمرانی هوش مصنوعی در ایران

برای گذار از وضعیت فعلی به حکمرانی مطلوب، اقدامات راهبردی زیر نه به‌صورت پراکنده، بلکه به‌عنوان یک منظومه واحد پیشنهاد می‌شود:

۷-۵-۱. تثبیت نهاد با جایگاه حاکمیتی و رویکرد فراقوه‌ای

تنظیم‌گری هوش مصنوعی به‌دلیل ماهیت فرابخشی (زیر بنایی و اقتصادی، آموزشی و فرهنگی، علمی و فناوری، حقوقی و قضایی، بهداشت و سلامت، امنیت و ...)، نمی‌تواند زیرمجموعه یک وزارتخانه خاص و حتی فقط یکی از قوا باشد. نتایج بررسی‌های این

تحقیق نشان می‌دهد درمجموع، شکل‌گیری مرکزی هم‌عرض با مرکز ملّی فضای مجازی ذیل شورای عالی فضای مجازی تصمیمی متقن و دقیق است. این گزینه مزیت‌های متعددی ازجمله ویژگی فرابخشی بودن و احتمال کم‌تر تعارض منافع را دارد.

۷-۵-۲. گذار از تنظیم‌گری سخت به تنظیم‌گری پویا و مشارکتی

ایران باید از مدل‌های سنتی قانون‌گذاری که سال‌ها به طول می‌انجامد فاصله گرفته و به سمت تنظیم‌گری پویا و مشارکتی حرکت کند. این راهکار به معنای ایجاد محیط‌های آزمایشی قانونی است که در آن استارتاپ‌ها و شرکت‌های بومی بتوانند تحت نظارت مستقیم، ایده‌های نوآورانه خود را پیش از عرضه عمومی آزمایش کنند. این کار باعث می‌شود قانون‌گذار پابه‌پای فناوری حرکت کرده و موانع نوآوری را به‌صورت موردی شناسایی و رفع کند.

۷-۵-۳. پیوست اخلاقی و بازنگری الگوریتمیک در پروژه‌های ملّی

صرفِ داشتن الگوریتم کافی نیست؛ هر سامانه هوش مصنوعی که با داده‌های شهروندان در مقیاس وسیع سروکار دارد (مانند سامانه‌های بانکی، قضایی یا استخدامی)، باید ملزم به دریافت تأییدیه اخلاقی و فنی باشد. این راهکار شامل اجباری کردن قابلیت توضیح است؛ یعنی در صورت رد درخواست وام یک شهروند توسط هوش مصنوعی، سامانه باید بتواند منطق تصمیم‌گیری خود را به زبان انسانی توضیح دهد تا امکان اعتراض و بازنگری فراهم باشد. این اقدام مستقیماً با هدف «شفافیت» و «عدالت اجتماعی» در ارتباط است.

۷-۵-۴. حمایت از پژوهش‌های بین‌رشته‌ای و ایجاد کنسرسیوم‌های اخلاق فناوری

شکاف میان مهندسان فنی و متخصصان علوم انسانی (دین و اخلاق، حقوق، فلسفه و جامعه‌شناسی) یکی از ریشه‌های چالش‌های فعلی ایران است. دولت باید مشوق‌هایی برای طرح‌های پژوهشی مشترک تعریف کند که در آن، پیوست‌های حقوقی و اخلاقی به‌اندازه کدهای فنی اهمیت داشته باشند. ایجاد «کنسرسیوم‌های ملّی» با حضور دانشگاه‌های برتر، راه را برای تدوین منشور ملّی اخلاق هوش مصنوعی هموار می‌کند

که با ارزش‌های فرهنگی و بومی ایران هم‌خوانی داشته باشد.

۷-۵-۵. توسعهٔ دیپلماسی فناوری و مشارکت در استاندارد گذاری بین‌المللی

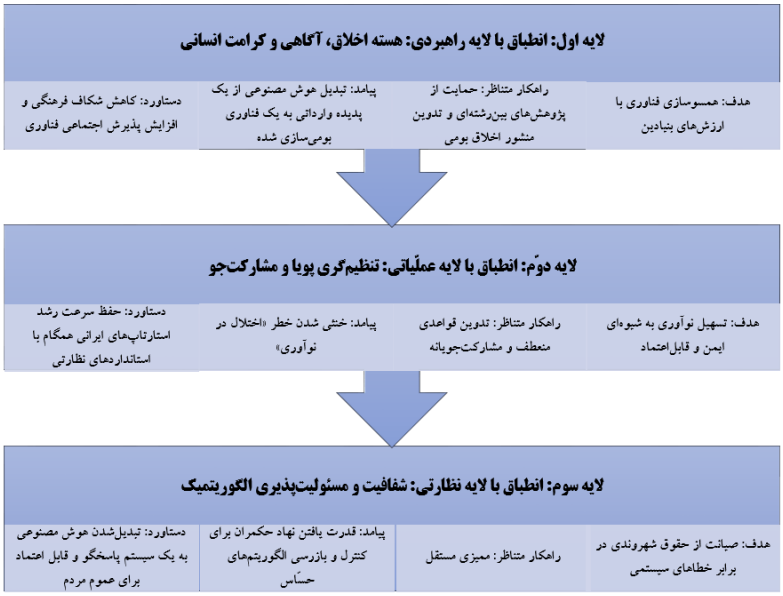
انزوای قانونی در حوزه هوش مصنوعی می‌تواند به معنای حذف از زنجیره ارزش جهانی باشد. ایران باید به‌طور فعال در سازمان‌هایی نظیر ITU (اتحادیه بین‌المللی مخابرات) و نشست‌های بین‌المللی یونسکو با موضوع «اخلاق هوش مصنوعی» حضور داشته باشد. هدف از این راهکار، اثرگذاری بر استانداردهای جهانی و جلوگیری از تحمیل قوانین فراملی است که ممکن است با منافع ملّی یا فرهنگی کشور در تضاد باشند.

۷-۵-۶. ارتقای سواد حکمرانی دیجیتال در سطوح مدیریتی

حکمرانی هوشمند نیازمند مدیرانی است که تفاوت میان «اتوماسیون ساده» و «هوش مصنوعی خودمختار» را درک کنند. برگزاری دوره‌های تخصصی «حکمرانی داده‌محور» برای مدیران ارشد و میانی، ریسک تصمیم‌گیری‌های غیرکارشناسی و نگاه صرفاً امنیتی به فناوری را کاهش داده و دیدگاهی «رفاه‌محور» و «توسعه‌گرا» را جایگزین می‌کند.

۷-۶. جمع‌بندی

تحلیل راهکارها نشان می‌دهد که چگونه می‌توان از مبانی تئوریک به اقدامات عملیاتی در ایران رسید. می‌توان ابعاد اجرایی را در سه لایه مدل پیشنهادی (راهبردی، عملیاتی و نظارتی) بازخوانی و تحلیل کرد؛ شکل زیر و تبیین ذیل آن، گویای این مهم است:



شکل ۲: ابعاد اجرایی مدل پیشنهادی تحقیق

لایه اوّل: انطباق با لایه راهبردی: «هسته اخلاق، آگاهی و کرامت انسانی»

در لایه راهبردی، هدف همسوسازی فناوری با ارزش‌های بنیادین است. راهکار متناظر برای این هدف، حمایت از پژوهش‌های بین‌رشته‌ای و تدوین منشور اخلاق بومی است. برای این‌که «هسته اخلاق» در ایران فعال شود، باید از ترجمه کدهای اخلاقی غربی عبور کرد و به مبانی بومی اخلاق مراجعه نمود.

ازاین‌رو می‌بایست حمایت از هم‌افزایی سه ظرفیت دانشگاه‌ها، حوزه‌های علمیه و صنعت در دستور کار قرار گیرد و مفاهیمی مانند «عدالت» و «کرامت انسانی» از منظر فرهنگ ایرانی-اسلامی در کدهای هوش مصنوعی بازتعریف شوند. پیامد این اقدام، تبدیل هوش مصنوعی از یک پدیده وارداتی به یک فناوری بومی‌سازی شده است. دستاورد این اقدام، کاهش شکاف فرهنگی و افزایش پذیرش اجتماعی فناوری است.

انطباق با لایه عملیاتی: «تنظیم‌گری پویا و مشارکت‌جو»

لایه عملیاتی مسئول ایجاد تعادل میان سرعت تکنولوژی و ثبات قانون است. در این

لایه ضروری است «ایجاد سندباکس‌های رگولاتوری» و «تعیین و تثبیت نهاد حکمران هوش مصنوعی» مورد حمایت قرار گیرند. چراکه «محیط‌های آزمایشی قانونی» (سندباکس)، ابزار اجرایی این لایه است. سندباکس رگولاتوری محیطی است که در آن شرکت‌ها می‌توانند نوآوری‌های جدید را تحت نظارت یک رگولاتور آزمایش کنند. هدف ساده است: تسهیل نوآوری به شیوه‌ای ایمن و قابل‌اعتماد. نهاد حکمران هوش مصنوعی به‌عنوان راهبر این لایه، به‌جای وضع قوانین صلب و همگانی، برای هر حوزه (مثلاً سلامت یا مالی) راهکار متناظر: قواعدی منعطف و مشارکت‌جویانه تدوین می‌کند. این رویکرد، پیامد: خطر «اختلال در نوآوری» را که قبلاً به‌عنوان یک چالش مطرح شد، خنثی می‌کند. این اقدام، موجب دستاورد: حفظ سرعت رشد استارتاپ‌های ایرانی همگام با استانداردهای نظارتی می‌شود.

انطباق با لایه نظارتی: «شفافیت و مسئولیت‌پذیری الگوریتمیک»

هدف: صیانت از حقوق شهروندی در برابر خطاهای سیستمی از مهم‌ترین وظایف نهاد حکمران هوش مصنوعی است. ازاین‌رو لازم است اجباری کردن ممیزی الگوریتمیک و قابلیت توضیح مورد توجه قرار گیرد. برای تحقق شفافیت در لایه نظارتی، راهکار عملیاتی متناظر این پژوهش، «ممیزی مستقل» است. پیامد: نهاد حکمران باید قدرت داشته باشد تا الگوریتم‌های حسّاس (مانند سیستم‌های رتبه‌بندی اعتباری در بانک‌های ایران) را بازرسی کند. این لایه تضمین می‌کند که اگر تصمیمی علیه یک شهروند اتخاذ شد، مسئولیت قانونی آن مشخص باشد. این راهکار مستقیماً چالش «جعبه سیاه» را هدف قرار می‌دهد. دستاورد: صیانت از حقوق شهروندی، هوش مصنوعی را به یک سامانه پاسخ‌گو و قابل اعتماد برای عموم مردم تبدیل می‌کند.

۸. نتیجه‌گیری

حکمرانی کارآمد هوش مصنوعی نیازمند ایجاد توازن دقیق میان نوآوری و ایمنی است. پژوهش حاضر نشان داد که حکمرانی هوش مصنوعی، فراتر از یک ضرورت بوروکراتیک، زیربنای تحقق «عدالت دیجیتال» و «اقتدار ملّی» در سده پیش رو است. واکاوی ابعاد مختلف این پدیده نشان می‌دهد که موفقیت در این عرصه مستلزم عبور از نگاه تک‌بُعدی فنی و حرکت به سمت یک نظام حکمرانی چندلایه، پویا و ارزش‌محور است. یافته‌های تحقیق تأکید می‌کنند که مزیّت‌های راهبردی نظیر ارتقای کارآمدی

دولت، پیش‌بینی‌پذیری اقتصادی و صیانت از حقوق شهروندی، تنها در سایه نظام تنظیم‌گری حاصل می‌شوند که بتواند تناقض میان «سرعت نوآوری» و «ثبات قانون» را مدیریت کند. درواقع، حکمرانی کارآمد با تبدیل تهدیدهای بالقوه به فرصت‌های اقتصادی، نه‌تنها مانع رشد نیست، بلکه با ایجاد «اطمینان حقوقی»، ریسک سرمایه‌گذاری را کاهش داده و شکوفایی اکوسیستم‌های دانش‌بنیان را تضمین می‌کند.

بااین‌حال، تحقق این مزیّت‌ها با چالش‌های بنیادینی روبروست؛ شکاف عمیق میان سرعت پیشرفت فناوری و شتاب تنظیم‌گری، پیچیدگی فنی مدل‌های «جعبه سیاه» و فقدان چهارچوب‌های مشخص برای مسئولیت‌پذیری حقوقی، لزوم بازنگری در مدل‌های سنتی نظارت را دوچندان کرده است. در مورد خاص ایران، این چالش‌ها با تعدد نهادهای موازی و فقدان نهاد تنظیم‌گر مستقل گره‌خورده است که منجر به‌نوعی بی‌ثباتی در سیاست‌گذاری ملّی شده است. غلبه نگاه توسعه‌گرایی فنیِ صِرف، بدون توجه به پیامدهای اجتماعی و اخلاقی، می‌تواند منجر به بحران مقبولیت و مقاومت اجتماعی در برابر فناوری شود.

در پاسخ به این وضعیت، این پژوهش با ارائه مدل پیشنهادی حکمرانی یکپارچه و ارزش‌مدار، راهکاری سه‌لایه را ترسیم کرد که در آن «هسته اخلاقی»، «تنظیم‌گری پویا» و «نظارت هوشمند» با یکدیگر هم‌سو شده‌اند. در این مدل، لایه راهبردی با تدوین منشور اخلاق بومی، ارزش‌های انسانی را در اولویت قرار می‌دهد؛ لایه عملیاتی با بهره‌گیری از «سندباکس‌های قانونی»، توازن میان نوآوری و قانون را برقرار می‌سازد؛ و لایه نظارتی با اجباری کردن ممیزی الگوریتمیک و قابلیت توضیح، شفافیت و پاسخ‌گویی را محقق می‌کند.

درنهایت، حکمرانی هوش مصنوعی در ایران باید به‌مثابهٔ یک «زیست‌بوم زنده و یادگیرنده» بازتعریف شود؛ نظامی که نه‌تنها از ریسک‌های احتمالی پیشگیری می‌کند، بلکه از داده‌های ملّی به‌عنوان ابزاری برای رفع تبعیض و ارتقای رفاه عمومی صیانت می‌نماید. پیشنهاد نهایی این پژوهش بر این اصل استوار است که برای دستیابی به تمدن دیجیتال بومی، باید «عقلانیت اخلاقی» را در کالبد «هوش مصنوعی» تلفیق کرد و روح ارزش‌های اسلامی-ایرانی در معماری سیستم‌های هوشمند و ایجاد توازن میان «آزادی نوآوری» و «مسئولیت‌پذیری اخلاقی» را در آن دمید تنها در این صورت است که فناوری به‌جای ابزار قدرت، به خادم کرامت انسانی و پیشران توسعهٔ پایدار ملّی تبدیل خواهد شد.

فهرست منابع

آقایاری، سعید (۱۴۰۳). ***حکمرانی و هوش مصنوعی: روایت علم‌سنجی از دو داستان درهم‌تنیده***. فصلنامه مطالعات راهبردی سیاست‌گذاری عمومی، ۱۴ (۵۰)، ۱۵۵-۱۷۸.

اکبری، محسن؛ جلائی‌نیا، راحله (۱۴۰۴). ***ارائه الگوی سیاست‌گذاری توسعه هوش مصنوعی در راستای برنامه هفتم توسعه***. مجله ایرانی سیاست عمومی، ۱۱(۲)، ۱۱۹-۱۳۶.

ترابی، محمدعلی؛ اقبال، حسین (۱۴۰۴)، ***پیشنهاد یک مدل حکمرانی هوش مصنوعی برای اداره دولتی در جمهوری اسلامی ایران***، مطالعات دولتی ایران معاصر ۹.

حسینی، محمدرضا؛ عزیزی مهماندوست، مهدی (۱۴۰۳). ***مطالعه تطبیقی قوانین و سیاست‌های هوش مصنوعی در کشورهای پیشرفته و ارائه پیشنهادهایی برای ایران***، فصلنامه حقوق فناوری‌های نوین، ۶ (۱۱).

دبیرخانه مجمع تشخیص مصلحت نظام (۱۳۹۲). ***سیاست‌های کلّی اقتصاد مقاومتی در مجموعه سیاست‌های کلّی نظام***.۱۴۰۲، ۱۲۷. دبیرخانه مجمع تشخیص مصلحت نظام.

دبیرخانه مجمع تشخیص مصلحت نظام (۱۴۰۲). ***سند چشم‌انداز جمهوری اسلامی ایران در افق ۱۴۰۴ در مجموعه سیاست‌های کلّی نظام***.۱۴۰۲، ۱۲۷. دبیرخانه مجمع تشخیص مصلحت نظام.

رئیسی، محمدابراهیم (۱۴۰۴). ***مقایسه سند ملی هوش مصنوعی ایران با سند ملی هوش مصنوعی سنگاپور و دلالت‌های آن بر امنیت اقتصادی***. فصلنامه امنیت اقتصادی، ۱۲ (۸)، ۴۳-۶۸.

شورای عالی انقلاب فرهنگی (۱۳۹۰). ***نقشه جامع علمی کشور*** (مصوّب جلسه ۶۷۹ مورخ ۱۳۸۹/۱۰/۱۴)

شورای عالی انقلاب فرهنگی (۱۴۰۳). ***سند ملّی هوش مصنوعی جمهوری اسلامی ایران*** (مصوب جلسه ۹۰۱ مورخ ۱۴۰۳/۳/۲۹)

شورای عالی فضای مجازی کشور (۱۴۰۱). ***سند راهبردی جمهوری اسلامی ایران در فضای مجازی*** (مصوب جلسه ۸۴ مورخ ۱۴۰۱/۵/۱۱)

صیادی، معصومه (۱۴۰۲، ۲۹ مرداد). ***حاکمیت هوش مصنوعی چیست و AI Governance چه اصولی دارد***. پایگاه خبری پیوست.

منصور، جهانگیر. ***کتاب قانون اساسی جمهوری اسلامی ایران***. (مصوب مجلس خبرگان قانون اساسی در سال ۱۳۵۸)، (۱۴۰۴) انتشارات دوران.

References

Agarwal, A., & Nene, M. J. (2025). *A five-layer framework for AI governance: Integrating regulation, standards, and certification*. Transforming Government: People, Process and Policy, 19(3), 535–555. https://doi.org/10.1108/TG-03-2025-0065

Aghayari, S. (2024). *Governance and artificial intelligence: A scientometric narrative of two intertwined stories*. Strategic Studies of Public Policy, 14(50), 155–178. (In Persian)

Akbari, M., & Jala'inia, R. (2025). *Presenting a policy-making model for artificial intelligence development in line with the Seventh Development Plan*. Iranian Journal of Public Policy, 11(2), 119–136. (In Persian)

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Atwood, B. (2025). *Artificial intelligence in Iran: National narratives and material realities*. Iranian Studies, 1–18. https://doi.org/10.1017/irn.2025.10

Batool, A., Zowghi, D., & Bano, M. (2025). *AI governance: A systematic literature review*. AI and Ethics, 5, 3265–3279. https://doi.org/10.1007/s43681-025-00456-8

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer.

Friedman, B. (1996). *Value-sensitive design*. Interactions, 3(6), 16–23. https://doi.org/10.1145/242485.242493

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

Friedman, B., & Kahn, P. H., Jr. (2003). *Human values, ethics, and design*. In J. A. Jacko & A. Sears (Eds.), The human-computer interaction handbook (pp. 1177–). [Publisher not provided].

Hosseini, M. R., & Azizi Mehmandoost, M. (2024). *A comparative study of artificial intelligence laws and policies in developed countries and providing suggestions for Iran*. Journal of Modern Technologies Law, 6(11), 351–375. (In Persian)

Kaminski, M. E. (2020). *The right to explanation, explained*. Berkeley Technology Law Journal, 34(1), 189–218.

Mansour, J. (2025). *The Constitution of the Islamic Republic of Iran* (Approved by the Assembly of Experts for the Constitution in 1979). Douran Publications. (In Persian)

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate*. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Prasad, S. B., Verma, A., & Singh, R. (2025). *Ethical frontiers and legal boundaries: Proposing a unified framework for AI regulation and accountability*. Next Research, 2(4), 101087. https://doi.org/10.1016/j.nexres.2025.101087

Raeisi, M. E. (2025). *A comparison of Iran's national artificial intelligence document with Singapore's national artificial intelligence document and its implications for economic security*. Economic Security Quarterly, 12(8), 43–68. (In Persian)

Sayyadi, M. (2023, August 20). *What is AI governance and what principles does it have?* Peivast News Agency. (In Persian)

Secretariat of the Expediency Discernment Council. (2013). *General policies of resistance economy*. In Collection of general policies of the system (p. 127). Secretariat of the Expediency Discernment Council. (In Persian)

Secretariat of the Expediency Discernment Council. (2023). *The vision document of the Islamic Republic of Iran in the horizon of 2025* (1404). In Collection of general policies of the system (p. 127). Secretariat of the Expediency Discernment Council. (In Persian)

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Supreme Council of Cyberspace of the Country. (2022). *Strategic document of the Islamic Republic of Iran in cyberspace* (Approved at Session 84, dated 11/5/1401 [August 2, 2022]). (In Persian)

Supreme Council of the Cultural Revolution. (2011). *Comprehensive scientific roadmap of the country* (Approved at Session 679, dated 14/10/1389 [January 4, 2011]). (In Persian)

Supreme Council of the Cultural Revolution. (2024). *National artificial intelligence document of the Islamic Republic of Iran* (Approved at Session 901, dated 29/3/1403 [June 19, 2024]). (In Persian)

Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). *Global governance of artificial intelligence: Next steps for empirical and normative research*. International Studies Review, 25(3). https://doi.org/10.1093/isr/viad030

Torabi, M. A., & Eghbal, H. (2025). *Proposing an artificial intelligence governance model for public administration in the Islamic Republic of Iran*. Contemporary Iranian State Studies, 9, 11–42. (In Persian)

Trik, M., Mozaffari, S., & Bidgoly, A. J. (2025). *AI in public decision-making: A philosophical and practical framework for assessing and weighing harm and benefit*. In Proceedings of the 2025

ACM Conference on Fairness, Accountability, and Transparency (FAccT '25) (pp. 1123–1138). https://doi.org/10.1145/3715275.3715323

Wittmann, V., & Meynhardt, T. (2025). *Human-centric AI governance: What the EU public values, what it really, really values*. Government Information Quarterly, 42(4), 102007. https://doi.org/10.1016/j.giq.2025.102007