PHIG PARTNERS

Public Health Infrastructure Grant

Literature Review: Examining the Key Considerations for the Use of Artificial Intelligence in Public Health and Health Care

Supplement to *Developing Artificial Intelligence Policies for Public Health Organizations: A Template and Guidance*

Health and Human Services Region 7

January 2025

PHIG PARTNERS
Public Health Infrastructure Grant

About This Initiative

Developing Artificial Intelligence Policies for Public Health Organizations: A Template and Guidance is a collaboration between the Kansas Health Institute (KHI), Health Resources in Action (HRiA), and the Wichita State University Community Engagement Institute (WSU CEI). The project scope included an environmental scan that informed the template's development, comprising a literature review and policy analysis at both state and city levels. The role of each organization varied by project component. This document, *Literature Review: Examining the Key Consideration for the Use of Artificial Intelligence in Public Health and Health Care*, was developed by all three organizations.

Authors

- **Emma Uridge, M.P.H.**, Analyst, KHI
- **Shelby C. Rowell, M.P.A.**, Analyst, KHI
- **Erika Gaitan, M.S.W.**, Associate Director of Community Impact, HRiA
- **Tatiana Y. Lin, M.A.**, Director of Business Strategy and Innovation, KHI
- **Taylor Carter, M.P.H.**, Public Health Program Specialist, WSU CEI
- **Emily Brinkman, M.P.H., R.D., L.D.**, Public Health Program Specialist, WSU CEI

Contributors

- **Ben Wood, M.P.H.**, Senior Director of Policy and Practice, HRiA
- **Kate Holmes, M.P.H.**, Associate Director of Health and Racial Equity, HRiA
- **AAron Davis, M.P.A.**, Director, Center for Public Health Initiatives, WSU CEI
- **Linda J. Sheppard, J.D.**, Team Leader, KHI

Acknowledgement

This work is supported by funds made available from the Centers for Disease Control and Prevention (CDC) of the U.S. Department of Health and Human Services (HHS), National Center for STLT Public Health Infrastructure and Workforce, through OE22-2203: Strengthening U.S. Public Health Infrastructure, Workforce, and Data Systems grant. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, CDC/HHS or the U.S. Government.

Disclosure Statement

During the literature review process, the research team utilized AI tools, specifically Petal and ChatGPT, to identify search terms, support the development of articles summaries and cross-articles analyses. These tools were used to generate initial drafts of the summaries. All content was subsequently reviewed and refined by the authors to ensure accuracy and quality. The authors take full responsibility for the final content presented in this document.

PHIG PARTNERS
Public Health Infrastructure Grant

Table of Contents

(This page intentionally left blank.)

Introduction

In 2024, the Kansas Health Institute (KHI), Health Resources in Action (HRiA) and Wichita State University Community Engagement Institute (WSU CEI) collaborated on a project which resulted in the document titled *Developing Artificial Intelligence Policies for Public Health Organizations: A Template and Guidance*. This template is designed to assist public health organizations, including nonprofits and government agencies at all levels, in creating policies or guidelines that facilitate ethical experimentation with artificial intelligence (AI) systems while addressing potential risks and promoting health equity and innovation.

To inform the development of the template, the research team conducted an environmental scan focused on considerations surrounding the use of AI, specifically identifying what should be included in the policies for public health organizations. This scan included a review of relevant literature, state-level policies that were introduced or passed, and policies or guidelines passed by local-level jurisdictions, specifically cities.

The primary goal of this document, *Literature Review: Examining the Key Consideration for the Use of Artificial Intelligence in Public Health and Health Care,* is to summarize findings from the literature regarding opportunities and challenges associated with AI. Key topics include bias, transparency, data privacy and the effects on diverse populations — issues frequently emphasized by the public health community and reflected in media and research publications.

The review's primary objective was to establish a robust foundation for developing AI policy and guidance templates tailored to public health needs. It addressed 10 research questions and examined 96 articles, with most studies focusing on the U.S. context, supplemented by insights from international settings. Both peer-reviewed studies and grey literature were included in the review.

During the initial phase of template development, the list of included provisions was based on findings from the review of policies. To ensure that the provisions in the template aligned with current research, the literature review findings were used to verify, modify, remove or propose new provisions. Additionally, the literature review findings were used to develop sections of the template that explain why specific issues are important to include in the policy and provide rationale for their inclusion.

Executive Summary

The literature review explored diverse dimensions of AI implementation in public health, encompassing its potential benefits and challenges. Key areas of focus included bias in AI systems, data privacy concerns, transparency issues, regulatory and ethical oversight and the impacts on historically marginalized populations. The findings offer critical insights into how public health policies can be shaped to maximize AI's benefits while addressing its potential risks. Through a balanced, human-centric approach that incorporates community engagement, these policies can guide the responsible and equitable deployment of AI in public health settings.

Key Findings Across Research

AI has been applied across various industries, including public health. The Centers for Disease Control and Prevention (CDC), for example, uses MedCoder, a system that utilizes natural language processing and machine learning to code causes of death, automating nearly 90 percent of records compared to the previous rate of less than 75 percent. Similarly, the Chicago Department of Public Health has used AI to identify children at high risk of lead poisoning, prioritizing home inspections through historical data analysis. To support real-time surveillance, AI can be used to analyze social media to monitor public sentiment, detect emerging health trends and identify potential disease outbreaks. By tracking keywords and discussions, AI could enable more rapid and effective responses to emerging health issues. However, to optimize the benefits of AI, it's essential to ensure models are unbiased and address data privacy concerns. Understanding potential sources of bias in AI algorithms and developing strategies to address them can help reduce the risk of reinforcing existing inequities.

Beyond this, AI holds potential for a range of applications. AI can be leveraged to assist in generating code or checking it for errors. It also can support community health needs assessments by performing tasks ranging from creating survey questions to summarizing results. Generative AI tools like ChatGPT can further enhance the public health workforce by aiding in administrative processes, creating health communications, drafting press releases and generating educational materials and training resources. In the use of AI tools, it is essential to maintain a human-centric approach to ensure ethical oversight, accuracy and the responsible application of AI in public health. Additionally, it is crucial to recognize the environmental impact of AI technologies, as well as address copyright concerns related to AI-generated content.

The literature review highlights several key considerations for using AI in public health, each with its own set of challenges and recommendations. Below is a focused summary based on the sections related to bias, transparency and data privacy.

Bias: AI's impact on health equity raises important considerations. The literature indicates that while AI has the potential to enhance public health, improve health services and reduce disparities, it can also perpetuate or amplify existing biases if not designed and implemented thoughtfully. If biases are embedded in training data, they can lead to inequitable public health interventions, misallocation of resources and the perpetuation of health disparities among underserved communities. To mitigate this, public health policies should incorporate inclusive development practices, use diverse datasets and establish continuous bias audits. The review emphasizes that addressing bias involves not only technical solutions but also the engagement of diverse stakeholders to ensure AI systems serve the entire community equitably.

Transparency: Transparency is crucial for public trust in AI applications in public health. However, the "black-box" nature of many AI systems creates significant challenges in ensuring explainability and accountability. The literature stresses that without clear and transparent decision-making processes, public confidence in AI-driven public health initiatives may diminish. To address this, public health policies should require AI systems to include explainable AI practices that make the decision-making processes understandable to non-expert users and stakeholders. Such transparency helps facilitate better oversight, enhances public trust and ensures that AI complements human expertise.

Data Privacy: Data privacy is a major focus in the use of AI for public health. AI systems are often trained on large volumes of data, which may involve sensitive information, thereby raising significant privacy and security challenges. The literature points out that breaches or misuse of data can severely undermine public trust and the effectiveness of public health programs. Public health policies should implement strong data governance frameworks, including secure data handling practices, informed consent protocols and privacy-preserving technologies.

Oversight: The literature highlights the crucial role of human oversight in automated decision-making, emphasizing its importance for accountability, safety and equity. Key oversight mechanisms, such as human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC), are designed to ensure ongoing human oversight, and intervention at critical stages of AI operations and decision making, particularly in sectors like health care and criminal

justice. However, challenges such as automation bias, superficial oversight practices and vague policy guidelines hinder effective implementation. According to the literature, ensuring transparency, addressing systemic biases and providing robust, actionable frameworks for oversight are vital to maintaining trust and ethical standards in AI systems.

Equity and Ethical Considerations: The literature emphasizes the critical need to address equity and ethical considerations in AI policies. Key themes include the importance of mitigating health disparities, preventing algorithmic bias and promoting fairness and transparency in AI systems. Ethical frameworks should prioritize inclusivity, accountability and human oversight to ensure AI's responsible use, particularly in sensitive sectors like health care and public health. Challenges such as the digital divide, lack of transparency and regulatory gaps underscore the need for comprehensive policies that close equity gaps and prevent systemic biases. Recommendations focus on embedding bias mitigation strategies, fostering community engagement, ensuring transparency and developing interdisciplinary collaborations to promote equitable and ethical AI deployment.

Individuals with Disabilities: The literature highlights both the opportunities and challenges that AI presents for individuals with disabilities, emphasizing the importance of inclusive design and equitable frameworks. On one hand, AI offers transformative potential: In health care, it can enhance diagnostics and personalized care; in education, it can provide tailored learning experiences; and in employment, it can support workplace accommodations and improve accessibility. However, concerns remain as AI systems could reflect biases stemming from the underrepresentation of disabled individuals in training datasets, potentially leading to unfair outcomes such as misdiagnoses, hiring exclusion and inequitable educational assessments. Ethical frameworks focused on fairness are critiqued for failing to address systemic inequalities, prompting calls for a justice-oriented approach that empowers individual with disabilities and acknowledges structural barriers. Inclusive design practices, such as engaging individuals with disabilities in AI development, are vital for addressing these challenges.

Older Adults: The literature underscores the significant potential of AI to improve the lives of older adults, particularly through technologies that enhance care, social connectivity and well-being. AI tools, such as telehealth services and socially assistive robots, offer benefits like personalized care, reduced loneliness and better engagement in daily activities. However, challenges such as algorithmic bias, mistrust, and digital ageism hinder widespread adoption. Digital ageism, in particular, encompasses biases against individuals based on their age in

digital contexts. This includes stereotypical views of older adults as technologically inept, assumptions about their inability to learn new technologies, and their exclusion from digital innovations due to perceived lack of adaptability. Older adults often face barriers like limited digital literacy and inequitable access to technology, especially in rural areas. The recommendations referenced in the literature include fostering inclusive design by involving older adults in AI development, improving digital literacy through community-based programs and addressing biases through audits and equitable data practices to ensure AI solutions cater to their diverse needs.

Racial and Ethnic Minorities: The literature highlights the dual impact of AI on racial and ethnic minorities, noting its potential to either perpetuate systemic inequities or promote fairness when carefully designed. Bias in AI systems is pervasive, stemming from non-representative training data, flawed model design and historical inequalities embedded in datasets. These biases lead to inequitable outcomes in health care, employment and criminal justice, such as misdiagnoses, hiring discrimination and harsher sentencing for minority groups. Challenges include regulatory gaps, lack of diversity in AI development teams and limited access to digital infrastructure. However, AI also offers opportunities to reduce disparities when paired with inclusive design, diverse datasets and equity-focused policies. The literature emphasizes the need for robust governance frameworks, bias mitigation strategies and community engagement to ensure AI systems are fair, accountable and beneficial for marginalized communities.

Community Engagement: The literature highlights that effective community engagement in AI policy is essential for equitable and culturally sensitive AI systems. Integrating community input during policy development enhances trust, ensures policies align with local needs and helps mitigate biases in AI applications. AI has the potential to improve public services such as health care and urban planning, but its benefits are contingent on equitable access and inclusive design. Challenges include infrastructure limitations, public mistrust and the risk of algorithmic bias disproportionately affecting marginalized groups. Recommendations include investing in digital infrastructure, fostering public-private partnerships, conducting inclusive consultations and promoting ethical AI practices with robust oversight and transparency. AI literacy and skill development also are crucial to empower communities to meaningfully engage with AI technologies.

Methodology

The literature review was conducted collaboratively by all three organizations. The purpose of the literature review was to examine considerations related to the utilization of artificial intelligence (AI) for various functions in public health. This review specifically focused on areas such as bias, transparency, data privacy and the impacts on different populations, among other relevant topics. The issues of focus were determined based on those typically raised by the public health community and cited in both media and research.

The primary purpose of this review was to establish a foundational basis for developing AI policy and guidance templates for public health organizations. The literature review encompassed both peer-reviewed literature and grey literature. A total of 10 research questions were explored.

Examined Questions:

1. What does the current landscape for the adoption and use of AI in public health look like?
2. How is bias mitigation addressed in AI policies?
3. How is data privacy addressed in AI policies?
4. How is transparency addressed in AI policies?
5. What role do AI policies assign to human oversight and intervention in automated decision-making processes?
6. What are the equity and ethical considerations of AI that should be addressed in policies?
7. What are the impacts of AI on individuals with disabilities and how should these issues be addressed in AI policies?
8. What are the impacts of AI on older adults and how should these issues be addressed in AI policies?
9. What are the impacts of AI on racial and ethnic minorities and how should these issues be addressed in AI policies?
10. What are the best practices for addressing community engagement in AI policies?

Most of the articles included in the review were published within the last five years. The research team focused on this timeframe due to the rapidly evolving landscape of AI and related research. However, a few articles from earlier periods also were included. All articles were published in English and mainly centered on the context of the United States, although some

articles focused on settings outside of the U.S. The literature sources included databases such as Google Scholar, PubMed, academic libraries and health and technology focused journals. Although the goal of the literature review was focused on public health, many articles centered on health care settings were included within the scope. The keywords employed for literature searches varied depending on the topic and included phrases such as "artificial intelligence and bias" and "generative AI and bias." These keywords were refined following an initial review of the articles to enhance the relevance and scope of the search process.

The number of articles reviewed per research question varied between 3 and 47. Due to the specificity of the research questions and the fact that articles often addressed more than one issue, certain articles were used to address multiple research questions. For example, an article exploring how bias mitigation is addressed in AI policies also might be relevant for a research question on transparency. The total number of unique articles included in the literature review was 96.

The articles were grouped by topic and detailed in a table referred to as the evidence table, which captured information such as the author, year, source, key findings relevant to the topic, recommendations and other pertinent details. An example of the evidence table can be found in *Appendix B*, page B-2

The research team utilized two AI tools, Petal and ChatGPT 4.0, for the literature review. The purpose was to evaluate the feasibility of using these systems for literature reviews and to facilitate the examination of a larger number of research questions and articles during the three-month project period. The tools were specifically used for several purposes: To suggest initial search terms, identify articles based on set search parameters, summarize articles for inclusion in the evidence table based on the inclusion criteria and support the creation of final summaries across all articles by topic.

The research team prioritized a human-in-the-loop approach in the quality assurance process, which was implemented throughout the review. Outputs were examined and validated against the available articles to ensure accuracy and reliability. For more information about the process of using these two AI tools for the literature review, see *Appendix B*, page B-1.

Section Structure Overview

In general, the literature review for each research question is structured into the following sections: Research questions examined in articles, summary of key findings, challenges, recommendations and bibliography. However, some research questions included additional sections based on their scope.

Limitations

The literature review has several limitations that should be acknowledged. It was not conducted systematically, which may have led to the omission of relevant articles related to the research questions, potentially limiting the comprehensiveness of the findings. Non-systematic reviews also can introduce selection bias, as the process of article selection may not be as rigorous or standardized as in systematic reviews. Furthermore, the use of generative AI tools, such as ChatGPT and Petal, presents certain limitations. The outputs generated by these tools were based on the documents and data provided to them, which might not capture all nuances or context-specific details. While a quality assurance (QA) process was implemented to verify the results, there remains the potential for nuanced information to be overlooked or misinterpreted during the generation process. Additionally, AI-generated summaries may lack the depth that human researchers bring, especially in critically analyzing and synthesizing complex research findings.

Research Question 1: What does the current landscape for the adoption and use of AI in public health look like?

Research Questions Examined in Articles

The reviewed literature examines how AI technologies are adopted and deployed to enhance public health outcomes. Some studies focus on specific use cases, such as improving health outcomes, public health surveillance and response. Others provide reviews of AI techniques, illustrating how AI contributes to disease outbreak prediction, patient diagnosis, treatment and resource optimization.

Ethical and Legal Considerations of AI

Multiple articles address ethical and legal challenges in AI implementation, emphasizing the need for responsible and transparent decision-making. The literature recommends the

establishment of regulatory frameworks, oversight mechanisms and ethical guidelines to govern AI development and mitigate bias. These recommendations aim to ensure that AI systems operate within a clear legal framework while upholding ethical principles and public trust.

Health Equity and AI

The impact of AI on health equity is a central theme across the literature, with research exploring both the potential for AI to reduce or exacerbate social inequalities. Two studies highlight the importance of fairness, inclusivity and transparency in AI design, especially in public health. The findings suggest that AI systems must incorporate rights-based approaches to prevent deepening existing disparities and to promote equitable health care outcomes.

AI in Public Health Communication

One study examines AI's role in generating effective public health communication, particularly using models like GPT-3 for pro-vaccination messaging. The study assesses AI's potential to shape public attitudes, emphasizing the need for transparency and ethical considerations to maintain public trust.

Summary of Key Findings – Adoption and Use of AI In Public Health

Artificial Intelligence (AI) is rapidly transforming public health, offering innovative solutions to complex challenges ranging from disease surveillance to personalized medicine. As AI technologies become more integrated into public health practices, they promise significant improvements in efficiency, accuracy and overall health care outcomes. However, these advancements are accompanied by a set of ethical, regulatory and practical challenges that must be carefully navigated to ensure equitable and responsible use.

Disease Outbreak Prediction and Surveillance

AI-powered systems are widely recognized for enhancing disease outbreak prediction and surveillance capabilities by analyzing large, complex datasets, such as social media, electronic health records and public health reports, to detect early signs of outbreaks and predict future occurrences (Kauffman, 2022). AI can enable earlier identification of outbreaks and trends than traditional methods by incorporating diverse data sources and managing temporal and spatial complexities (Zeng et al., 2021). The use of AI also facilitates global monitoring through rapid internet-based surveillance (Zeng et al., 2021).

Patient Diagnosis and Treatment

AI significantly improves diagnostic accuracy by analyzing medical images, thereby reducing invasive procedures and supporting decision-making in treatment optimization (Kauffman, 2022). Machine learning algorithms excel in image recognition, significantly enhancing the ability to diagnose diseases, for example, acromegaly — a rare hormonal disorder caused by the pituitary gland producing excessive amounts of growth hormone (GH) during adulthood. Early detection is critical to managing this condition effectively (Thomasian et al., 2021).

Health Equity and Bias Correction

AI has demonstrated the ability to promote health equity by improving access to care and resource allocation in underrepresented communities (Thomasian et al., 2021). By auditing and adjusting data labels, AI systems can help correct racial and social biases, advancing fairness in health care delivery (Thomasian et al., 2021). Continuous bias surveillance ensures the accuracy and fairness of AI models, improving their impact on public health interventions.

Clinical Trials and Research

AI streamlines the clinical trial process by predicting outcomes and identifying patients more likely to respond to treatments, thereby reducing time and costs (Kauffman, 2022). Additionally, AI-generated public health reports and the automation of health data summarization contribute to streamlining research efforts (Jungwirth & Haluza, 2023).

Personalized Medicine and Precision Care

These technologies support rehabilitation and care robotics and assist in interventions and the communication needs of disabled individuals (Giansanti, 2022). Wearable technologies further enhance personalized care by continuously monitoring individual medical information (Giansanti, 2022).

Public Health Monitoring and Decision-Making

AI plays a critical role in monitoring the effectiveness of public health programs by analyzing data to identify areas needing improvement and by providing feedback on the efficacy of proposed public health policies (Kauffman, 2022; Jungwirth & Haluza, 2023). It also enhances decision-making processes in public health action, allowing for better response planning and resource allocation (Davis et al., 2024). These technologies contribute to enhancing access to

care, particularly in underserved regions (Kauffman, 2022) and can significantly reduce costs associated with public health programs by targeting efforts more efficiently (Kauffman, 2022).

AI in Public Health Messaging

AI, particularly large language models like GPT-3, has shown potential in rapidly generating public health content and supporting the development of public health messaging (Karinshak et al., 2023). These models can contribute to effective and tailored communication in public health campaigns, particularly around vaccination and preventive measures.

Simulation of Public Health Policies

AI's ability to simulate the effects of public health interventions allows policymakers to evaluate proposed policies before implementation, providing valuable feedback on potential outcomes (Jungwirth & Haluza, 2023). This capability helps improve the effectiveness of interventions and contributes to better planning and execution of public health strategies.

Challenges

The reviewed literature identifies several challenges for incorporating AI into the public health and health care discipline.

Data Privacy

In the reviewed literature on the integration of artificial intelligence (AI) in public health, several challenges are consistently identified across different sources. A significant concern is data privacy, as the vast amounts of sensitive health data required for AI systems pose risks for breaches and misuse. To protect sensitive information, existing laws like the federal Health Insurance Portability and Accountability Act (HIPAA) may not fully address the new risks introduced by AI.

Bias and Discrimination

Bias and discrimination in AI models can perpetuate or exacerbate existing inequalities in health care. AI systems are at risk of reinforcing racial and socioeconomic biases due to training on unbalanced or biased datasets (Thomasian et al., 2021). AI algorithms often replicate biases inherent in training data, which can result in discriminatory outcomes (Zeng et al., 2021). In public health surveillance, there is potential for AI to perpetuate bias if not properly designed and monitored, making equitable access to technology and model interpretability critical (Jungwirth & Haluza, 2023).

Model Accuracy and Ethical Considerations

Accurate and reliable results from AI models are essential, especially in health contexts where errors could lead to serious consequences (Kauffman, 2022). The "black-box" nature of AI poses ethical concerns by hindering transparency and accountability, making it difficult for users to understand or challenge AI-driven decisions (Thomasian et al., 2021). Additionally, AI-generated misinformation, such as the production of outdated or incorrect information, threatens trust in these technologies (Karinshak et al., 2023).

Regulation

Ethical guidelines highlight the importance of external validation of AI models, transparency and adherence to existing regulatory standards to ensure the safe and equitable application of AI in public health (Davis et al., 2024).

Technical Requirements and Operability

The technical and operational challenges of AI in public health are multifaceted. These include difficulties related to data sparsity, the development of baselines and the creation of effective computational frameworks (Zeng et al., 2021). Ensuring data quality and system scalability remains a significant challenge, particularly in large-scale implementations (Kauffman, 2022). Additional technical hurdles include developing effective prompts and conducting iterative testing to refine AI systems (Karinshak et al., 2023).

Societal and Infrastructure Barriers

Societal and infrastructure barriers also present key obstacles to AI adoption in public health. Limited broadband access in regions with underdeveloped digital infrastructure impedes effective AI deployment (Davis et al., 2024).

In conclusion, while AI holds great potential for advancing public health, substantial challenges persist in areas such as data privacy, bias, model accuracy, regulatory gaps, technical limitations and societal acceptance. Addressing these issues is critical for the ethical and equitable deployment of AI.

Recommendations

Transparency and Human Oversight in AI Applications

Strict human oversight is essential to complement AI systems in public health, ensuring they serve as augmentative tools rather than replacements for human judgment (Jungwirth & Haluza,

2023). Maintaining human oversight in AI-generated public health messaging is critical to ensuring accuracy and relevance (Karinshak et al., 2023). Additionally, iterative review processes and the development of effective prompts are recommended to improve the quality of AI contributions in public health communication (Karinshak et al., 2023). Transparency is also a key factor in building public trust. Clear communication about the use of AI-generated messages and assurances that AI supports, rather than replaces, human expertise is necessary (Karinshak et al., 2023).

Bias Mitigation and Health Equity

Mitigating algorithmic bias is an essential component of ethical AI implementation. Comprehensive frameworks to address bias throughout the AI lifecycle, from data collection to implementation, are necessary (Thomasian et al., 2021). Strategies such as using diverse datasets and federated learning techniques can enhance data diversity while preserving privacy (Thomasian et al., 2021). Continuous bias audits are also recommended to identify and mitigate intersectional biases related to factors such as gender, age and socioeconomic status (Thomasian et al., 2021).

AI in Public Health Surveillance and Strategic Implementation

AI-enabled public health surveillance shows potential for detecting local outbreaks and monitoring global epidemics, though applications remain in the early stages (Zeng et al., 2021). Further research is needed to resolve technical and ethical challenges, such as ensuring privacy and improving model interpretability (Zeng et al., 2021).

A strategic approach is essential for implementing generative AI in public health. Starting with a small number of highly visible, impactful, and strategically prioritized applications of generative AI use cases that align with organizational priorities and offer measurable outcomes is advised (Davis et al., 2024). Organizations must assess their technological and talent capacities before scaling AI projects, with collaborations between public health agencies and private or academic partners playing a critical role in development, testing and scaling (Davis et al., 2024). Effective risk management practices focusing on fairness, privacy and regulatory compliance are necessary, along with adherence to evolving global guidelines such as the EU AI Act and U.S. Department of Health and Human Services regulations (Davis et al., 2024).

Bibliography

1. Davis, S., Singh, S., Srinidhi, N., & Wilson, M. (2024). Public health's inflection point with generative AI. McKinsey & Company. https://www.mckinsey.com/industries/social-sector/our-insights/public-healths-inflection-point-with-generative-ai#/
2. Giansanti, D. (2022). Artificial intelligence in public health: Current trends and future possibilities. *International Journal of Environmental Research and Public Health*, *19*(11907), 1−4. https://doi.org/10.3390/ijerph191911907
3. Jungwirth, D., & Haluza, D. (2023). Artificial intelligence and public health: An exploratory study. *International Journal of Environmental Research and Public Health*, 20(4541). https://doi.org/10.3390/ijerph20054541
4. Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Article 116. https://doi.org/10.1145/3579592
5. Kauffman, H. (2022). Artificial Intelligence and Public Health: A Descriptive Review of Use Cases. *Applied Research in Artificial Intelligence and Cloud Computing,* 5(1), 29–37. https://researchberg.com/index.php/araic/article/view/88
6. Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy*, 42(4), 602-611. https://doi.org/10.1057/s41271-021-00319-5
7. Zeng, D., Cao, Z., & Neill, D. B. (2020). Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. *Artificial Intelligence in Medicine*, 437–453. https://doi.org/10.1016/B978-0-12-821259-2.00022-3

Research Question 2: How is bias mitigation addressed in AI policies?

Research Questions Examined in Articles

The literature identified and examined how AI researchers and policymakers address bias mitigation in AI policies, revealing critical disconnects that complicate regulatory efforts and ethical considerations. Literature frequently recommends the establishment of ethical guidelines, regulatory frameworks and oversight mechanisms to govern AI development and deployment. Addressing bias in AI systems is a central research question across the literature identified.

Bias, Equity and Inclusivity

One exploration is how AI can exacerbate or reduce social inequalities. Literature stressed the importance of designing AI applications with fairness, inclusivity and transparency around mitigating bias. This is particularly relevant in public health, where AI has the potential to either mitigate or deepen health inequities based on inherent bias. Discussions also extend to AI's impact on populations and communities of focus, underrepresented groups and global health disparities.

Common Policy Recommendations and Governance Approaches

The literature frequently recommends establishing oversight mechanisms to govern AI development and deployment. This includes implementing policies that ensure transparency, explainability to lay audiences and accountability, particularly in efforts to mitigate bias in AI systems.

Summary of Key Findings – Bias Mitigation

Mitigating bias in AI systems, as well as reviewing AI-generated outputs for bias, is a central concern for researchers exploring the implications of AI in public health and other sectors.

Policy and Governance

The lack of alignment between different conceptions of AI poses a risk to the field, particularly concerning the real-world implications of algorithmic bias (Krafft et al., 2020). There is a need for algorithms, the predetermined way AI systems operate, to reflect human values, such as fairness and accountability, while managing the balance between "applied inequality" and "consequential decision-making" (Calo, 2017). Public concerns over fairness and transparency in AI applications, especially in criminal justice and hiring practices, underscore the absence of robust regulations to safeguard personal data and address bias (Robles & Mallinson, 2023). Existing AI policies across countries often exhibit significant differences, though they frequently emphasize ethical principles like justice and fairness, as seen in private sector initiatives and guidelines to address algorithmic bias, such as IEEE P7003 (a draft standard that provides processes and methodologies to address issues of bias in algorithms) (Biersmith & Laplante, 2022). Multi-stakeholder and diverse participation and the use of both technical and non-technical measures are necessary for effectively addressing bias in real-world AI applications (Stix, 2021). The literature calls for proactive measures to tackle inherent algorithmic bias, social

inequalities and discriminatory impacts in AI development and deployment (Fukuda-Parr & Gibbons, 2021).

Health Care and Public Health

In health care and public health, there is a substantial focus on addressing the biases that can be perpetuated by AI that can impact decision making and produce misinformation (Davis et al., 2024). AI technologies should not encode biases to the disadvantage of identifiable groups, especially groups that are already marginalized. Bias is a threat to inclusiveness and equity, as it can result in a departure, often arbitrary, from equal treatment (WHO, 2021). AI-driven health interventions must be guided by local needs and input, rather than data availability, to avoid perpetuating biases in datasets related to ethnicity, socioeconomic status and gender (Schwalbe & Wahl, 2020). Biased algorithms in health care can exacerbate health disparities by misjudging risks in certain patient populations, necessitating transparent decision-making in AI applications (Reddy et al., 2020). The risk of systemic inequities being embedded in AI algorithms calls for the use of multidisciplinary approaches to design ethical AI systems that adhere to current health care standards (Murphy, Di Ruggiero, & Upshur, 2021), while also recognizing these biases. Transparent decision-making procedures and community engagement with underrepresented groups are crucial to mitigate health inequities in AI applications (Smith et al., 2020). AI regulation in public health should include adherence to findable, accessible, interoperable and reusable (FAIR) data principles to ensure ethical data collection and prevent unnoticed biases (Verma et al., 2020). Additionally, effective collaboration among diverse experts and implementing bias audits are essential steps in public health surveillance to ensure AI does not reinforce existing disparities (Flores, Kim, & Young, 2023).

AI Development and Deployment Ethics

The ethics of AI development and deployment emphasize the unintended consequences that arise from biases in training data or flawed model design, such as favoring one gender over another in job recruitment applications (Eitel-Porter, 2021). Some research reflects on existing ethics guidelines and notes they are often too abstract, fail to address structural issues and mainly reflect the values of the experts chosen to create them; highlighting the need for diversity and inclusivity to effectively combat bias in guidance (Hickok, 2020). The potential for biased algorithms to influence social equity through digital advertising, housing, job opportunities and gender representation indicates the broader social implications of AI (Engler, 2022). Furthermore, it presents an argument that while bias in AI exists, it can mitigate the high level of

bias and inconsistency demonstrated by individuals. For example, the "filter bubble" theory suggests that personalized content based on individuals' interests and engagement level can decrease information diversity and encourage polarization (Engler, 2022). Addressing ingrained biases in AI research, such as those in pathology, requires ethical guidelines to account for racial bias in algorithmic outcomes (Jackson et al., 2021).

Technical Strategies and Tools for Bias Mitigation

Technical strategies for addressing bias in AI are critical, including transparency, regular conformance testing and scheduled audits to detect and correct biases (Methnani et al., 2021). Bias testing and algorithmic fairness measures are essential to prevent discriminatory impacts, particularly in applications like facial recognition—which are widely used (Shneiderman, 2020). Implementing monitoring mechanisms in public administration can help mitigate unfair outcomes produced by predictive algorithms, such as unjust distinctions and the undermining of public administration's efforts to ensure individualized and equal justice (Bodó & Janssen, 2022). Bias management strategies such as fairness constraints in machine learning and publicly available, clear decision-making processes can help address ethical challenges in AI systems (Khan et al., 2022). The importance of global standards for AI in health care is underscored to ensure inclusiveness and to address disparities in performance across different socio-economic contexts (WHO, 2021).

Multidisciplinary and Participatory Approaches

Multidisciplinary approaches to AI development emphasize the use of tools like AI Fairness 360 to help developers identify and mitigate bias in models across different applications (Rossi, 2018). The concept of "algorithm-in-the-loop" systems, where humans retain ultimate oversight and final approval of content, is suggested to reduce bias risks associated with AI-generated public health messaging (Karinshak et al., 2023). High-quality, representative datasets and diverse development teams are seen as crucial for addressing bias throughout the AI lifecycle (Seppälä, Birkstedt, & Mäntymäki, 2021).

Autonomous Systems and Human Oversight

Automation bias refers to the tendency for people to favor or rely excessively on automated systems or technology, even when these systems are flawed or provide incorrect information. Individuals may perceive automated decisions or recommendations as more reliable or accurate than those made by humans, leading them to overlook errors or to disregard their own judgment

(Green, 2022). A potential mitigation technique is integrating human review of AI-generated content and outputs.

Addressing bias in autonomous systems involves ensuring that AI trained with uncurated datasets adheres to standards, such as IEEE P7003, to prevent bias (Winfield et al., 2019). Despite growing awareness of algorithmic bias risks, automation bias persists in human-AI interactions within public sector decision-making, suggesting the need for more effective oversight (Alon-Barkat & Busuioc, 2023). Oversight challenges include the potential for untrained overseers to be more susceptible to automation bias, underscoring the importance of understanding AI's limitations (Laux, 2023).

Challenges

The reviewed literature identifies several challenges for adequately addressing and recognizing bias in AI outputs and systems.

Definitional Ambiguity and Policymaking

Barriers to bias mitigation in AI often stem from definitional ambiguity, as the lack of consensus on what constitutes AI complicates conversations about effective regulatory policies and bias mitigation strategies (Krafft et al., 2020). Proprietary AI systems pose additional challenges; legal barriers such as trade secret laws and other regulations create transparency issues that hinder the detection and mitigation of biases in these systems (Calo, 2017). A lack of public trust driven by limited engagement, insufficient transparency and the absence of clear information creates "blind spots" in AI policy-making, particularly regarding concerns from populations of focus (Robles & Mallinson, 2023). Technical knowledge gaps and the rapid advancement of AI technology also overwhelm policymakers, making it difficult for them to keep pace with AI assurance initiatives and harmonize ethical principles with existing legislation (Biersmith & Laplante, 2022).

Health Care and Public Health Barriers

Bias mitigation in health care and public health is challenged by issues such as the lack of representativeness in training data, which leads to biased AI outputs, and the opaque nature of AI algorithms, which complicates the identification of biases, particularly in deep learning models (Reddy et al., 2020; Murphy, Di Ruggiero, & Upshur, 2021), which is a form of machine learning in which the computer network rapidly teaches itself to understand a concept without

human intervention. Ethical and regulatory challenges arise from the proprietary nature of AI technologies and concerns about data privacy and security, which obstruct efforts to implement bias mitigation (Thomasian et al., 2021; Kasula, 2021). In low- and middle-income countries, the absence of standardized methods for health technology assessments, as well as the challenge of avoiding ethnic, socio-economic and gender biases, further hinder bias mitigation efforts (Schwalbe & Wahl, 2020). Additionally, addressing bias in health care AI involves dealing with the lack of diverse training data, especially for rare diseases, and navigating the logistical challenges of continual bias auditing and surveillance (Thomasian et al., 2021).

Ethical Considerations in AI Development and Deployment

The black-box nature of AI systems presents a major barrier to bias mitigation, as developers may find it difficult to identify the sources of bias or apply appropriate fairness criteria for different applications (Rossi, 2018). Trade-offs between fairness and accuracy present dilemmas when trying to balance ethical considerations with the technical performance of AI systems (Jaume-Palasi, 2019). The dominance of private companies in discussions about AI ethics often diverts attention from underlying social, racial and economic issues, with abstract ethics guidelines failing to offer practical solutions for bias mitigation (Hickok, 2020). Bias also persists in widely used systems like facial recognition, where developers may resist acknowledging these biases due to their applications in law enforcement or commercial settings (Shneiderman, 2020).

Data-Related Barriers

Several challenges arise from data-related issues, including the perpetuation of societal biases embedded in algorithmic training data, which can be labor-intensive to correct through data pre-processing or post-processing techniques (Methnani et al., 2021). This training involves the process of teaching an algorithm, typically a machine learning model, to perform a specific task by exposing it to data. During this process, the algorithm learns patterns, features and relationships within the data to make predictions, classify information or perform other tasks. Ensuring data diversity is particularly challenging, as AI models trained on data from high-income countries may not perform effectively in different socio-economic contexts, thereby increasing the risk of bias (WHO, 2021). In public health surveillance, issues like data misrepresentation, unreliable annotators and reinforcement of societal biases through algorithmic training further complicate the process of bias mitigation (Flores, Kim, & Young, 2023).

Human Element in AI Oversight

Human oversight in AI presents additional barriers, such as automation bias, where decision-makers may overly rely on AI recommendations or selectively adhere to them based on existing stereotypes (Alon-Barkat & Busuioc, 2023). Implementing effective human oversight is challenging due to losing situational awareness, over-trust in AI systems and the difficulties of balancing socio-legal control (Tsamados, Floridi, & Taddeo, 2024). Furthermore, untrained overseers may be more prone to automation bias and that expert overseers may hurt their accuracy by over-relying on their own judgment versus the algorithm's recommendation (Laux, 2023). The effectiveness of bias mitigation can diminish if periodical mandates for auditors review of systems reduce their competence over time (Laux, 2023).

Global and Regional Limitations

The global distribution of AI development is affected by the concentration of financial resources and cultural dominance in a few countries, limiting the ability to adapt AI systems to different local contexts, leading to lagging deployment on smaller scales (Fournier-Tombs, 2023). Legal and intellectual property barriers further complicate the transfer and adaptation of AI technologies across various regions, presenting obstacles to implementing effective bias mitigation strategies early-on (Qin, 2024).

Recommendations

To address bias mitigation in AI policies, the literature offers several recommendations.

Definitional Clarity and Policy Alignment

The need for a clear, accessible definition of AI is crucial for effective policymaking. The research recommends developing a policy-facing definition that aligns with both current and future AI applications, facilitating the implementation of oversight procedures and regulations (Krafft et al., 2020). Additionally, the disconnect between policymakers and researchers on AI definitions should be addressed to avoid overlooking present AI technologies while also anticipating future innovations (Krafft et al., 2020). One recommendation to consider is to incorporate widely accepted definitions. One example to consider is the Organization for Economic Cooperation and Development (OECD), to guide policy development. The OECD defines an artificial intelligence (AI) system as a “machine-based system that can, for a given set of human defined objectives, make predictions, recommendations or decisions influencing

real or virtual environments." (OECD, 2019). There is also a need for new oversight mechanisms on the use and effects of AI in clinical practice, which must incorporate reflexive assessment of the scientific and social merits of AI-driven research and governance (Blasimme, 2020).

Ethics and Trust in AI Development

Several recommendations emphasize the need for ethical frameworks that uphold trust and accountability in AI. Policies should integrate transparency, explainability and bias detection to ensure AI systems can be trusted (Rossi, 2018). Proposals also suggest fostering public trust through citizen engagement, prioritizing public values in decision-making processes and conducting public education campaigns to increase awareness about AI technologies (Robles & Mallinson, 2023; Engler, 2022). Ethical AI development also should involve strong governance controls, including mandated training on AI for all organizational levels and establishing metrics to track adherence to AI principles (Eitel-Porter, 2021). The adoption of rights-based approaches grounded in international legal standards can further strengthen the ethical foundation for AI policies (Fukuda-Parr & Gibbons, 2021).

AI Governance and Regulatory Recommendations

Effective governance of AI requires the development of clear regulatory frameworks and ethical guidelines that address the unique challenges of AI deployment in various sectors, especially health care (Leimanis, et al. (2021); Reddy et al., 2020). Recommendations include establishing standards for data collection and quality management and strengthening regulatory oversight mechanisms to ensure compliance with ethical principles (Verma et al., 2020). The importance of collaborative approaches involving stakeholders from different fields also is highlighted, ensuring diverse perspectives inform regulatory policies (Biersmith & Laplante, 2022). In public health, AI governance should incorporate equity-focused frameworks to prevent the exacerbation of health disparities and to promote the use of AI for ethical, transparent and people-centered applications (Couture, V., et al., 2023; Silva Jr. et al., 2024; Flores et al., 2023).

Human Oversight and Decision-Making in AI

Recommendations suggest implementing policies that balance automation and human oversight, particularly in high-stakes settings like public health and government decision-making (Tsamados et al., 2024; Green, 2022). Ensuring that human decision-makers have a meaningful role in the oversight of AI systems can help mitigate risks associated with automation bias and

over-reliance on AI-generated recommendations (Alon-Barkat & Busuioc, 2023). Guidelines also should specify levels of human intervention and clearly define the responsibilities of human monitors. This can ensure that AI-enhanced decision-making processes are ethical and maintain a human-centered approach to technology adoption (Sele & Chugunova, 2024).

Data Management and Bias Mitigation

Addressing bias in AI requires robust data governance practices that promote transparency, representativeness and accountability (Kasula, 2021). Recommendations include developing auditing systems to evaluate bias, establishing universal guidelines for addressing algorithmic bias and creating cross-disciplinary teams to oversee AI deployments (Flores et al., 2023). Improving data literacy within organizations and engaging diverse communities can help identify and address sources of bias in AI models (Fisher & Rosella, 2022). Additionally, systematic risk analyses should be conducted to assess potential harms associated with specific AI deployments (Hickok, 2020).

Public Health and Equitable AI Use

AI policies in public health should prioritize ethical considerations and inclusive processes to ensure the benefits of AI are widely shared while avoiding harm to historically marginalized populations (Smith et al., 2020). Recommendations focus on addressing health inequities, promoting fairness in AI model development and ensuring that algorithms are tested across diverse population sub-groups (Thomasian et al., 2021). Proposals also include updating regulatory frameworks to account for the unique challenges of AI in health care, with a strong emphasis on data protection and algorithmic transparency (WHO, 2021).

Place-Based Policy Considerations

The international dimension of AI policy requires global cooperation to establish ethical guidelines and regulatory frameworks that uphold shared values, such as fairness and human rights (Khan et al., 2022). Regional policies, such as those recommended for the EU's AI Act, emphasize flexibility and adaptability to accommodate rapid technological advancements and varying local contexts (Laux, 2023; Ghimire & Edwards, 2023). Recommendations also stress the importance of capacity building and long-term financing for local AI ecosystems, particularly in low- and middle-income countries, to ensure sustainable development and ethical AI use (Fournier-Tombs, 2023).

Bibliography

1. Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1), 153−169. https://doi.org/10.1093/jopart/muac007
2. Biersmith, L., & Laplante, P. (2022). Introduction to AI Assurance for Policy Makers. 2022 IEEE 29th Annual Software Technology Conference (STC), 51–56. https://doi.org/10.1109/STC55697.2022.00016
3. Bodó, B., & Janssen, H. (2022). Maintaining trust in a technologized public sector. *Policy and Society, 41*(3), 414−429. https://doi.org/10.1093/polsoc/puac019
4. Calo, R. (2017). Artificial intelligence policy: Primer and roadmap. *U.C. Davis Law Review*, 51(2), 399-436. https://lawreview.sf.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/51-2_Calo.pdf
5. Green, B. Chander, A., Fjeld, J., Jobin, A., & Schwartz, P. M. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681. https://doi.org/10.1016/j.clsr.2022.105681
6. Couture, V., Roy, M. C., Dez, E., Laperle, S., & Bélisle-Pipon, J. C. (2023). Ethical implications of artificial intelligence in population health and the public's role in Its governance: Perspectives from a citizen and expert panel. *Journal of Medical Internet Research, 25*, e44357. https://www.jmir.org/2023/1/e44357
7. DSIT. (2023). *Emerging Processes for Frontier AI Safety*. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety
8. Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics, 1*, 73−80. https://doi.org/10.1007/s43681-020-00011-6
9. Engler, A. (February 1, 2022). The EU and U.S. are Starting to Align on AI Regulation. Brookings. https://www.brookings.edu/articles/the-eu-and-u-s-are-starting-to-align-on-ai-regulation/
10. Fisher, S., & Rosella, L. (2022). Priorities for successful use of artificial intelligence by public health organizations. *BMC Public Health, 22*(1), 2146. https://doi.org/10.1186/s12889-022-14422-z
11. Flores, L., Kim, S., & Young, S. D. (2024). Addressing bias in artificial intelligence for public health surveillance. *Journal of Medical Ethics*, 50(3), 190–194. https://doi.org/10.1136/jme-2022-108875

12. Fournier-Tombs, E. (2023). Local transplantation, adaptation, and creation of AI models for public health policy. *Frontiers in Artificial Intelligence, 6*:1085671. https://doi.org/10.3389/frai.2023.1085671
13. Fukuda-Parr, S., & Gibbons, E. (2021). Emerging consensus on 'Ethical AI': Human rights critique of stakeholder guidelines. *Global Policy, 12*(3), 254-265. https://doi.org/10.1111/1758-5899.12965
14. Ghimire, A., & Edwards, J. (2023). From Guidelines to Governance: A Study of AI Policies in Education. https://arxiv.org/html/2403.15601v1
15. Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, 1(1), 41-47. https://doi.org/10.1007/s43681-020-00008-1
16. Jackson, B. R., Ye, Y., Crawford, J. M., Becich, M. J., Roy, S., Botkin, J. R., de Baca, M. E., & Pantanowitz, L. (2021). The ethics of artificial intelligence in pathology and laboratory medicine: principles and practice. *Academic Pathology, 8*, 2374289521990784. https://doi.org/10.1177/2374289521990784
17. Jaume-Palasi, L. (2019). Why we are failing to understand the societal impact of artificial intelligence. *Social Research: An International Quarterly, 86*(2), 477-498. https://muse.jhu.edu/article/732186
18. Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), Article 116. https://doi.org/10.1145/3579592
19. Kasula, B. (2021). Ethical and Regulatory Considerations In AI-Driven Healthcare Solutions. *International Meridian Journal, 3*(3), 1-8. https://meridianjournal.in/index.php/IMJ/article/view/23
20. Kauffman, H. (2022). Artificial Intelligence and Public Health: A Descriptive Review of Use Cases. *Applied Research in Artificial Intelligence and Cloud Computing,* 5(1), 29–37. https://researchberg.com/index.php/araic/article/view/88
21. Khan, A.A., Badshah, S., Liang, P., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges, 383-392. *International Conference on Evaluation and Assessment in Software Engineering*. https://www.researchgate.net/publication/361385839_Ethics_of_AI_A_Systematic_Literature_Review_of_Principles_and_Challenges

22. Krafft, P. M., Young, M., Huang, K., Katell, M., & Bugingo, G. (2020). Defining AI in Policy versus Practice. Paper Presentation AIES '20, February 7–8, 2020, New York, NY, USA. https://doi.org/10.1145/3375627.3375835
23. Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a democratic design of AI governance under the European Union AI Act. Oxford Internet Institute. https://link.springer.com/article/10.1007/s00146-023-01777-z
24. Leimanis, A., & Palkova, K. (2021). Ethical Guidelines for Artificial Intelligence in Healthcare from the Sustainable Development Perspective. *European Journal of Sustainable Development*, 10(1), 90−102. https://doi.org/10.14207/ejsd.2021.v10n1p90
25. Davis, S., Singh, S., Srinidhi, N., & Wilson, M. (2024). Public health's inflection point with generative AI. McKinsey & Company. https://www.mckinsey.com/industries/social-sector/our-insights/public-healths-inflection-point-with-generative-ai#/
26. Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4, Article 737072. https://doi.org/10.3389/frai.2021.737072
27. Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics,* 22, Article 14. https://doi.org/10.1186/s12910-021-00577-8
28. Organization for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
29. Qin, H., & Li, Z. (2024). A study on enhancing government efficiency and public trust: The transformative role of artificial intelligence and large language models. *International Journal of Engineering and Management Research, 14*(3). https://doi.org/10.5281/zenodo.12619360
30. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. https://doi.org/10.1093/jamia/ocz192
31. Renda, A. (2019). Artificial intelligence: Ethics, Governance and Policy Challenges. Centre for European Policy Studies (CEPS) Task Force Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3420810

32. Robles, P., & Mallinson, D. J. (2023). Artificial intelligence technology, public trust, and effective governance. *Review of Policy Research, 40*, 1−18. https://doi.org/10.1111/ropr.12555
33. Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs, 72*(1), 127−134. https://www.jstor.org/stable/10.2307/26588348
34. Schwalbe, N., & Wahl, B. (2020). Artificial intelligence and the future of global health. *The Lancet*, 395. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30226-9/fulltext
35. Sele, D., & Chugunova, M. (2024). Putting a human in the loop.: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0298037
36. Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021). From ethical AI principles to governed AI. In *Proceedings of the Forty-Second International Conference on Information Systems* (ICIS 2021), Austin, TX. https://www.researchgate.net/publication/358234837_From_Ethical_AI_Principles_to_Governed_AI
37. Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4), Article 26. https://doi.org/10.1145/3419764
38. Silva Jr., J. B., Lima, N. T., Haddad, A. E., Galiano, S. G., Saiso, S. G., Valdez, M. L., ... & Kohan, P. (2024). From national and regional commitments to global impact: Artificial intelligence for equitable public health at the G20. *Revista Panamericana de Salud Pública 48*, e73. https://doi.org/10.26633/RPSP.2024.73
39. Smith, M. J., Axler, R., Bean, S., Rudzicz, F., & Shaw, J. (2020). Four equity considerations for the use of artificial intelligence in public health. Bulletin of the World Health Organization, 98(4), 290–292. https://pmc.ncbi.nlm.nih.gov/articles/PMC7133473/
40. Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics, 27*(15). https://doi.org/10.1007/s11948-020-00277-3
41. Blasimme, A., & Vayena, E., (2020). The ethics of AI in biomedical research, patient care, and public health. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford handbook of ethics of artificial intelligence.* Oxford University Press. https://bioethics.jhu.edu/wp-content/uploads/2021/10/Oxford_handbook.pdf

42. Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy*, 42(4), 602-611. https://doi.org/10.1057/s41271-021-00319-5
43. Tsamados, A., Floridi, L., & Taddeo, M. (2024). Human control of AI systems: from supervision to teaming. *AI and Ethics*, 4(4). https://doi.org/10.1007/s43681-024-00489-4
44. Verma, A., Rao, K., Eluri, V., & Sharma, Y. (2020). Regulating AI in public health: Systems challenges and perspectives. *Observer Research Foundation Occasional Paper, 261*. https://www.orfonline.org/public/uploads/posts/pdf/20230719010608.pdf
45. World Health Organization. (2021, June 28). *WHO Issues First Global Report on Artificial Intelligence (AI) in Health and Six Guiding Principles for its Design and Use.* https://www.who.int/news/item/28-06-2021-who-issues-first-global-report-on-ai-in-health-and-six-guiding-principles-for-its-design-and-use
46. Winfield, A. F., & Michael, K., Pitt J., & Evers V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE, 107*(3), 1-14. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8662743
47. World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO guidance.* https://www.who.int/publications/i/item/9789240029200

Research Question 3: How is data privacy addressed in AI policies?

Research Questions Examined in Articles

The literature examines how AI governance can address concerns related to data privacy, especially in sectors like public health, criminal justice, and business. A key focus is ensuring that AI systems respect individuals' privacy while maintaining transparency and accountability in data handling.

Bias, Trust and Data Privacy

The research highlights public concern over data privacy in AI applications, particularly in sensitive areas like hiring practices, criminal justice and health care. It explores how AI systems collect, process and store data, emphasizing the need for privacy protection mechanisms. Discussions include the balance between leveraging data for AI advancements and protecting

individual rights to privacy. Trust in AI is tied to how effectively data is handled and how transparent systems are about data use.

Common Policy Recommendations and Governance Approaches

The literature recommends establishing strong governance models to safeguard data privacy in AI applications. This includes creating regulatory frameworks that enforce transparency in AI data processing and storage, ensuring that AI systems comply with privacy laws and protecting sensitive personal data. Specific guidelines are suggested for health care AI, where privacy concerns are heightened, emphasizing the need for secure data infrastructure and ethical data usage in public health and clinical practice.

Summary of Key Findings – Data Privacy

The literature notes several data privacy challenges for AI used in public health and health care settings and contexts.

Data Privacy Concerns in AI Systems

The rapid growth of artificial intelligence (AI) brings significant data privacy challenges, particularly in the context of freely shared information and pattern recognition capabilities that can expose sensitive personal details. This shift in privacy discourse emphasizes the need for mechanisms to achieve data parity, without compromising personal privacy. AI's acceleration is leading to challenges in managing privacy within data-intensive environments such as large and complex datasets and databases (Calo, 2017).

Concerns about data privacy extend to personal identifiable information (PII), especially when AI automates data processes. One study, involving survey findings pertaining to perception of AI, show that 68 percent of respondents anticipate problems with privacy and civil rights due to AI, with 74 percent rating data privacy as "very important" (Robles & Mallinson, 2023). Additionally, education levels and gender influence these concerns, with higher education correlating to greater worry about privacy issues. Female respondents show 32 percent higher odds of considering AI safe, 37 percent higher odds of supporting its development and 35 percent higher odds of perceiving AI-related harms as unlikely (Robles & Mallinson, 2023).

Regulatory Frameworks and Data Handling Practices

The European Union's General Data Protection Regulation (GDPR) has been found to be a foundational framework in regulating data privacy within AI systems. This policy also applies to U.S. businesses that process personal data about EU or European Economic Area (EEA) citizens, or that target those citizens for goods or services. The framework establishes essential standards for transparency, fairness and traceability, mandating that AI system performance data and design choices be recorded in an "AI factsheet" to ensure compliance and accountability (Rossi, 2018). California, Colorado, Connecticut, Utah and Virginia have consumer privacy laws inspired by the GDPR (Troutman, 2022). Additionally, some data privacy laws discussed in the literature require that only essential data be collected to avoid non-compliance risks. The risks associated with improper data combinations further complicate regulatory adherence, highlighting the need for careful data governance practices, review and oversight (Eitel-Porter, 2021).

To address privacy risks, approaches like differential privacy and the use of public datasets have been suggested. Differential privacy can minimize the risks of analyzing sensitive information and transparency remains crucial in managing patient data (Reddy, Allan, Coghlan, & Cooper, 2020). In health care, the ethical handling of patient data, including obtaining explicit consent from patients for data usage, is critical to minimizing privacy breaches and ensuring legal compliance (Rai, 2020).

AI-Driven Public Health: Ethical Considerations and Risks

The integration of AI into public health introduces various data privacy and security risks, especially concerning the large-scale use of personal health information.

One significant risk involves the re-identification of anonymized data, which remains a concern even when data is de-identified. For instance, hacking incidents like the one in Mumbai, India, where 35,000 medical records were leaked, illustrate the potential harm when sensitive health data is leaked. This case reinforces the need for robust privacy protection measures and facilitating public trust in AI applications (Murphy et al., 2021).

Challenges in Ensuring Transparency and Accountability

Ensuring transparency in AI decision-making is a significant challenge, especially given the "black box" nature of many AI systems, where the internal workings and decision-making

processes are complex, opaque, and not easily understood or accessible even to experts. The lack of clear explainability and the opaque nature of some AI decision processes make it difficult to meet legal requirements for transparency and explicability (Leimanis & Palkova, 2021). If modelled, compliance with GDPR stipulates that AI systems should be explainable and accountable, but many organizations struggle to implement these requirements effectively (Burrell, 2016). The need for transparency extends to maintaining data protection standards, especially regarding health data, which is categorized as sensitive information and demands strong privacy measures (Rai, 2020).

Regulations like GDPR and the California Privacy Rights Act (CPRA) provide a legal framework for safeguarding personal data in AI applications. They impose obligations such as securing data through encryption, implementing transparency measures and ensuring data subjects' rights are protected, including the right to human intervention in automated decision-making (Domingo, 2022; Renda, 2019).

Data Privacy in AI-Driven Decision-Making

The risks associated with the circulation of confidential health data in AI-driven health care systems are ever-growing. AI models need large datasets for training, which heightens the risk of privacy violations and brings attention to the challenge of maintaining data anonymity while ensuring the accuracy of predictive models (Murphy, Di Ruggiero, & Upshur, 2021). Moreover, data governance issues arise when health data is shared with multiple entities, making it difficult to maintain control and prevent re-identification of individuals (Rossi, 2018).

The resource limitations faced by new enterprises attempting to meet stringent data privacy standards, such as those required for anonymization and de-identification, further complicate the development of AI systems. As a result, compliance becomes both cost-intensive and challenging for smaller organizations with limited resources and capacity (Rai, 2020).

Public Health Implications of Data Privacy and Ethical Governance

AI's role in public health also brings into focus the ethical implications of using personal health data without adequate privacy protections. One example cited in the literature is the use of COVID-19 contact tracing applications that utilized Bluetooth to notify users if they had been within two meters of an infected person. The Bluetooth-based framework was designed to avoid storing or sharing personally identifiable information. However, researchers attributed the application's low adoption rate to concerns about privacy, data governance, and human rights,

as well as technological and practical challenges such as battery drainage on older phones and difficulties verifying positive diagnoses through health codes. (Fournier-Tombs, 2023).

As the volume of health data collected by AI systems increases, the potential for breaches and misuse rises, highlighting the critical need for robust ethical frameworks and regulatory guidelines that protect privacy while enabling responsible innovation in AI-driven health care solutions (Burrell, 2016).

Challenges

The reviewed literature identifies several challenges in addressing risks for data privacy and security in AI systems and the use of AI tools:

Challenges in Implementing Data Privacy and Security Measures

The implementation of AI presents significant challenges in addressing data privacy and security concerns, including crafting policies to avoid legal scrutiny in the U.S., managing data sharing without sacrificing privacy and overcoming barriers to accountability related to laws like the U.S. Digital Millennium Copyright Act of 1998 (Calo, 2017). The lack of a comprehensive legal framework for regulating AI algorithms and ensuring data privacy once data is in AI systems further complicates these efforts (Robles & Mallinson, 2023). Implementing privacy-preserving techniques without compromising AI performance and ensuring compliance with existing regulations also pose considerable difficulties (Rossi, 2018). Additional challenges arise from risks associated with improper data combination, reluctance to report concerns, rushed development cycles and the use of AI outside its original context (Eitel-Porter, 2021). Moreover, health care AI applications face specific issues in obtaining genuine patient consent, preventing data breaches and managing risks associated with using public datasets (Reddy et al., 2020).

Transparency, Accountability and Regulatory Compliance

The need for transparency in AI systems is a recurring challenge. Ensuring stronger privacy measures for health data sharing, especially given the cumbersome nature of anonymization processes, adds to the implementation complexity (Rai, 2020). The risks associated with using proprietary software for specific smart health devices that collect health data, further complicate efforts to maintain data security and transparency (Murphy et al., 2021). A lack of understanding in machine learning processes and systems complicates efforts to monitor compliance and address privacy concerns, making it difficult to ensure ethical data management (Burrell, 2016).

Data Governance and Ethical Concerns

The integration of AI into public health systems brings unique challenges related to data governance, particularly when balancing transparency and privacy in blockchain, information-sharing systems (Bodó & Janssen, 2022). Blockchain is a decentralized digital ledger that electronically records, stores, and verifies data or transactions across a network. It ensures transparency, immutability, and security, making it a reliable tool for managing and sharing information while protecting privacy. Anonymizing and securing health data while navigating policies like European Union's General Data Protection Regulation GDPR) and Health Insurance Portability and Accountability Act (HIPAA) regulations and ensuring ethical use and informed consent adds layers of complexity. Protecting personal information is not only a legal requirement but also an ethical necessity, especially given challenges such as capturing high-volume data securely and the monopolization of data by large technology giants (Shneiderman, 2020; Fukuda-Parr & Gibbons, 2021). Achieving transparency, particularly in explaining AI decisions to users, is another significant challenge, requiring a balance between legal norms and user expectations (Felzmann et al., 2019).

Risks in AI-Driven Health Care Solutions

Health care applications of AI face specific hurdles in ensuring data privacy and security. For instance, sharing sensitive datasets, such as facial images, may not be feasible due to privacy concerns (Thomasian, et al., 2021). Unauthorized access to health-related data and ethical challenges in extracting detailed information from AI-processed data also pose risks (Hamet & Tremblay, 2017). The difficulty in achieving meaningful human oversight and review, especially in high-risk AI systems, complicates efforts to ensure ethical data management practices (Domingo, 2022; Laux, 2023).

Balancing Innovation with Privacy Regulations

Meeting policy requirements, addressing algorithmic biases and providing comprehensive training to keep up with rapid technological changes are crucial for mitigating privacy risks (Khan et al., 2022; Elendu, C., et al., 2023). The ethical implications of using personal health data necessitate robust governance frameworks and continuous adaptation to evolving regulatory landscapes (Murphy et al., 2021; Leimanis & Palkova, 2021).

Recommendations

To address data privacy and security challenges in AI-driven public health and health care, the following recommendations were referenced in the reviewed articles.

Policy and Governance Recommendations

Recommendations emphasize crafting policies that address data parity challenges and ensure responsible data-sharing practices while avoiding infringements on free speech. To support privacy, interventions such as legal safeguards and incentives for compliance with privacy standards, even as AI becomes more democratized, are necessary. Furthermore, creating legal mechanisms to foster data parity without compromising privacy and to overcome barriers to accountability — such as those posed by trade secret law or anti-circumvention regulations — Is advised (Calo, 2017). Public input is recommended as a core principle for AI governance, with transparent methods and active citizen engagement in policymaking to address the lack of ideal regulatory frameworks for AI. This includes considering public concerns about privacy, civil rights and integrating input and public engagement to facilitate public trust alongside transparency and risk mitigation as equal governance priorities (Robles & Mallinson, 2023).

Technical and Ethical Standards

Actionable measures involve implementing data minimization techniques by collecting only the data necessary for AI models to function, thereby reducing risks of non-compliance with regulations like GDPR (Domingo, 2022). Establishing structured mechanisms for employees to report AI-related concerns without repercussions is critical, as is documenting AI model development with a focus on system integrity, bias and transparency in a customizable sign-off process. Continuous testing of AI systems for ethical compliance throughout their lifecycle also is recommended (Eitel-Porter, 2021). In the health care domain, constituting a data governance panel with representatives from patients, clinicians and AI experts to review training datasets for representativeness is advised. Additionally, differential privacy techniques should be employed to protect sensitive patient data, while public datasets should be prioritized to reduce privacy breach risks (Reddy et al., 2020).

Transparency, Accountability and Citizen Engagement

Ensuring transparency across all stages of the AI lifecycle is recommended to foster fairness and accountability. Technical measures such as explainable AI (XAI) to clarify decision-making

processes and privacy-preserving techniques to secure personal data are critical. Actionable steps include setting requirements for comprehensive testing protocols, audit trails and adversarial testing, which is the evaluation of how a system behaves when given harmful or malicious inputs to assess the robustness of AI systems. Non-technical actions should involve algorithmic impact assessments and engaging civil society in decision-making (Stix, 2021). To build public trust, continuous monitoring for bias in AI systems, clear communication strategies about data usage and citizen involvement in AI governance are essential. Furthermore, integrating transparency into AI system components, such as datasets, models and frameworks, can facilitate local adaptation and control (Qin & Li, 2024; Couture et al., 2023).

Data Security and Privacy Preservation Techniques

Data anonymization and de-identification are recommended to protect individual privacy, especially in health data applications. Implementing secure data storage and transfer protocols, including encryption and access controls, is essential to safeguard against unauthorized access. Informed consent procedures should be robust, clearly communicating the purpose, scope and risks associated with AI data processing. Collaborative approaches such as federated learning can be used to enable AI model training without transferring sensitive patient data, while techniques like cryptographic methods and differential privacy can be employed to reconcile data protection with data availability (Thomasian., et al., 2021; Renda, 2019).

Ethical AI Development and Human Rights Considerations

Establishing comprehensive ethical guidelines focused on principles like transparency, accountability and fairness is recommended for ethical AI use in public health. Legal frameworks should mandate transparency in AI algorithms, including the disclosure of data sources, methodologies and metrics used to inform decisions. Policies also should prioritize protecting human rights, such as privacy and data protection, and include mechanisms for redress for individuals adversely affected by AI decisions. Incorporating diverse perspectives during AI development and addressing biases through debiasing techniques and retraining on representative datasets are crucial actions. Furthermore, the adoption of governance frameworks that explicitly address ethical AI practices can mitigate risks and uphold human dignity and rights (Fukuda-Parr & Gibbons, 2021; Murphy et al., 2021; Leimanis & Palkova, 2021).

Bibliography

1. Bodó, B., & Janssen, H. (2022). Maintaining trust in a technologized public sector. *Policy and Society, 41*(3), 414−429. https://doi.org/10.1093/polsoc/puac019
2. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society.* https://doi.org/10.1177/2053951715622512
3. Calo, R. (2017). Artificial intelligence policy: Primer and roadmap. *U.C. Davis Law Review*, 51(2), 399-436. https://lawreview.sf.ucdavis.edu/sites/g/files/dgvnsk15026/files/media/documents/51-2_Calo.pdf
4. Chen, H., Zeng, D., Buckeridge, D. L., Izadi, M. I., Verma, A., Okhmatovskaia, A., Hu, X., Shen, X., Cao, Z., & Wang, F. Y., et al. (2009). AI for global disease surveillance. *IEEE Intelligent Systems, 24,* 66−82. https://ailab-ua.github.io/courses/MIS510/6c-globalsurveillance-2009.pdf
5. Couture, V., Roy, M. C., Dez, E., Laperle, S., & Bélisle-Pipon, J. C. (2023). Ethical implications of artificial intelligence in population health and the public's role in its governance: Perspectives from a citizen and expert panel. *Journal of Medical Internet Research, 25,* e44357. https://doi.org/10.2196/44357
6. Domingo, S. (2022). Human Intervention and Human Oversight in the GDPR and AI Act. *Trilateral Research.* https://trilateralresearch.com/emerging-technology/human-intervetion-in-gdpr-and-ai
7. Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, 1, 73−80. https://doi.org/10.1007/s43681-020-00011-6
8. Elendu, C., Amaechi, D. C., Elendu, T. C., Jingwa, K. A., Okoye, O. K., Okah, M. J., Ladele, J. A., Farah, A. H., & Alimi, H. A. (2023). Ethical implications of AI and robotics in healthcare: A review. Medicine, 102(50), e36671. https://doi.org/10.1097/MD.0000000000036671
9. Felzmann, H., Fosch Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 5*(2), 1−14. https://doi.org/10.1177/2053951719860542
10. Fournier-Tombs, E. (2023). Local transplantation, adaptation, and creation of AI models for public health policy. *Frontiers in Artificial Intelligence, 6,* 1085671. https://doi.org/10.3389/frai.2023.1085671

11. Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy, 12*(Suppl.6), S6. https://doi.org/10.1111/1758-5899.12965
12. Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism, 69(Supplement)*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011
13. Khan, A.A., Badshah, S., Liang, P., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges, 383-392. *International Conference on Evaluation and Assessment in Software Engineering*. https://www.researchgate.net/publication/361385839_Ethics_of_AI_A_Systematic_Literature_Review_of_Principles_and_Challenges
14. Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a democratic design of AI governance under the European Union AI Act. Oxford Internet Institute. https://link.springer.com/article/10.1007/s00146-023-01777-z
15. Leimanis, A., & Palkova, K. (2021). Ethical Guidelines for Artificial Intelligence in Healthcare from the Sustainable Development Perspective. *European Journal of Sustainable Development*, 10(1), 90−102. https://doi.org/10.14207/ejsd.2021.v10n1p90
16. Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4, Article 737072. https://doi.org/10.3389/frai.2021.737072
17. Mitchell, M., et al. (2019). Model cards for model reporting. *Conference on Fairness, Accountability, and Transparency,* 328−596. https://doi.org/10.1145/3287560.3287596
18. Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics,* 22, Article 14. https://doi.org/10.1186/s12910-021-00577-8
19. Qin, H., & Li, Z. (2024). A study on enhancing government efficiency and public trust: The transformative role of artificial intelligence and large language models. *International Journal of Engineering and Management Research, 14*(3), 1−10. https://doi.org/10.5281/zenodo.12619360
20. Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137−141. https://doi.org/10.1007/s11747-019-00710-5
21. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. https://doi.org/10.1093/jamia/ocz192

22. Renda, A. (2019). Artificial intelligence: Ethics, Governance and Policy Challenges. Centre for European Policy Studies (CEPS) Task Force Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3420810
23. Robles, P., & Mallinson, D. J. (2023). Artificial intelligence technology, public trust, and effective governance. *Review of Policy Research, 40*, 1−18. https://doi.org/10.1111/ropr.12555
24. Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs, 72*(1), 127−134. https://www.jstor.org/stable/10.2307/26588348
25. Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems, 10(4), Article 26. https://doi.org/10.1145/3419764
26. Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics, 27*(15). https://doi.org/10.1007/s11948-020-00277-3
27. Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy*, 42(4), 602-611. https://doi.org/10.1057/s41271-021-00319-5
28. Troutman Pepper. (2022, December 13). U.S. state privacy laws in 2023: California, Colorado, Connecticut, Utah, and Virginia. https://www.troutman.com/insights/us-state-privacy-laws-in-2023-california-colorado-connecticut-utah-and-virginia.html

Research Question 4: How is transparency addressed in AI policies?

Research Questions Examined in Articles

Key research questions center on the transparency of AI technologies, particularly in the context of governance and ethical implications. Studies emphasize the need for clear guidelines to improve AI systems' transparency, particularly in public health applications. The literature highlights the importance of making AI decision-making processes explainable to both policymakers and the public to ensure accountability and enhance public trust.

AI Transparency in Public Health Surveillance and Communication

Transparency in AI-driven public health initiatives, such as disease surveillance and public communication, is a significant focus of the literature. Researchers call for transparent audit

systems and guidelines that make AI outputs in public health understandable and accessible, especially when using natural language processing (NLP) techniques and large language models (LLMs). These measures are necessary to prevent misinformation and improve the effectiveness of AI-powered health promotion and emergency response efforts.

Enhancing Public Trust Through Transparent AI Systems

Multiple studies explore how transparent AI systems can foster public trust, particularly when AI is integrated into public health operations. Findings indicate that transparency, combined with clear communication of AI's role and limitations, is essential for building trust and legitimacy among users. The literature underscores that transparency in AI models helps identify biases, supports ethical AI deployment and aligns with public expectations for fairness and accountability.

Challenges in AI Transparency and Policy Implementation

The research also identifies challenges related to implementing transparency in AI policy. While transparency is crucial for informed consent and ethical oversight, the complex and opaque nature of deep learning models often limits transparency. As a result, studies suggest the need for dynamic regulatory frameworks that address these limitations while promoting greater transparency in AI's design, deployment and decision-making processes within public health systems.

Summary of Key Findings – Transparency

The regulation of AI technologies faces a critical challenge due to the lack of a clear and universally accepted definition of AI. This conceptual ambiguity complicates the alignment between research and policy, potentially leading to unintended consequences in AI governance. Despite the prevalence of policy recommendations, a significant proportion of regulatory documents fail to define AI explicitly, underscoring the need for precise, policy-facing definitions to ensure effective implementation. Additionally, the integration of equity and human-centered considerations into AI definitions is essential for addressing economic inequality and discrimination, which are central concerns in AI data transparency. Understanding and harmonizing these definitional and regulatory challenges across regions is fundamental to fostering responsible AI development and adoption.

Regulatory Clarity and AI Definition

There is a significant lack of conceptual clarity surrounding AI's definition for regulatory purposes, which contributes to a disconnect between how researchers and policymakers define AI, potentially leading to harmful and unintended consequences in policy implementation. Notably, 38 percent of the analyzed policy documents did not define AI, yet still issued policy recommendations, highlighting the need for policy-facing definitions to ensure that guidance is relevant and effectively implemented (Krafft et al., 2020). Studies further reveal that equity is a central concern in AI data transparency; 82 percent of AI researchers recognize economic inequality as a relevant issue, while 83 percent agree that discrimination and oppression are also relevant concerns (Krafft et al., 2020). These insights underscore the need for AI definitions and frameworks that incorporate human-centered considerations to address equity in research and policy (Krafft et al., 2020). The Bletchley Declaration emphasizes that the EU AI Act mandates transparency for AI systems interacting with humans or generating synthetic content, requiring such systems to be explicitly identified as AI (Bletchley Declaration, 2023). Different regions adopt varied regulatory approaches, with the EU focusing on stringent regulations and foresight, the U.S. leaning toward pragmatic, industry-aligned regulation and China emphasizing macro-level policies promoting innovation (Hu & Li, 2024; Qin & Li, 2024).

Public Trust and AI Transparency

Algorithmic transparency, especially for AI used in public sector decision-making, is crucial for legitimizing AI outcomes among citizens. While transparency can enhance legitimacy, not all opacity is unjustified, as AI's inherent complexity often makes full algorithmic transparency impractical or unintelligible to the public (Robles & Mallinson, 2023). Additionally, disparities in AI support and concerns over potential risks vary significantly across different demographics — such as education levels, race, gender and political ideologies — highlighting the importance of equity in data transparency to build public trust and support effective governance (Robles & Mallinson, 2023). Developing procedural fairness in AI application could help build public confidence without requiring complete transparency. Transparency and explainability are essential for achieving public trust, which includes disclosing when AI is being used and providing insight into how decisions are made within AI systems (Fisher, et al., 2022; Taeihagh, 2021). However, the relationship between AI policies and public trust is not consistently observed across studies, suggesting a need for more data and time to explore this relationship fully (Taeihagh, 2021). In some cases, transparency can reduce trust if AI predictions, even

when accurate, do not align with users' intuitive understanding, indicating that the effects of transparency on trust are more nuanced than previously thought (Schmidt et al., 2020).

AI in Health Care

The use of AI in health care presents unique transparency challenges, with a significant concern being the "black box" nature of AI models, which can reduce trustworthiness and impair the validation of clinical recommendations (Reddy et al., 2020). There is a noted concern among health care professionals about their inability to scrutinize AI systems' outputs, especially as these systems become more complex, making their inner workings harder to understand (Murphy et al., 2021). Transparency in health care AI involves not only making decisions intelligible to medical practitioners and patients but also ensuring that AI-driven tools are clear about limitations and risks (Leimanis & Palkova, 2021). AI transparency is linked to trust, where issues such as algorithmic opacity can reinforce biases, lead to discrimination and perpetuate inequity in health care (Kasula, 2021). Additionally, flexible regulatory approaches, such as the use of regulatory sandboxes, can allow experimentation with AI regulations in controlled environments under regulatory supervision, enabling organizations to test new technologies while ensuring safety and compliance. This approach facilitates the adaptation of models to local needs while maintaining transparency and mitigating potential risks (Verma et al., 2020; Fournier-Tombs, 2023).

Ethical AI and Governance Frameworks

Ethical guidelines frequently prioritize transparency, although interpretations and implementations differ across frameworks. For instance, a meta-analysis of AI ethics guidelines found that over half emphasized themes like transparency, justice and privacy (Hickok, 2021). Further exploration of AI policies across 25 countries found that ethical principles such as justice, fairness, transparency and privacy are emphasized frequently, with distinct variations between private sector and government policy discussions. This variation underscores the need for diverse perspectives to mitigate risks and promote responsible AI strategies (Biersmith & Laplante, 2022). Transparency is not just about making AI systems explainable but also involves dynamic task allocation, communication of performance metrics and other strategies that provide meaningful insights to users (Zerilli et al., 2022). Explainability in AI helps identify potential biases, supports user control and enhances the accuracy of decisions (Shneiderman, 2020). The use of transparency-by-design approaches, such as embedding privacy, security and explainability from the beginning, is essential for ensuring ethical AI (Seppälä et al., 2021). Furthermore, multi-stakeholder feedback mechanisms and cross-sectoral collaborations can

help bridge the gap between ethical principles and practical policy recommendations by supporting the actionability of transparency guidelines through testing and validation protocols (Stix, 2021).

Policy Interventions and AI Governance

The regulatory landscape for AI governance differs significantly across regions. In the EU, comprehensive regulations like the AI Act emphasize harmonization and stringent requirements, while in the U.S., policies are more pragmatic and closely aligned with industry innovation to avoid stifling technological growth. China's regulatory approach focuses on promoting AI development and fostering an ecosystem conducive to innovation (Hu & Li, 2024; Qin & Li, 2024). Transparency is a foundational requirement for human oversight in AI, including the disclosure of whether AI is involved in the decision-making process and details about the AI's development, such as training data and model characteristics (Laux, 2023).

Algorithmic bias in public sector decision-making remains a critical concern, as it can perpetuate discrimination when unchecked. For example, the Dutch childcare benefits scandal highlighted the dangers of algorithmic predictions aligning with existing stereotypes against minority groups, leading to unjust discrimination (Alon-Barkat & Busuioc, 2023). However, studies suggest that providing algorithmic explanations does not necessarily improve human oversight; in fact, it can lead to overreliance on AI recommendations even when incorrect (Green, 2022). Collaborative, future-oriented policy frameworks are needed to navigate the complexities of AI governance and to ensure AI is ethically integrated into society (Renda, 2019).

Challenges and Limitations of Transparency

Complete transparency in AI systems may not always be feasible or desirable due to technical, social or economic factors. For instance, deep learning models are often considered "black boxes" because their inner workings are complex and difficult to interpret, which presents challenges for ensuring transparency and accountability (Data Governance Institute, n.d.; Khan et al., 2022). Transparency's impact can vary by context; it can enhance user satisfaction in some cases, such as with music recommendations, but may not always lead to increased trust in other contexts, like social media algorithms (Felzmann et al., 2019). Moreover, incorrect or overly simplistic transparency may lead to automation complacency, information overload or misinterpretation, which diminishes the intended benefits of transparency efforts (Zerilli et al., 2022). Thus, transparency should be implemented with a focus on the specific context and requirements of different AI applications (Methnani et al., 2021). Transparent AI systems should

not only provide explanations but also ensure that users understand the significance of the transparency measures and how to apply the information provided (Schmidt et al., 2020).

Challenges

Ambiguity and Lack of Clear Definitions

The lack of a clear, policy-facing definition of AI presents significant implementation challenges, as it creates uncertainty about which systems fall under regulatory frameworks. This ambiguity makes it difficult to develop and enforce appropriate and effective regulatory policies and can result in a disconnect between how policymakers and AI researchers define AI. Such discrepancies can lead to harmful and unintended consequences in policymaking, as the absence of a unified definition hampers the development of meaningful guidelines that apply to the relevant technologies (Krafft et al., 2020). Additionally, defining transparency in AI is itself a challenge, as it encompasses more than just disclosing when AI is used. It involves explaining the system's purpose, the factors influencing decision-making and how these decisions are made in a manner that is understandable to the public (Taeihagh, 2021).

Balancing Transparency with Privacy and Intellectual Property

A key challenge in implementing data transparency is balancing the need for transparency with the protection of sensitive data and proprietary information. While transparency can foster public trust, it must be carefully managed to avoid compromising privacy, especially in areas like health care where patient data is involved. The use of public datasets to train AI models is one approach to mitigate privacy risks, but it raises concerns about data quality and the representativeness of datasets used for AI training (Reddy et al., 2020). Additionally, organizations face difficulties in disclosing sufficient information about AI systems without revealing proprietary details, as legal constraints related to intellectual property can limit the level of transparency achievable (Renda, 2019; Fisher & Rosella, 2022). Businesses may also need to update policies to stay compliant with evolving laws governing AI usage.

Technical Complexity and the Black-Box Problem

The black-box nature of many AI models, particularly those employing deep learning, poses significant challenges for data transparency. This opacity makes it difficult to interpret the data, processes and logic underlying AI decisions, limiting the ability of users and regulators to scrutinize outputs. The proprietary nature of many AI systems adds another layer of complexity, as it often restricts access to critical information needed for evaluating AI models (Rossi, 2018; Flores et al., 2023). In health care, this lack of interpretability impairs the validation of clinical

recommendations, reduces trust in AI-generated decisions and poses safety risks due to the difficulty in identifying errors or biases in the algorithms (Murphy et al., 2021). Commercially available AI systems frequently exhibit these transparency issues, making it challenging for organizations to ensure accountability and public trust (Crossnohere et al., 2022).

Harmonizing Standards and Keeping Pace with Technological Change

The rapid pace of AI advancements often outstrips regulatory capabilities, creating gaps between existing laws and the emerging technologies that need oversight. Harmonizing various legal requirements, professional standards and AI-related mandates across jurisdictions is a persistent challenge, particularly for complex and high-risk AI models known as "frontier AI" (Biersmith & Laplante, 2022). These technologies, due to their novelty and potential risks, can be deployed quickly without thorough verification and validation, leaving a regulatory lag. Policymakers must therefore engage in ongoing efforts to keep AI regulations aligned with technological developments, while also standardizing methods of transparency across different sectors and applications (Taeihagh, 2021).

Providing Meaningful Explanations

While transparency is often associated with making AI systems explainable, the process of providing meaningful and understandable explanations for AI decisions remains difficult. Explainable AI (XAI) methods like Local Interpretable Model-Agnostic Explanations (LIME), Partial Dependence Plots and SHapley Additive exPlanations aim to improve interpretability but may not always be sufficient to convey the underlying mechanisms of AI models to non-experts (Flores et al., 2023). In addition to the technical complexity of these methods, there is the risk of information overload, where too much information or overly detailed explanations can overwhelm users and decrease the effectiveness of transparency efforts. Therefore, AI explanations must be designed in a way that is accessible and relevant to the intended audience to avoid overtrust and ensure that users accurately understand the limitations and capabilities of AI systems (Zerilli et al., 2022).

Ethical Considerations and Accountability

Ensuring ethical transparency involves more than merely providing explanations for AI decisions; it also requires addressing biases in data, monitoring AI systems throughout their lifecycle and maintaining accountability for AI-driven outcomes. The abundance of bias metrics and fairness notions complicates the selection of suitable measures for specific contexts, making bias detection and mitigation a challenging process (Rossi, 2018). In health care,

ensuring transparency in AI systems necessitates making clinical decisions and AI functionalities intelligible to medical practitioners and patients. This can involve regulatory considerations that add further complexity to implementing data transparency (Leimanis & Palkova, 2021). Establishing robust accountability frameworks also is crucial, as responsibility for AI decisions often spans multiple stakeholders, including developers, health care professionals and regulators (Kasula, 2021; Winfield et. al., 2019).

Resource and Expertise Limitations

Implementing effective transparency measures often requires significant resources, including financial investment, specialized expertise and ongoing training for personnel. Smaller organizations and public institutions may find it difficult to allocate the necessary resources for developing and maintaining robust transparency practices. Technical knowledge gaps, coupled with the rapid evolution of AI technologies, necessitate continuous education and capacity-building efforts for employees to stay up-to-date on best practices (Eitel-Porter, 2021; Seppälä et al., 2021). Furthermore, keeping up with the latest transparency tools and methodologies can be a demanding task for organizations with limited budgets and staff (Bodó & Janssen, 2022).

Fragmentation and Data Interoperability Issues

Data fragmentation and lack of interoperability between different systems present significant obstacles to implementing data transparency, particularly in sectors like health care and public administration. AI models rely on comprehensive and high-quality datasets for accurate functioning, but inconsistencies in data formats, standards and terminologies can hinder the integration of diverse data sources (Verma et al., 2020).

Potential for Misplaced Trust and Overreliance

Transparency in AI systems does not always lead to improved outcomes; it can sometimes result in misplaced trust or overreliance on AI-generated recommendations. Studies indicate that explanations for algorithmic decisions can cause users to trust AI outputs even when recommendations are flawed, emphasizing the importance of designing transparency measures that account for human cognitive limitations and biases (Green, 2022). Poorly designed transparency initiatives may inadvertently cause harm by giving users a false sense of security, making it necessary to carefully calibrate the information provided (Zerilli et al., 2022).

Cost and Complexity of Auditing

Auditing AI systems to ensure transparency can be resource-intensive and time-consuming,

particularly when it involves third-party reviews. Companies may resist external audits due to concerns about protecting proprietary information, and the process of continuous monitoring to ensure compliance with transparency standards requires specialized skills and significant investment (Stone et al., 2016; WHO, 2021). Additionally, addressing transparency throughout the entire AI lifecycle, including the design, deployment and operational phases, is an ongoing challenge for organizations seeking to maintain regulatory compliance (WHO, 2021).

Addressing Bias and Ensuring Fairness

Implementing data transparency also involves addressing bias and ensuring fairness in AI models, which can be challenging due to the complex nature of AI algorithms and the dynamic nature of real-world data. Different approaches to debiasing and conforming to fairness standards require extensive resources and continuous testing (Rossi, 2018). Ensuring that transparency is maintained across the entire AI lifecycle, including monitoring for biases that may emerge over time, is essential for building trustworthy AI systems (Methnani et al., 2021). This ongoing effort is crucial for mitigating social inequities and ensuring that AI applications do not perpetuate or amplify existing disparities. The following recommendations were referenced in the reviewed articles.

Recommendations

The reviewed literature identifies several challenges for ensuring transparency around the use of AI and generated outputs from AI systems:

AI Governance and Public Trust

The importance of data transparency and integrity in building public trust in AI systems is well-documented. Ensuring algorithmic transparency, particularly in public sector applications, is crucial. However, not all algorithmic opacity is inherently negative, especially when sensitive data is involved. Public education on data use in government decision-making is recommended, as well as legal protections for personal data handled by AI systems. Involving public input in AI governance is emphasized as a co-equal principle alongside transparency and risk mitigation (Robles & Mallinson, 2023).

A comprehensive approach to AI governance includes fostering an environment conducive to innovation and public trust. This involves developing technical standards for AI, incorporating privacy design principles and ensuring strict scientific integrity. Recommendations also stress

the need for guidelines on AI system development, accountability measures and education on compliance with relevant regulations and policies (Biersmith & Laplante, 2022).

Ethical AI Development

Ethical AI development prioritizes transparency, accountability, fairness and safety. IBM's principles advocate for AI to augment rather than replace human intelligence and to ensure data policies are transparent to build trust. Google's principles call for AI to be beneficial, fair and accountable, while the World Economic Forum's ethical guidelines emphasize fairness, data protection and opposition to autonomous weaponry. These guidelines collectively aim to ensure that AI systems are socially responsible and uphold human values (Rossi, 2018).

Data transparency and integrity can be maintained by documenting AI operations thoroughly and making AI model explanations accessible in non-technical language. Companies should only collect data necessary for AI functions to reduce privacy risks and establish mechanisms for employees to report ethical concerns. Recording decisions related to trade-offs in AI development ensures accountability (Eitel-Porter, 2021).

Implementing audit trails, conducting bias testing and documenting training datasets and testing histories are crucial steps to enhance transparency and address potential biases. Clear metrics should guide the development and operation of AI systems to ensure they are aligned with ethical principles (Shneiderman, 2020).

AI in Health Care

In health care, data transparency and integrity are essential for patient safety and outcomes improvement. Equitable distribution of AI-driven technologies is particularly pressing in public health, where AI's implementation could exacerbate existing health disparities if not handled responsibly (WHO, 2021). Establishing data governance panels to review AI training datasets ensures that the data is representative and sufficient. Regular audits should be conducted to assess bias, accuracy and predictability, with a focus on setting clear clinical objectives for AI applications. The use of public datasets is recommended to minimize privacy breaches, and professional bodies should issue guidelines on where AI can be used in diagnosis and treatment (Reddy et al., 2020).

Transparency in AI decision-making processes also is critical for maintaining equity, as the opaque nature of algorithmic decision-making poses risks to equitable health care delivery and governance. Research underscores that health care AI systems must prioritize procedural and

distributive justice to prevent discrimination and health inequities, ensuring fair access to AI-driven resources (Reddy et al., 2020). Transparency in AI applications should also involve educating patients about how their data is utilized to build trust in AI technologies. An inclusive approach to AI development helps address biases and improve AI tool accuracy and fairness. Regular assessment of AI-driven health care projects is recommended to maintain data integrity and ensure alignment with clinical standards (Murphy, Di Ruggiero & Upshur, 2021; Hamet & Tremblay, 2017).

The principles of explainable AI (XAI) are recommended for health care, as they allow users to understand AI decisions while avoiding explanations that reduce human-AI collaboration accuracy. It is essential that AI systems maintain data transparency and integrity through robust testing and conformance checks (Schmidt, Biessmann & Teubner, 2020).

Regulatory and Legal Considerations

Ensuring transparency in AI development is a key aspect of regulatory compliance. Regulations such as the EU AI Act require AI developers to disclose algorithms, data sources and development processes, which are necessary for effective human oversight. Sharing empirical testing results publicly is recommended to enhance accountability and transparency (Laux, 2023).

Accurate documentation of AI systems' capabilities, as required by regulatory frameworks, and regular monitoring of AI agents can help prevent risks associated with defective products. The EU AI Act and other regulatory bodies have established obligations to ensure that AI systems provide accurate risk assessments and that their intended uses are well-documented (Bletchley Declaration, 2023).

Ethical impact assessments should be conducted prior to AI deployment, especially in public health. Disclosing data sources, development methodologies and performance metrics fosters informed decision-making and builds trust in AI-driven public health initiatives (Chitramala, 2024).

Addressing Bias and Ensuring Fairness

AI governance should include establishing audit systems to assess bias in AI implementations, particularly in sectors such as public health surveillance. Guidelines to address algorithmic bias are essential to ensure fairness, and efforts should be made to include diverse populations in data collection to mitigate under-representation (Flores, Kim & Young, 2023).

Ensuring transparency in AI models used in medicine is vital for independent evaluations and external validation. Open disclosure of AI development processes helps validate AI tools and ensure their reproducibility (Crossnohere et al., 2022). Regular conformance testing and robust monitoring mechanisms can help identify biases and maintain data integrity (Methnani et al., 2021; Bodó & Janssen, 2022).

Data Sharing and Open Collaboration

Promoting open collaboration and data sharing in AI development can facilitate localization and adaptation to various contexts while maintaining transparency. Open-source AI components, including datasets and frameworks, allow for customization and adaptation. Transferring ownership and using open data standards ensure that AI systems remain accessible and adaptable (Fournier-Tombs, 2023).

Scalable data-sharing initiatives can accelerate the adoption of AI across industries by avoiding fragmentation. Standardizing global data interoperability practices can enable effective decision-making across different levels of governance (Shaheen, 2021).

Bibliography

1. Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1), 153−169. https://doi.org/10.1093/jopart/muac007
2. Biersmith, L., & Laplante, P. (2022). Introduction to AI Assurance for Policy Makers. 2022 IEEE 29th Annual Software Technology Conference (STC), 51–56. https://doi.org/10.1109/STC55697.2022.00016
3. Bletchley Declaration. (2023). The Bletchley Declaration by countries attending the AI safety summit, 1−2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
4. Bodó, B., & Janssen, H. (2022). Maintaining trust in a technologized public sector. *Policy and Society, 41*(3), 414−429. https://doi.org/10.1093/polsoc/puac019
5. Crossnohere, N. L., Elsaid, M., Paskett, J., Bose-Brill, S., & Bridges, J. F. P. (2022). Guidelines for artificial intelligence in medicine: Literature review and content analysis of frameworks. *Journal of Medical Internet Research, 24*(8), e36823. https://doi.org/10.2196/36823

6. Data Governance Institute. (n.d.). Data governance definition. https://datagovernance.com/the-data-governance-basics/definitions-of-data-governance/
7. Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics, 1*, 73−80. https://doi.org/10.1007/s43681-020-00011-6
8. Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681. https://doi.org/10.1016/j.clsr.2022.105681
9. Kasula, B. (2021). Ethical and Regulatory Considerations In AI-Driven Healthcare Solutions. *International Meridian Journal, 3*(3), 1-8. https://meridianjournal.in/index.php/IMJ/article/view/23
10. Felzmann, H., Fosch Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1). https://doi.org/10.1177/2053951719860542
11. Fisher, S., & Rosella, L. (2022). Priorities for successful use of artificial intelligence by public health organizations. *BMC Public Health, 22*(1), 2146. https://doi.org/10.1186/s12889-022-14422-z
12. Flores, L., Kim, S., & Young, S. D. (2024). Addressing bias in artificial intelligence for public health surveillance. Journal of Medical Ethics, 50(3), 190–194. https://doi.org/10.1136/jme-2022-108875
13. Fournier-Tombs, E. (2023). Local transplantation, adaptation, and creation of AI models for public health policy. *Frontiers in Artificial Intelligence, 6*, 1085671. https://doi.org/10.3389/frai.2023.1085671
14. Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism, 69(Supplement)*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011
15. Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, 1(1), 41-47. https://doi.org/10.1007/s43681-020-00008-1
16. Hu, S., & Li, Y. (2024). Policy interventions and regulations on generative artificial intelligence: Key gaps and core challenges. In *25th Annual International Conference on Digital Government Research* (DGO 2024). https://doi.org/10.1145/3657054.3659122
17. Khan, A.A., Badshah, S., Liang, P., Khan, B., Ahmad, A., Fahmideh, M., Niazi, M., & Akbar, A. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges, 383-392. *International Conference on Evaluation and Assessment in*

Software Engineering. https://www.researchgate.net/publication/361385839_Ethics_of_AI_A_Systematic_Literature_Review_of_Principles_and_Challenges

18. Krafft, P. M., Young, M., Huang, K., Katell, M., & Bugingo, G. (2020). Defining AI in policy versus practice. *AIES '20: Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 73–79. https://doi.org/10.1145/3375627.3375835
19. Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a democratic design of AI governance under the European Union AI Act. Oxford Internet Institute. https://link.springer.com/article/10.1007/s00146-023-01777-z
20. Leimanis, A., & Palkova, K. (2021). Ethical Guidelines for Artificial Intelligence in Healthcare from the Sustainable Development Perspective. *European Journal of Sustainable Development*, 10(1), 90−102. https://doi.org/10.14207/ejsd.2021.v10n1p90
21. Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4, Article 737072. https://doi.org/10.3389/frai.2021.737072
22. Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics,* 22, Article 14. https://doi.org/10.1186/s12910-021-00577-8
23. Qin, H., & Li, Z. (2024). A study on enhancing government efficiency and public trust: The transformative role of artificial intelligence and large language models. *International Journal of Engineering and Management Research*. 14(3). https://doi.org/10.5281/zenodo.12619360
24. Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association*, 27(3), 491–497. https://doi.org/10.1093/jamia/ocz192
25. Renda, A. (2019). Artificial intelligence: Ethics, Governance and Policy Challenges. Centre for European Policy Studies (CEPS) Task Force Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3420810
26. Robles, P., & Mallinson, D. J. (2023). Artificial intelligence technology, public trust, and effective governance. *Review of Policy Research, 40*, 1−18. https://doi.org/10.1111/ropr.12555
27. Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs, 72*(1), 127−134. https://www.jstor.org/stable/10.2307/26588348

28. Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29*(4), 260−278. https://doi.org/10.1080/12460125.2020.1819094
29. Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021). From ethical AI principles to governed AI. In *Proceedings of the Forty-Second International Conference on Information Systems* (ICIS 2021), Austin, TX. https://www.researchgate.net/publication/358234837_From_Ethical_AI_Principles_to_Governed_AI
30. Shaheen, Y. M. (2021). Applications of artificial intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1
31. Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems, 10*(4), Article 26. https://doi.org/10.1145/3419764
32. Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. *Science and Engineering Ethics, 27*(15). https://doi.org/10.1007/s11948-020-00277-3
33. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial Intelligence and Life in 2030: One hundred year study on artificial intelligence,* Stanford University. https://ai100.stanford.edu/sites/g/files/sbiybj18871/files/media/file/ai100report10032016fnl_singles.pdf
34. Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society, 40*(2), 137−157. https://doi.org/10.1080/14494035.2021.1928377
35. Verma, A., Rao, K., Eluri, V., & Sharma, Y. (2020). Regulating AI in public health: Systems challenges and perspectives. *Observer Research Foundation Occasional Paper, 261*. https://www.orfonline.org/public/uploads/posts/pdf/20230719010608.pdf
36. Winfield, A. F., & Michael, K., Pitt J., & Evers V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE, 107*(3), 1-14. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8662743
37. World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health: WHO guidance.* https://www.who.int/publications/i/item/9789240029200
38. Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(1), Article 100455. https://doi.org/10.1016/j.patter.2022.100455

Research Question 5: What role do AI policies assign to human oversight and intervention in automated decision-making processes?

Research Questions Examined in Articles

The literature extensively examines how AI systems can be designed with human-centered oversight, focusing on policies and frameworks that ensure accountability, safety and equity. Studies investigate mechanisms that prioritize human involvement at key stages of AI decision-making, aiming to maintain ethical standards and user trust in sensitive sectors like health care and criminal justice. Research questions emphasize how human oversight can prevent unintended consequences and enhance AI's role as a supportive, rather than autonomous, tool.

Regulation of Human Intervention in AI Decision-Making

The research highlights how existing AI policies, such as the GDPR and the proposed EU AI Act, regulate human intervention within AI-driven decision-making processes. These studies explore the practical implementation of human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC) models across various sectors. The literature assesses how these regulatory frameworks are structured to enforce human accountability, ensuring that AI systems remain aligned with safety, fairness and legal compliance.

Addressing Biases in Human Interaction with AI

The literature also explores how biases, such as automation bias or selective adherence to AI recommendations, impact human interactions with AI systems. Research questions focus on the potential of these biases to skew decision-making, particularly in public sector applications, and examine how oversight mechanisms can be designed to minimize such biases, thereby promoting fairness and accuracy.

Challenges and Limitations of Human Oversight Mechanisms

The research also identifies the challenges and limitations of human oversight in AI systems, particularly concerning issues of accountability, responsibility and preventing discrimination. Studies raise questions about the adequacy of current oversight models in mitigating AI's risks while ensuring ethical decision-making, underscoring the need for more robust, transparent and inclusive policies.

Summary of Key Findings – Human Oversight and Intervention

Human oversight is a cornerstone of global AI governance frameworks, serving as a safeguard to ensure accountability, mitigate risks and foster public trust in AI technologies. By embedding mechanisms for human intervention and ensuring meaningful oversight, these frameworks aim to align AI operations with societal values and ethical standards. Several key findings emerge from the literature concerning the role of human oversight in AI systems:

Human Oversight as a Core Principle

Human oversight is a foundational element in many global AI governance frameworks. Documents such as the Organization for Economic Cooperation and Development (OECD) *AI Principles* and the EU AI Act emphasize human oversight as essential to maintaining accountability, preventing harm and ensuring public trust in AI technologies (Cihon, 2024; Laux, 2023; Shneiderman, 2020). Human involvement in AI governance is generally presented as a safeguard against the potential risks posed by AI, such as errors or biases, and is often linked to notions of transparency and accountability (Shneiderman, 2020; Fukuda-Parr & Gibbons, 2021; Hickok, 2021).

Mechanisms for Human Intervention

Across the literature, human intervention is implemented through various oversight mechanisms, including human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC) models. These approaches allow humans to intervene in AI decision-making at different stages, ensuring that AI does not operate unchecked, especially in critical applications like health care and law enforcement (Methnani et al., 2021; Shneiderman, 2020; Green, 2022).

Bias and Automation Challenges

Human oversight also is linked to the mitigation of biases within AI systems. Studies show that without proper oversight, AI systems can exacerbate pre-existing social biases, particularly in high-stakes sectors such as criminal justice and public services (Green, 2022; Fukuda-Parr & Gibbons, 2021; Hickok, 2021). However, there also are concerns about the efficacy of human oversight in mitigating these biases. For example, human decision-makers may exhibit selective adherence to AI recommendations when these align with their own biases, potentially compounding the problem rather than resolving it (Alon-Barkat & Busuioc, 2023; Green, 2022; Sele & Chugunova, 2024).

Accountability and Transparency

Several articles emphasize the need for transparency in AI systems to support effective human oversight. Mechanisms such as audit trails and "ethical black boxes" are proposed to track and analyze the decision-making processes of AI systems, thereby enabling humans to review, correct or contest AI outputs. An ethical black box refers to a mechanism or system integrated into artificial intelligence (AI) technologies that records and stores data about the AI's decision-making processes, interactions, and behaviors. Its purpose is to create transparency, accountability, and traceability in the functioning of AI systems, much like a flight recorder or "black box" in aviation. This is particularly important in ensuring that human actors can be held accountable for the operation and outcomes of AI systems (Winfield et al., 2019; Shneiderman, 2020; Methnani et al., 2021).

Human Oversight in Policy

Although human oversight is frequently cited as necessary in AI policies, there is criticism regarding its actual implementation. Some policies rely on superficial forms of human oversight, such as rubber-stamping algorithmic decisions, without providing meaningful opportunities for human intervention or correction. This has led to concerns that such policies may create a false sense of security, legitimizing flawed AI systems without addressing their underlying issues (Green, 2022; Laux, 2023; Cihon, 2024).

Challenges

The implementation of effective human oversight in AI systems faces several challenges:

Complexity of AI Systems:

As AI systems become increasingly complex, ensuring consistent and effective human oversight is a significant challenge. AI systems often evolve dynamically, creating difficulties in monitoring and adjusting their behavior, particularly in real-time (Shneiderman, 2020; Laux, 2023). Moreover, the sheer volume of data generated by AI systems complicates oversight efforts, raising concerns about scalability and the cost of continuous monitoring (Shneiderman, 2020; Methnani et al., 2021).

Automation Bias

Automation bias, where humans defer to AI recommendations even when they are flawed, is a persistent challenge. Research indicates that human decision-makers tend to trust algorithmic outputs more than human-generated advice, even in cases where the algorithm is incorrect (Green, 2022; Alon-Barkat & Busuioc, 2023). This over-reliance on AI undermines the effectiveness of human oversight and can lead to poor decision-making outcomes (Green, 2022; Sele & Chugunova, 2024).

Vagueness in Policy Guidelines

Policies such as the EU AI Act provide for human oversight but often lack specific guidance on how this oversight should be implemented. The absence of clear standards for the roles and responsibilities of human overseers can lead to confusion and inconsistency in oversight practices (Domingo, 2022; Laux, 2023).

Bias and Inequity

AI systems are prone to perpetuating systemic biases, particularly against marginalized communities. Without robust oversight mechanisms, these biases can become entrenched in AI decision-making processes, exacerbating social inequities (Fukuda-Parr & Gibbons, 2021; Hickok, 2021). The lack of diversity among those tasked with overseeing AI systems further compounds this problem, as decision-making processes may reflect the biases and priorities of a narrow group of stakeholders (Hickok, 2021; Methnani et al., 2021).

Superficial Human Oversight

In many cases, human oversight is more symbolic than substantive. The practice of rubber-stamping AI decisions is highlighted as a major flaw in current oversight policies. This superficial

involvement fails to provide the safeguards necessary to prevent errors or biases from influencing outcomes (Green, 2022; Laux, 2023).

Recommendations

To address the challenges identified, several recommendations emerge from the literature:

Stronger Regulatory Frameworks

AI policies should include more detailed and enforceable regulations regarding the level and extent of human oversight required for AI systems. This includes clear definitions of roles and responsibilities for human overseers and the development of standards for auditing and reviewing AI systems (Domingo, 2022; Laux, 2023; Methnani et al., 2021)

Increased Transparency and Accountability

Transparent decision-making processes, supported by audit trails and explainable AI systems, are critical for ensuring that human oversight remains meaningful. The use of "ethical black boxes" and similar mechanisms can help trace AI decisions and hold the relevant human actors accountable (Winfield et al., 2019; Shneiderman, 2020; Methnani et al., 2021). An ethical black box refers to a mechanism or system integrated into artificial intelligence (AI) technologies that records and stores data about the AI's decision-making processes, interactions, and behaviors

Diverse and Inclusive Oversight

Policymakers should prioritize the inclusion of diverse voices in AI oversight, particularly those from marginalized communities who are most affected by AI systems. This can help ensure that oversight processes are equitable and reflective of the broader public interest (Fukuda-Parr & Gibbons, 2021; Hickok, 2021; Laux, 2023).

Training and Competency Development

Human overseers must be adequately trained and knowledgeable about the AI systems they are monitoring. This includes understanding the limitations and risks associated with AI, as well as the ability to intervene effectively when necessary (Laux, 2023; Methnani et al., 2021; Shneiderman, 2020).

Combining Human and AI Oversight

In some cases, human oversight may need to be augmented by AI tools to address the complexity of modern AI systems. Hybrid oversight systems, where AI assists humans in monitoring and correcting AI outputs, may provide a more effective approach to maintaining control over automated decision-making processes (Methnani et al., 2021; Winfield et al., 2019).

Bibliography

1. Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1), 153−169. https://doi.org/10.1093/jopart/muac007
2. Cihon, P. (2024). *Chilling Autonomy: Policy enforcement for human oversight of AI agents*. Workshop on Generative AI and Law (GenLaw '24) at the 41st International Conference on Machine Learning. https://blog.genlaw.org/pdfs/genlaw_icml2024/79.pdf
3. Domingo, S. (2022). Human Intervention and Human Oversight in the GDPR and AI Act. *Trilateral Research.* https://trilateralresearch.com/emerging-technology/human-intervetion-in-gdpr-and-ai
4. Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy, 12*(S6), 34−43. https://doi.org/10.1111/1758-5899.12965
5. Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 105681. https://www.sciencedirect.com/science/article/pii/S0267364922000292
6. Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics, 1*(1), 41-47. https://doi.org/10.1007/s43681-020-00008-1
7. Laux, J. (2023). Institutionalized Distrust and Human Oversight of Artificial Intelligence: Toward a democratic design of AI governance under the European Union AI Act. Oxford Internet Institute. https://link.springer.com/article/10.1007/s00146-023-01777-z
8. Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4, Article 737072. https://doi.org/10.3389/frai.2021.737072

9. Sele, D., & Chugunova, M. (2024). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *PLOS ONE, 19*(2), e0298037. https://doi.org/10.1371/journal.pone.0298037
10. Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems, 10(4), Article 26. https://doi.org/10.1145/3419764
11. Winfield, A. F., & Michael, K., Pitt J., & Evers V. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE, 107*(3), 1-14. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8662743

Research Question 6: What are the equity and ethical considerations of AI that should be addressed in policies?

Research Questions Examined in Articles

Across the reviewed literature, the focus is consistently placed on exploring how artificial intelligence (AI) intersects with health equity and ethics, particularly in public health, workplace environments and machine learning (ML) systems. Common themes include the role of AI in either mitigating or perpetuating health disparities, the need for ethical guidelines to prevent algorithmic bias, and strategies to promote fairness and transparency in AI systems. Several authors advocate for stronger human rights-based frameworks to ground AI ethics in international law.

Equity and AI Development

A significant theme involves how AI systems either mitigate or exacerbate health disparities, particularly for racial minorities and low-income groups. Throughout the review, a consistent emphasis was placed on ensuring that AI systems do not exacerbate existing health disparities but instead promote fairness and equitable outcomes for all populations.

Ethical Guidelines and Fairness

Many articles examine the ethical implications of AI, including transparency, accountability, fairness and the prevention of algorithmic bias (Thomasian et al., 2021; Cachat-Rosset & 2023). The discussions focus on creating clear ethical standards for AI systems to avoid perpetuating inequalities.

AI's Role in Public Health

Several works highlight how AI is currently being integrated into public health initiatives, stressing the need for equitable deployment to avoid widening health disparities (Smith et al., 2020AI's potential to transform public health, especially in diagnostics and personalized care, is contrasted with the risks posed by data bias and inequitable access.

Summary of Key Findings – Equity and Ethics

AI technologies in public health and health care have the potential to transform service delivery, but poorly designed systems risk exacerbating existing health disparities. Biases in data collection, model design and deployment often disadvantage marginalized communities, particularly racial minorities and low-income populations. Addressing these risks requires proactive strategies, including transparent accountability mechanisms, inclusive development practices and targeted investments in digital infrastructure to bridge the digital divide.

Health Equity Risks

Many AI systems used in public health and health care risk exacerbating existing health disparities if not carefully designed and implemented. Biases in data collection, model specification and deployment can significantly disadvantage marginalized communities, particularly racial minorities and low-income populations (Dankwa-Mullan et al., 2021; Thais et al., 2023). For instance, health care AI tools often fail to account for diverse demographic needs, leading to skewed health outcomes for underserved populations (Smith et al., 2020).

To address these risks, several articles recommend frameworks that explicitly prioritize health equity and racial justice at every stage of the AI lifecycle. Dankwa-Mullan et al. (2021) proposed a structured framework that integrates health equity and racial justice principles into the development, deployment and governance of AI systems. This approach ensures that AI models are not just optimized for technical accuracy but also for promoting equitable health outcomes across different population groups.

Bias and Accountability

Algorithmic bias is a recurring theme across all reviewed articles, with a focus on the need for accountability mechanisms to ensure fairness in AI models. Bias can manifest at multiple stages, from data collection to model deployment, making it critical to establish safeguards that address these biases proactively (Baum, 2023; Hendricks-Sturrup et al., 2023). Several authors

emphasized the importance of regular bias audits and equity-sensitive metrics to monitor AI systems' performance and prevent discriminatory outcomes.

In the health care context, it is particularly important to ensure that AI systems are transparent and accountable, given the high-stakes nature of health-related decision-making. Ethical frameworks must prioritize fairness and require that AI decisions be interpretable by both developers and end-users. Regular audits, stakeholder engagement and interdisciplinary collaboration are essential components of a robust accountability system that aligns AI systems with health equity goals (Thomasian et al., 2021).

Ethical Governance

Ethical concerns such as transparency, fairness and accountability are widely discussed across the literature. Many ethical guidelines for AI development are currently voluntary, and the lack of enforceable standards poses a significant risk of corporate-driven decision-making that may not prioritize equity (Fukuda-Parr & Gibbons, 2021). Human oversight is seen as a critical component in preventing harm and ensuring that AI systems are used responsibly in public health settings.

Inclusion and Community Engagement

Inclusivity in AI development is critical to ensuring that AI systems reflect the needs of diverse populations. Co-created solutions involving underrepresented groups and community-led decision-making are essential for fostering trust and mitigating the risks of biased AI systems (Hendricks-Sturrup et al., 2023; Simmons et al., 2023). Engaging communities from the initial stages of AI development can help identify potential biases early on and ensure that the resulting models are designed with diverse perspectives in mind, thereby reducing bias and increasing fairness.

Transparency and Human Oversight

Many studies emphasize that AI systems, particularly those used in decision-making contexts such as health care and public health, should remain transparent and maintain human oversight to prevent harm. Ethical frameworks must prioritize fairness, ensure that AI decisions are interpretable by users and involve regular bias audits (Baum, 2023; Thomasian et al., 2021).

Digital Divide and Inequitable Access

The literature highlights the risk of the digital divide exacerbating health disparities. Populations in low-income areas may lack access to the digital infrastructure required to benefit from AI innovations in health care and public health, leading to further inequities (Smith et al., 2020; Kayode, 2024). Addressing this challenge requires targeted investment in digital infrastructure, particularly in rural and underserved urban areas.

Government agencies and public health institutions must prioritize closing the digital divide to ensure equitable access to AI-driven health care solutions. Policies should aim to provide equitable access to AI tools and ensure that underrepresented populations can benefit from AI innovations in public health. This includes investing in digital infrastructure, enhancing digital literacy and creating alternative access points for AI-driven health services.

Challenges

The literature identifies several key challenges to integrating equity and ethical principles into AI policies:

Algorithmic Bias

A recurring challenge is the presence of bias in AI models, which can arise from non-representative data, problem specification and deployment contexts. These biases can have far-reaching consequences, particularly for racial minorities, women and low-income groups. AI systems can unintentionally perpetuate discrimination in public health interventions, health care diagnostics and workplace hiring practices (Baum, 2023; Gichoya et al., 2021).

Lack of Transparency and Accountability

Many AI systems operate as "black boxes," meaning their decision-making processes are unclear, difficult to interpret, or audit. This lack of transparency can lead to mistrust, particularly in health care and public health applications where human well-being is at stake (Fukuda-Parr & Gibbons, 2021). Without transparent systems and accountability mechanisms, AI systems can make decisions that disproportionately harm marginalized communities.

Regulatory Gaps

The rapid advancement of AI technologies has outpaced existing legal frameworks, leaving significant regulatory gaps. In health care and public health, regulations such as HIPAA and

GDPR do not fully address the ethical risks posed by AI, including algorithmic bias and data privacy concerns (Thais et al., 2023). This lack of cohesive regulation leaves AI systems vulnerable to exploitation and misuse.

Data Quality and Infrastructure

AI systems rely on high-quality data to function effectively, but data quality issues are common, particularly in low-income areas with limited digital infrastructure (Smith et al., 2020). The digital divide continues to present a challenge for equitable AI deployment, with certain populations being disproportionately affected by the lack of access to AI-driven health care solutions (Bharel et al., 2024).

Trust and Resistance

Many stakeholders, including health care professionals and the public, remain skeptical of AI systems, particularly when AI-generated decisions conflict with human judgment. Resistance to adopting AI is often rooted in fears of technology replacing human workers, as well as concerns about AI's ability to provide fair and accurate outcomes (Thomasian et al., 2021; Kanter et al., 2023).

Recommendations

To address the equity and ethical challenges in AI, the literature offers several policy recommendations:

Comprehensive Regulatory Frameworks

Strong regulatory frameworks are essential to govern AI systems, particularly in health care and public health contexts. These frameworks should address algorithmic bias, ensure transparency and provide accountability mechanisms for AI-driven decisions. Regulatory bodies should consider establishing AI-specific guidelines, rather than relying on existing frameworks that do not fully address AI's unique ethical risks (Fukuda-Parr & Gibbons, 2021). International collaboration is necessary to create cohesive regulations that prevent the exploitation of AI technologies.

Bias Mitigation and Equity Audits

AI developers and public health organizations should prioritize equity by embedding bias mitigation strategies into the AI lifecycle. Bias audits should be conducted regularly to ensure

that AI models do not perpetuate systemic inequities (Thomasian et al., 2021). Additionally, equity-sensitive metrics should be developed to evaluate AI systems' fairness, focusing on reducing disparities in health care outcomes. Organizations can achieve this by using diverse datasets, federated learning and ongoing audits (Dankwa-Mullan et al., 2021; Baum, 2023).

Inclusive Data Practices and Community Engagement

AI systems should be developed using inclusive data practices, ensuring that data sets are representative of diverse populations. To achieve this, AI developers should engage with marginalized communities through co-created solutions, ensuring that AI systems reflect the needs and perspectives of those most at risk of being overlooked (Hendricks-Sturrup et al., 2023). Authentic and ongoing engagement with these communities is critical for building trust and ensuring that AI systems promote equitable outcomes (Smith et al., 2020).

Human Oversight and Transparency

Transparency and governance models in organizations should be a foundational principle in AI systems, particularly in health care and public health. AI models must be interpretable by users, and human oversight should be maintained to prevent AI from making harmful or biased decisions (Baum, 2023). In addition to transparency, fairness must be integrated into AI governance models, with interdisciplinary collaborations between AI developers, ethicists and health care professionals (Berdahl et al., 2023).

Addressing the Digital Divide

Governments and public health institutions must prioritize closing the digital divide to ensure equitable access to AI-driven health care solutions. Investment in digital infrastructure, particularly in low-income areas, is essential for preventing AI systems from worsening existing health disparities (Smith et al., 2020). Policies should aim to provide equitable access to AI tools and ensure that underrepresented populations can benefit from AI innovations in public health.

Ethical AI in Nonprofit and Workplace Environments

In nonprofit organizations and workplaces, AI should augment, not replace, human work. Nonprofits should adopt human-centered AI guidelines and ensure that AI tools are used responsibly, with small-scale pilots recommended before broader implementation (Kanter et al., 2023). In workplaces, AI tools should be designed to support diversity, equity and inclusion

(DEI) principles, with regular bias audits and cross-functional collaboration among technical, legal and HR teams (Baum, 2023).

Interdisciplinary Collaboration

AI systems require collaboration between diverse fields, including AI development, public health, law and ethics. Successful implementation of ethical AI systems depends on interdisciplinary approaches that align AI technologies with public health goals, while ensuring that equity and ethical standards are prioritized (Berdahl et al., 2023). This collaboration should also extend to global partnerships to ensure that AI systems benefit all populations.

Ongoing Monitoring and Ethical Learning

AI systems, particularly those deployed in public health, should undergo continuous monitoring to evaluate their impact on health outcomes and ensure that emerging ethical issues are addressed (Simmons et al., 2023). Continuous improvement of ethical and legal frameworks is necessary to respond to new challenges posed by AI technologies (Gichoya et al., 2021). This iterative process will help ensure that AI systems evolve to better serve marginalized populations and promote equity.

Bibliography

1. Baum, B. (2023). AI challenges in the workplace: Are artificial intelligence policies meeting DEI thresholds? *Journal of Business and Behavioral Sciences, 3*5(3), 3−13. https://asbbs.org/files/2023-24/JBBS_35.3_Fall_2023.pdf
2. Berdahl, C. T., Baker, L., Mann, S., Osoba, O., & Girosi, F. (2023). Strategies to improve the impact of artificial intelligence on health equity: Scoping review. *JMIR AI*, 2, e42936. https://doi.org/10.2196/42936
3. Bharel, M., Auerbach, J., Nguyen, V., & DeSalvo, K. B. (2024). Transforming public health practice with generative artificial intelligence. *Health Affairs, 43*(6), 776−782. https://doi.org/10.1377/hlthaff.2024.00050
4. Birhane, A., Ruane, E., Laurent, T., Brown, M. S., Flowers, J., Ventresque, A., & Dancy, C. L. (2022). The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 948-958). ACM. https://doi.org/10.1145/3531146.3533157

5. Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. *Applied Artificial Intelligence, 37*(1), e2176618. https://doi.org/10.1080/08839514.2023.2176618
6. Dankwa-Mullan, I. (2024). Health equity and ethical considerations in using artificial intelligence in public health and medicine. *Preventing Chronic Disease, 21*, E64. https://doi.org/10.5888/pcd21.240245
7. Dankwa-Mullan, I., Scheufele, E. L., Matheny, M. E., Quintana, Y., Chapman, W. W., Jackson, G., & South, B. R. (2021). A proposed framework on integrating health equity and racial justice into the AI development lifecycle. *Journal of Health Care for the Poor and Underserved, 32*(2), 300−317. https://doi.org/10.1353/hpu.2021.0065
8. Fukuda-Parr, S., & Gibbons, E. (2021). Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines. *Global Policy, 12*(S6), 32–41. https://doi.org/10.1111/1758-5899.12965
9. Gichoya, J. W., McCoy, L. G., Celi, L. A., & Ghassemi, M. (2021). Equity in essence: A call for operationalizing fairness in machine learning for healthcare. *BMJ Health Care Inform, 28*(e100289). https://doi.org/10.1136/bmjhci-2020-100289
10. Hendricks-Sturrup, R., Simmons, M., Anders, S., Aneni, K., Clayton, E. W., Coco, J., Collins, B., Heitman, E., Hussain, S., Joshi, K., Lemieux, J., Novak, L. L., Rubin, D. J., Shanker, A., Washington, T., Waters, G., Harris, J. W., Wagner, T., Yin, R., Yin, Z., & Malin, B. (2023). Developing ethics and equity principles, terms, and engagement tools to advance health equity and researcher diversity in AI and machine learning: Modified Delphi approach. *JMIR AI*, 2, e52888. https://doi.org/10.2196/52888
11. Kanter, B., Fine, A., & Deng, P. (2023, September 7). 8 steps nonprofits can take to adopt AI responsibly. *Stanford Social Innovation Review.* https://ssir.org/articles/entry/8_steps_nonprofits_can_take_to_adopt_ai_responsibly
12. Simmons, M., Hendricks-Sturrup, R., Waters, G., Novak, L., Were, M., & Hussain, S. (2023). An Expert Panel Discussion: Embedding ethics and equity in artificial intelligence and machine learning infrastructure. *Big Data and Society, 11*(S1), S1-S13. https://doi.org/10.1089/big.2023.29061.rtd
13. Smith, M. J., Axler, R., Bean, S., Rudzicz, F., & Shaw, J. (2020). Four equity considerations for the use of artificial intelligence in public health. Bulletin of the World Health Organization, 98(4), 290–292. https://pmc.ncbi.nlm.nih.gov/articles/PMC7133473/

14. Thais, S., Shumway, H., & Iglesias Saragih, A. (2023). Algorithmic bias: Looking beyond data bias to ensure algorithmic accountability and equity. *Journal of Artificial Intelligence Research, 58,* 59−64. https://sciencepolicyreview.org/2023/08/mitspr-191618004007/
15. Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy, 42,* 602−611. https://doi.org/10.1057/s41271-021-00319-5

Research Question 7: What are the impacts of AI on individuals with disabilities and how should these issues be addressed?

Research Questions Examined in Articles

The literature focuses on how AI intersects with the lives of individuals with disabilities, addressing several critical research questions:

Fair Treatment for Individuals with Disabilities

AI systems have the potential to either perpetuate or mitigate societal inequalities for individuals with disabilities. A significant concern is the presence of biases encoded within AI systems, which can fail to accommodate the diverse needs of disabled individuals (Trewin, 2018). Inclusive design is emphasized as a critical strategy for ensuring fair treatment and addressing the risks of reinforcing existing biases (Trewin, 2018).

Risks and Opportunities of AI Across Sectors

AI presents both opportunities and risks for individuals with disabilities in sectors such as employment, education, health care and public safety. Depending on how design flaws are addressed, AI can either empower or marginalize these individuals (Trewin et al., 2019). Evaluating AI's impact across these contexts is essential to promote meaningful inclusion and to ensure that benefits are equitably distributed (Trewin et al., 2019).

Impact of Disability Models on AI Development

The influence of different models of disability — medical, social and relational — on AI development has significant implications for its outcomes. These conceptual frameworks shape AI design and deployment and can lead to biases if not critically examined (Newman-Griffis et al., 2024). Integrating a comprehensive understanding of disability into AI models is necessary to prevent the perpetuation of inequities (Newman-Griffis et al., 2024).

Ethical Considerations in AI for Disabled Individuals

The ethical use of AI for individuals with disabilities requires a shift from a basic fairness approach to a justice-oriented framework. Addressing power dynamics and structural inequalities in AI ethics is crucial to developing a more nuanced, equity-driven ethical model that guides AI applications (Bennett & Keyes, 2019). This perspective emphasizes the importance of prioritizing justice and equity to ensure AI systems effectively support disabled individuals (Bennett & Keyes, 2019).

Summary of Key Findings – Impacts to Individuals With Disabilities

AI technologies can reflect and reinforce societal biases, particularly against individuals with disabilities. These biases often stem from datasets that lack diverse input, resulting in AI systems that fail to accurately represent the needs of disabled populations. Across domains such as employment, education and health care, AI systems can either empower or disadvantage individuals with disabilities, depending on their design and implementation. Addressing these issues requires inclusive development processes, meaningful engagement with disabled communities and a shift from fairness-based to justice-oriented approaches to mitigate structural inequalities and promote equitable outcomes.

Bias in AI Systems

AI technologies often reflect and reinforce societal biases against individuals with disabilities. Many AI models fail to accurately represent disabled individuals due to their reliance on datasets lacking diverse input (Trewin, 2018). For instance, facial recognition software may struggle to identify individuals with disabilities or atypical features because the training data predominantly consists of images of able-bodied individuals (Trewin, 2018).

Impact on Different Domains

AI technologies can both enhance and hinder the lives of disabled individuals across various sectors, including employment and health care. In the employment sector, AI-driven recruitment tools may inadvertently exclude qualified candidates who require accommodations, as these systems often rely on rigid criteria that fail to consider the specific needs of disabled applicants (Trewin et al., 2019). For example, an AI tool analyzing resumes might undervalue the qualifications of a deaf candidate who uses an interpreter, misinterpreting communication styles as deficiencies (Trewin et al., 2019).

Employment

The introduction of AI in recruitment processes poses significant risks for disabled individuals. AI tools that analyze video interviews may misinterpret the body language or facial expressions of candidates with autism or other neurological conditions, leading to unjust assessments of their capabilities (Whittaker et al., 2019). Similarly, AI systems used in hiring may exclude qualified applicants with cognitive disabilities based on preconceived notions about productivity or work performance (Whittaker et al., 2019).

Education

In education, AI can offer personalized learning experiences but may exacerbate disparities if not carefully designed. For instance, educational AI systems may misjudge the capabilities of students with learning disabilities, leading to outcomes that fail to reflect their true potential (Guo et al., 2019). Time constraints implemented by AI-driven learning platforms can disadvantage students who require additional time due to processing speed variations, unfairly impacting their performance (Guo et al., 2019).

Health Care

AI applications in health care provide opportunities to improve access to care for individuals with disabilities but also present significant challenges. Diagnostic tools may lead to misdiagnoses if they fail to account for the diverse manifestations of disabilities (Bennett & Keyes, 2019). For example, machine learning models trained primarily on data from non-disabled individuals may struggle to recognize symptoms in patients with rare disabilities, resulting in inadequate or incorrect treatment (Bennett & Keyes, 2019). Additionally, biases in health care AI systems can lead to the misallocation of resources, as algorithms may prioritize treatments based on prevailing norms that do not address the unique health care needs of disabled individuals (Bennett & Keyes, 2019).

Engagement of Disabled Individuals in AI Development

Involving individuals with disabilities in the AI development process is critical to ensuring that their diverse needs are addressed. Engaging disabled individuals as active participants in designing and testing AI systems can result in more inclusive and effective solutions (Newman-Griffis et al., 2024). Collaboration between AI developers and disability advocacy organizations can provide valuable insights into the challenges faced by disabled individuals, enabling user-centered design approaches (Newman-Griffis et al., 2024).

Fairness vs. Justice

The discourse surrounding AI ethics often emphasizes fairness, but this focus can inadvertently reinforce existing power dynamics. A justice-oriented approach recognizes the broader structural inequalities affecting disabled individuals and addresses systemic barriers that prevent equitable access to AI technologies (Bennett & Keyes, 2019). This shift involves acknowledging the historical and social contexts in which these technologies operate to create more inclusive and equitable outcomes (Bennett & Keyes, 2019).

Challenges

The integration of AI technologies for individuals with disabilities faces several significant challenges:

Diversity of Disabilities

The wide spectrum of disabilities complicates the development of universally effective AI solutions. Each disability presents unique challenges that must be considered during the design process. For instance, the features of individuals with cerebral palsy may affect their mobility and communication styles, necessitating specialized input in AI systems designed to assist them (Whittaker et al., 2019).

Privacy Concerns

Privacy issues related to the collection of sensitive data about disabilities pose significant barriers to participation in AI systems. Individuals may be reluctant to disclose their disabilities due to concerns about confidentiality and potential discrimination (Trewin, 2018). Ensuring data protection and establishing trust are critical for encouraging disabled individuals to engage with AI technologies.

Underrepresentation in Data

Many AI systems are trained on datasets that lack representation of individuals with disabilities, leading to harmful misinterpretations of their needs and capabilities. For instance, an AI system that analyzes job applications might rely on data predominantly sourced from able-bodied individuals, which can skew hiring practices and perpetuate discrimination against disabled candidates (Guo et al., 2019).

Complexities of AI Systems

The "black box" nature of AI can obscure understanding and accountability, complicating efforts to mitigate biases. Users may find it difficult to challenge decisions made by AI systems, particularly when those decisions directly impact their lives, as seen in health care and employment contexts (Trewin et al., 2019).

Recommendations

To improve AI's impact on individuals with disabilities, the literature presents several key recommendations.

Promote Inclusive Design Practices

Engage individuals with disabilities in all stages of AI development to ensure their perspectives shape the technology. For example, co-design workshops that involve disabled individuals can help identify specific needs and challenges that AI systems should address (Trewin et al., 2019). Such inclusive practices could lead to AI applications that better serve the needs of disabled individuals, as evidenced by successful initiatives in accessible technology development.

Shift Ethical Frameworks

Transition from a fairness-centered approach to a justice-oriented framework that considers broader power dynamics and societal implications. This shift requires a commitment to addressing systemic inequalities and empowering disabled individuals within the AI development process (Bennett & Keyes, 2019).

Regular Bias Audits and Monitoring

Implement ongoing monitoring and bias testing throughout the AI development lifecycle. Tools like IBM's AI Fairness 360 Toolkit can be employed to evaluate the impacts of AI systems on diverse populations and identify areas for improvement (Whittaker et al., 2019). Regular audits can help ensure that AI technologies do not reinforce existing biases or inequities.

Develop Inclusive Datasets

Create datasets that accurately reflect the diversity of disabilities while addressing privacy concerns. This includes utilizing techniques such as federated learning, which allows data to be

used for training AI models without compromising individual privacy (Newman-Griffis et al., 2024). Collaborative efforts with disability advocacy organizations can facilitate the collection of diverse data that informs AI system design.

Bibliography

1. Bennett, C. L., & Keyes, O. (2019). *What is the Point of Fairness? Disability, AI, and the complexity of justice*. Proceedings of the ASSETS '19: The 21st International ACM SIGACCESS Conference on Computers and Accessibility. DOI: 10.1145/3308561.3353790. https://www.sigaccess.org/newsletter/2019-10/bennet.html
2. Guo, A., Kamar, E., Vaughan, J. W., Wallach, H., & Morris, M. R. (2019). Toward Fairness in AI for People with Disabilities: A research roadmap. ACM ASSETS 2019 Workshop on AI Fairness for People with Disabilities. https://dl.acm.org/doi/10.1145/3386296.3386298
3. Newman-Griffis, D., Rauchberg, J. S., Alharbi, R., Hickman, L., & Hochheiser, H. (2024). Definition drives design: Disability models and mechanisms of bias in AI technologies. *Journal of Disability Studies and AI Ethics. 28(102).* https://firstmonday.org/ojs/index.php/fm/article/view/12903
4. Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N., & Manser, E. (2019). Considerations for AI fairness for people with disabilities. *AI Matters, 5*(3), 40−53. https://doi.org/10.1145/3362077.3362086
5. Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Ringel Morris, M., Rankin, J., Rogers, E., Salas, M., & West, S. M. (2019). *Disability, bias, and AI.* AI Now Institute at NYU. https://ainowinstitute.org/publication/disabilitybiasai-2019

Research Question 8: What are the impacts of AI on older adults and how should these issues be addressed?

Research Questions Examined in Articles

The literature explores the specific benefits of AI technologies for older adults, focusing on their ability to enhance care, social connectivity and overall well-being. Studies investigate how AI tools, such as socially assistive robots and telehealth services, contribute to improved physical and mental health outcomes in various care settings. Socially assistive robots are AI-driven systems designed to provide support and companionship, focusing on enhancing physical and

mental well-being through social interaction and assistance in tasks, particularly in healthcare and caregiving settings.

Research questions emphasize how AI can facilitate personalized care, support daily activities and foster engagement with the care environment.

Psychological and Social Implications of AI for Older Adults

Research also examines the psychological and social implications of AI use among older adults. The studies highlight both positive outcomes, such as reduced loneliness and increased social interaction, and potential concerns, including over-reliance on AI and the need for emotional adaptability when interacting with AI tools. Questions address how AI affects social dynamics, engagement and the overall mental well-being of older adults, emphasizing the importance of evaluating AI's impact beyond physical care.

Integrating Ethical Considerations in AI Design for Older Adults

Ethical considerations, such as cultural competence and inclusivity, are central to research on AI design and implementation for older adults. The literature investigates how AI tools can be developed to respect cultural norms, individual preferences and varying levels of technological literacy among older populations. Studies focus on how ethical design can prevent biases, improve user experience and ensure that AI solutions are both accessible and beneficial to this demographic.

Summary of Key Findings – Impact on Older Adults

AI technologies offer significant potential to enhance the well-being of older adults by addressing emotional, social and accessibility challenges. From socially assistive robots that reduce loneliness to telehealth services that bridge social isolation, AI is fostering greater connectivity and psychological resilience in aging populations. However, ensuring equitable access remains a critical challenge, as many older adults face barriers such as inadequate internet infrastructure and limited technological literacy. Addressing these disparities is vital to fully realize AI’s potential to improve the quality of life for older individuals.

Enhancements in Well-being

The CARESSES Randomized Controlled Trial highlights that culturally competent socially assistive robots can significantly improve emotional well-being among older adults in care

settings. Participants reported decreased feelings of loneliness and enhanced engagement with their care environment, suggesting that tailored interactions can foster psychological resilience (Papadopoulos et al., 2022).

Facilitating Social Connectivity

AI technologies, including telehealth services and virtual companionship platforms, can bridge the gap of social isolation commonly experienced by older adults. For example, studies have demonstrated that video conferencing tools enable older adults to maintain family connections, which are crucial for mental health and social engagement (Chu et al., 2022).

Equitable Access

The literature emphasizes that for AI to be effective for older populations, equitable access must be prioritized. Many older adults, particularly in rural areas, face barriers in utilizing telehealth services due to inadequate internet access and technological proficiency (Rubeis et al., 2022). Addressing these gaps is essential for leveraging the full potential of AI in enhancing the quality of life for older adults.

Challenges

The reviewed literature identifies several challenges in addressing AI's negative impacts on older adults.

Algorithmic Bias

AI systems often rely on training datasets that inadequately represent older adults. For instance, health care diagnostic algorithms predominantly trained on younger populations may lead to misdiagnoses for older patients, as the algorithms fail to account for age-related health issues (Chu et al., 2022).

Digital Ageism

Stereotypical views of older adults as technologically inept can result in the development of AI systems that do not meet their nuanced needs. This lack of understanding can hinder the adoption of AI technologies intended to improve their lives (Rubeis et al., 2022).

Transparency and Trust

Opacity of AI algorithms can breed mistrust, particularly among older adults who have historically faced discrimination in health care settings (Papadopoulos et al., 2022).

Recommendations

To mitigate the challenges associated with AI's impact on older adults, the following recommendations emerged from the reviewed literature.

Establish Comprehensive Regulatory Frameworks

Policymakers should create robust regulatory frameworks tailored to AI technologies targeting older adults. These frameworks should mandate regular audits to identify and mitigate biases within AI systems (Rubeis et al., 2022).

Implementing Inclusive Design Practices

Engaging older adults in the design and development of AI technologies is critical. Participatory design workshops can ensure that technologies are tailored to meet the diverse needs and preferences of older users (Chu et al., 2022).

Enhance Digital Literacy Programs

Training initiatives aimed at improving digital literacy among older adults are essential for fostering confidence in using AI technologies. Community centers and libraries can serve as valuable resources for such educational efforts (Rubeis et al., 2022).

Bibliography

1. Chu, C. H., Nyrup, R., Leslie, K., Shi, J., Bianchi, A., Lyn, A., McNicholl, M., Khan, S., Rahimi, S., & Grenier, A. (2022). Digital ageism: Challenges and opportunities in artificial intelligence for older adults. *The Gerontologist,* 62(7), 947–955. https://doi.org/10.1093/geront/gnab167
2. Papadopoulos, C., Castro, N., Nigath, A., Davidson, R., Faulkes, N., Menicatti, R., Khaliq, A. A., Recchiuto, C., Battistuzzi, L., Randhawa, G., Merton, L., Kanoria, S., Chong, N. Y., Kamide, H., Hewson, D., & Sgorbissa, A. (2022). The CARESSES randomised controlled trial: Exploring the health-related impact of culturally competent artificial intelligence embedded into socially assistive robots and tested in older adult care homes. *International Journal of Social Robotics,* 14(1), 245–256. https://link.springer.com/article/10.1007/s12369-021-00781-x

3. Rubeis, G., Fang, M. L., & Sixsmith, A. (2022). Equity in AgeTech for ageing well: The role of social determinants in designing AI-based assistive technologies. *Science and Engineering Ethics*, 28, Article 49. https://link.springer.com/article/10.1007/s11948-022-00397-y

Research Question 9: What are the impacts of AI on racial and ethnic minorities and how should these issues be addressed?

Research Questions Examined in Articles

Across the reviewed literature, a common thread is the examination of how artificial intelligence (AI) impacts racial and ethnic minorities, with a focus on both direct and indirect effects.

Bias in AI systems

Articles consistently address the prevalence of algorithmic bias and its implications, specifically for historically marginalized groups. This bias manifests in areas like health care, criminal justice and employment, creating unequal outcomes for these communities (Timmons et al., 2023; Hernandez-Boussard et al., 2021).

How AI Can Perpetuate Discrimination

The literature explores various ways biases enter AI systems, including biased data collection, flawed model design and deployment in biased environments. These mechanisms cause AI to replicate and even exacerbate existing social inequities (West et al., 2019; Prince & Schwarcz, 2020).

Mitigating Harmful Impacts of AI on Racial and Ethnic Minorities

Recommendations span the adoption of equity-focused AI policies, regular bias audits and community engagement to ensure AI systems are inclusive and fair (Gupta et al., 2022; Karanicolas, 2024).

Summary of Key Findings - Impacts of AI On Racial and Ethnic Minorities

Bias in AI systems is pervasive, reflecting and amplifying existing societal inequalities across sectors such as health care, employment and criminal justice. Historical data embedded with

systemic inequities often drives AI models, leading to biased outcomes. These disparities are further exacerbated by the lack of diversity in AI development teams and the absence of enforceable governance frameworks, resulting in systems that fail to account for the needs of diverse populations. Addressing these issues requires robust ethical governance, inclusive development practices and accountability measures to align AI systems with equity and human rights standards.

Systemic Bias and Inequitable Outcomes

Bias in AI systems is pervasive and deeply embedded in multiple sectors, from health care to criminal justice and employment (Roösli et al., 2021). AI tools often rely on historical data that reflect existing inequalities, leading to biased outcomes against racial and ethnic minorities. For instance, in health care, AI models trained on predominantly White patient data may fail to accurately diagnose conditions in minority populations, resulting in inadequate treatment (Timmons et al., 2023).

Disparities in Health and Economic Sectors

AI systems contribute to racial disparities by affecting decisions related to health care resource allocation, hiring and law enforcement. In one study, AI tools used for health care diagnostics were found to misclassify racial minorities, leading to poorer health outcomes (Hernandez-Boussard et al., 2021). In employment, automated recruitment systems were shown to favor White candidates over equally qualified minority applicants (Prince & Schwarcz, 2020).

Amplification of Structural Inequities

AI not only mirrors but also magnifies existing structural inequalities, particularly when deployed in sensitive areas like criminal justice. For example, risk assessment tools used in judicial settings often predict higher recidivism rates for Black individuals compared to White individuals, even when controlling for similar offense histories (Gupta et al., 2022). This pattern exacerbates racial disparities in incarceration rates and judicial outcomes.

Lack of Representation in AI Development

The lack of diversity in AI development teams is a significant factor in the creation of biased systems (West et al., 2019). The majority of AI research and development is conducted by homogenous groups, leading to the omission of minority perspectives in AI design and testing.

This lack of inclusivity results in models that fail to capture the needs of diverse populations, perpetuating biases (Karanicolas, 2024).

Ethical Governance

Ethical concerns such as transparency, fairness and accountability are widely discussed. Current AI governance frameworks lack enforceability, which allows for the deployment of biased systems without repercussions (Timmons et al., 2023). There is an urgent need for stronger governance mechanisms to ensure that AI systems are aligned with human rights and ethical standards.

Challenges

The reviewed literature identifies several challenges in addressing AI's negative impacts on racial and ethnic minorities.

Algorithmic Bias and Data Limitations

One of the main challenges is the persistence of algorithmic bias, often stemming from non-representative data. When training datasets lack diversity, AI systems produce outcomes that disadvantage minority groups (Roösli et al., 2021). Furthermore, existing datasets may reflect historical discrimination, making it difficult to develop unbiased models without overhauling the underlying data infrastructure.

Regulatory Gaps and Insufficient Oversight

Rapid advancements in AI have outpaced existing regulatory frameworks, leaving significant gaps in oversight. Current regulations like the GDPR in Europe and HIPAA in the United States do not fully address the unique risks posed by AI, such as algorithmic discrimination and data privacy concerns (Prince & Schwarcz, 2020). Without cohesive regulations, AI systems can be misused, further harming historically marginalized populations.

Digital Divide and Inequitable Access

The digital divide remains a significant barrier to equitable AI deployment. Populations in low-income areas, which often include racial and ethnic minorities, lack access to the digital infrastructure needed to benefit from AI-driven innovations (Smith et al., 2020). This lack of access exacerbates disparities in health, education and economic opportunities.

Trust and Public Perception

Mistrust in AI is widespread, particularly among minority communities that have historically faced discrimination from institutions adopting these technologies (Hernandez-Boussard et al., 2021). This skepticism poses a barrier to the adoption of AI tools, especially in health care and public services, where trust is crucial.

Recommendations

To mitigate the negative impacts of AI on racial and ethnic minorities, the literature offers several recommendations.

Establish Comprehensive Regulatory Frameworks

Strong regulatory frameworks tailored to AI are essential. These frameworks should include guidelines for algorithmic transparency, regular equity audits and stringent data privacy standards (Gupta et al., 2022). Policymakers should work toward harmonized international regulations that address the unique ethical risks of AI and ensure that AI systems do not disproportionately harm historically marginalized populations (Karanicolas, 2024).

Implement Bias Mitigation Strategies

AI developers should integrate bias mitigation strategies throughout the AI lifecycle, including diverse training data, regular bias audits and the use of fairness metrics (Timmons et al., 2023). These measures should be supplemented with transparency reports to ensure that AI systems are continuously monitored and adjusted to prevent discriminatory outcomes.

Promote Inclusivity in AI Development

Increasing diversity in AI research and development teams is critical for creating systems that reflect the needs of all communities. This involves recruiting more minority researchers, fostering inclusive work environments and prioritizing community engagement in AI design (West et al., 2019). Such inclusivity will help ensure that AI systems are better equipped to serve diverse populations.

Address the Digital Divide

Policymakers and public institutions must prioritize closing the digital divide by investing in digital infrastructure in underserved communities. This will enable equitable access to AI-driven innovations and prevent the exacerbation of existing inequalities (Smith et al., 2020).

Foster Community Engagement and Trust

Building trust with minority communities requires ongoing dialogue and collaboration. AI developers should engage with affected communities through participatory design processes, ensuring that their voices are heard and their concerns are addressed (Karanicolas, 2024). This approach will promote the co-creation of AI systems that are fair, inclusive and trusted by all stakeholders.

Bibliography

1. Gupta, M., Parra, C. M., & Dennehy, D. (2022). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? *Information Systems Frontiers, 24(*5), 1465−1481. https://pubmed.ncbi.nlm.nih.gov/34177358/
2. Karanicolas, M. (2024). Challenging minority rule: Developing AI standards that serve the majority world. *UCLA Law Review, 71*, 2−19. https://www.uclalawreview.org/wp-content/uploads/securepdfs/2024/05/05-Karanicolas-No-Bleed-2.pdf
3. Prince, A. E. R., & Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review, 105*, 1257−1318. https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data
4. Roösli, E., Rice, B., & Hernandez-Boussard, T. (2021). Bias at warp speed: How AI may contribute to the disparities gap. *Journal of the American Medical Informatics Association, 28*(1), 190–192. https://pmc.ncbi.nlm.nih.gov/articles/PMC7454645/
5. Smith, M. J., Axler, R., Bean, S., Rudzicz, F., & Shaw, J. (2020). Four equity considerations for the use of artificial intelligence in public health. Bulletin of the World Health Organization, 98(4), 290–292. https://pmc.ncbi.nlm.nih.gov/articles/PMC7133473/
6. Timmons, A., Duong, J. B., & Simo Fiallo, N. (2023). A call to action on assessing and mitigating bias in AI applications for mental health. *Perspectives on Psychological Science, 18*(5), 1062–1096. https://pubmed.ncbi.nlm.nih.gov/36490369/

7. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race, and Power in AI. AI Now Institute. https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2

Research Question 10: What are the best practices for addressing community engagement in AI policies?

Research Questions Examined in Articles

The literature addresses various facets of community engagement in the context of AI policy development, implementation and governance. These studies explore how AI policies can better integrate the needs and concerns of diverse communities, ensure equitable outcomes and mitigate the potential negative impacts of AI technologies.

Community Engagement During Policy Development

The involvement of local communities in the AI policy-making process is critical to ensuring that AI systems are culturally sensitive and aligned with community needs. Community engagement fosters policies that reflect local values and address specific challenges unique to each community (Aderibigbe et al., 2023; Robles & Mallinson, 2023).

Community Efforts to Mitigate Negative Impacts

AI's impact on various sectors has been explored in diverse contexts. For example, in health care within developing countries, community efforts focus on leveraging AI to improve access and outcomes while addressing local resource constraints (Aderibigbe et al., 2023). In urban planning for smart cities, community engagement helps guide AI's role in sustainable development and equitable resource allocation (Chauncey & McKenna, 2024). Similarly, in local governments, efforts to adopt AI involve ensuring that systems meet the specific administrative and social needs of the community (Yigitcanlar et al., 2024).

Community Engagement Challenges

Ensuring effective community representation in AI systems presents ethical and practical challenges. Facial recognition technology highlights issues such as bias and the lack of inclusivity in algorithmic design, leading to potential misuse or harm (Ruhrmann, 2019). In global health, challenges include addressing disparities and ensuring that AI systems are equitable and accessible across diverse populations (Guzmán, 2024).

Integrating Community Engagement in Governance

Concrete strategies for integrating community engagement in AI governance include public consultations, the adoption of ethical AI practices and the development of inclusive policy frameworks. These approaches help ensure that AI technologies are developed and implemented in ways that reflect community values and promote equity (Faerron Guzmán, 2024; Chauncey & McKenna, 2024).

Summary of Key Findings – Community Engagement

Community involvement is essential for the successful implementation of AI technologies, ensuring that they are effective, equitable and aligned with public needs. By incorporating significant community input in regions with infrastructural challenges, AI systems can better address local priorities and foster public trust. Engagement with diverse stakeholders also helps mitigate biases, promote inclusivity and enhance the flexibility of AI applications, leading to improved outcomes in areas such as health care, urban planning and public services. Prioritizing community participation ensures that AI technologies are not only technically robust but also culturally and socially appropriate. The literature offers several key findings related to community engagement in AI policies:

Community Involvement as a Crucial Element

Community involvement is widely recognized as essential for the successful implementation of AI technologies. Significant community input is critical in the development of AI systems, particularly in regions with prominent infrastructural challenges, such as developing countries (Aderibigbe et al., 2023). Public trust in AI governance is significantly enhanced when active public engagement is prioritized, fostering greater acceptance and effectiveness (Robles & Mallinson, 2023).

The Role of AI in Enhancing Public Services

AI has the potential to improve public services when community engagement is incorporated into its design and implementation. For instance, AI can optimize health care delivery in resource-limited settings by tailoring solutions to local needs (Aderibigbe et al., 2023). In urban planning, AI can enhance outcomes by facilitating community-inclusive decision-making processes, ensuring that development aligns with public priorities (Chauncey & McKenna, 2024).

Addressing Bias and Inequities Through Community Engagement

Bias in AI systems is a persistent concern in sensitive areas such as law enforcement and health care. Community engagement is essential to identifying and mitigating these biases, especially when involving marginalized groups in the AI development process (Ruhrmann, 2019; Guzmán, 2024). Early community involvement ensures that AI systems are not only technically effective but also equitable and culturally appropriate.

Benefits of Inclusivity and Flexibility

The flexibility of AI systems improves when community engagement is prioritized. For example, AI chatbots in smart cities can enhance cognitive flexibility and promote creative urban solutions when community needs are considered (Chauncey & McKenna, 2024). Similarly, AI systems designed with significant community input can bridge gaps in health care delivery and improve public health outcomes (Balogun et al., 2023).

Challenges

The reviewed literature identifies several challenges for engaging communities and the public about AI.

Infrastructure and Resource Constraints

Infrastructural limitations, such as inadequate internet access and unreliable power supply, are significant barriers to effective community involvement in AI development. These issues are particularly pronounced in developing countries, where they hinder both participation and equitable deployment of AI technologies (Aderibigbe et al., 2023). The digital divide further exacerbates inequities in health care delivery and access to AI tools (Balogun et al., 2023).

Bias and Ethical Concerns

Preventing bias in AI systems is a significant challenge, particularly in sectors like law enforcement, where algorithmic discrimination risks exacerbating societal inequities (Ruhrmann, 2019; Eiras et al., 2024). Addressing these biases requires a concerted effort to ensure that AI algorithms are developed and tested in diverse and representative contexts.

Public Trust and Mistrust of AI

Public mistrust remains a key barrier to AI adoption in governance and other sectors. Transparency and accountability are critical to building trust, but the complexity and opacity of AI decision-making processes make achieving these goals difficult (Robles & Mallinson, 2023).

Over-Reliance on AI Systems

Over-reliance on AI without adequate human oversight can lead to problematic outcomes. Maintaining a balance between automation and human judgment is essential, particularly in strategic planning and decision-making processes (Chukhlomin, 2024).

Recommendations

Recommendations for Addressing Community Engagement Challenges in AI Policies

The literature provides several strategies for addressing the challenges of community engagement in AI policies, focusing on fostering inclusive AI development, ensuring transparency and accountability and bridging the digital divide to promote equitable outcomes.

Investing in Infrastructure and Bridging the Digital Divide

Significant investments in infrastructure are necessary to ensure that AI technologies benefit all communities, particularly in developing countries. Governments and private sector partners must collaborate to improve internet connectivity and power supply, which are essential for effective AI deployment (Aderibigbe et al., 2023). Addressing the digital divide is critical to ensuring equitable health care delivery, particularly in underserved regions such as Africa (Balogun et al., 2023).

Promoting Public-Private Partnerships

Collaboration between government entities, private companies and community organizations is essential for overcoming resource constraints and aligning AI technologies with community needs. Public-private partnerships can foster innovation and enhance resource allocation, while global cooperative approaches to AI governance can ensure consistency across jurisdictions (Aderibigbe et al., 2023; Faerron Guzmán, 2024).

Strengthening Community Engagement Practices

Integrating community engagement into AI policy development and deployment ensures diverse perspectives are represented. Public consultations during policy formation can help incorporate the voices of marginalized and underrepresented groups (Robles & Mallinson, 2023). Designing AI tools with inclusivity and flexibility in mind makes them more accessible to a wide range of community members (Chauncey & McKenna, 2024).

Developing Ethical AI Policies

Ethical considerations must be central to AI policy development. Robust regulatory frameworks

addressing data privacy, algorithmic bias and transparency are necessary for ensuring fairness and accountability (Ruhrmann, 2019; Eiras et al., 2024). Continuous monitoring and auditing mechanisms should be implemented to identify and address potential biases and maintain trust in AI systems.

Fostering AI Literacy and Skill Development

Promoting AI literacy and skill development among communities and policymakers is critical for meaningful engagement with AI technologies. Initiatives such as educational programs at historically black colleges and universities (HBCUs) and other educational institutions can prepare students for the AI-driven workforce and enhance community understanding of AI applications (Long et al., 2024). Ongoing training and stakeholder involvement are necessary for ensuring effective governance and strategic planning in AI implementation (Chukhlomin, 2024).

Integrating Human Oversight in AI Systems

AI systems should not operate autonomously without appropriate human oversight. Maintaining a balance between AI automation and human judgment helps prevent negative outcomes and ensures that AI decisions align with institutional values and community needs (Chukhlomin, 2024; Aderibigbe et al., 2023).

Bibliography

1. Aderibigbe, A. O., Ohenhen, P. E., Nwaobia, N. K., Gidiagba, J. O., & Ani, E. C. (2023). Artificial intelligence in developing countries: Bridging the gap between potential and implementation. *Computer Science & IT Research* Journal, 4(3), 185–199. https://doi.org/10.51594/csitrj.v4i3.629
2. Balogun, O. D., Ayo-Farai, O., Ogundairo, O., Maduka, C. P., Okongwu, C. C., Babarinde, A. O., & Sodamade, O. T. (2023). Integrating AI into health informatics for enhanced public health in Africa: A comprehensive review. *International Medical Science Research* Journal, 3(3), 127-144. https://doi.org/10.51594/imsrj.v3i3.643
3. Chauncey, S. A., & McKenna, H. P. (2024). Creativity and innovation in civic spaces supported by cognitive flexibility when learning with AI chatbots in smart cities. *Urban Science, 8*(1), 16. https://doi.org/10.3390/urbansci8010016

4. Chukhlomin, V. (2024). Exploring the Use of Custom GPTs in Higher Education Strategic Planning: A Preliminary Field Report. SUNY Empire State University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4793697
5. Faerron Guzmán, C. A. (2024). Global health in the age of AI: Safeguarding humanity through collaboration and action. *PLOS Global Public Health, 4*(1), e0002778. https://doi.org/10.1371/journal.pgph.0002778
6. Robles, P., & Mallinson, D. J. (2023). Artificial intelligence technology, public trust, and effective governance. *Review of Policy Research, 40*, 1−18. https://doi.org/10.1111/ropr.12555
7. Ruhrmann, H. (2019). Facing the Future: Protecting Human Rights in Policy Strategies for Facial Recognition Technology in Law Enforcement. Goldman School of Public Policy. https://citrispolicylab.org/wp-content/uploads/2019/09/Facing-the-Future_Ruhrmann_CITRIS-Policy-Lab.pdf
8. Yigitcanlar, T., David, A., Li, W., Fookes, C., Bibri, S. E., & Ye, X. (2024). Unlocking AI adoption in local governments: Best practice lessons from smart cities. *Preprints*. https://doi.org/10.20944/preprints202406.0159.v1
9. Eiras, F., Petrov, A., Vidgen, B., Schroeder de Witt, C., Pizzati, F., Elkins, K., Mukhopadhyay, S., Bibi, A., Purewal, A., Csaba, B., Steibel, F., Keshtkar, F., Barez, F., Smith, G., Guadagni, G., Chun, J., Cabot, J., Imperial, J. M., Nolazco-Flores, J. A., Landay, L., Jackson, M., Torr, P. H. S., Darrell, T., Lee, Y. S., & Foerster, J. (2024). Risks and opportunities of open-source generative AI. *Preprints*. https://arxiv.org/html/2405.08597v1
10. Long, K. C., Gunder, A., Robinson, B., Davis, V. L., & Barth, D. (2024). Leading the AI Revolution: The Crucial Role of HBCUs in Steering AI Leadership. Online Learning Consortium. https://files.eric.ed.gov/fulltext/ED657299.pdf
11. Green Cities Artificial Intelligence. (2024). City of Salem, Sustainable City Year Program. https://scholarsbank.uoregon.edu/xmlui/handle/1794/29247.

Appendix A. Literature Review Methodology

Figure A-1 provides a detailed overview of the key areas examined in the literature review, including research questions, objectives, keywords used and sources of information. Each section of the table highlights specific questions that guided the review process, such as the current landscape of AI adoption in public health, strategies for bias mitigation and the role of human oversight in AI policies. The table also includes insights into the types of studies considered.

Figure A-1. **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	Inclusion Criteria
1.What does the current landscape for the adoption and use of AI in public health look like?	To provide an overview of the current status and trends in AI adoption within the public health sector. To identify key applications and case studies where AI has been successfully implemented. To explore challenges and barriers that may hinder broader adoption of AI in public health.	"AI adoption in public health", "Use of AI in public health", "Current landscape of AI in public health", "Public health AI applications", "AI integration in health policy", "AI implementation in public health", "AI in public health practice", "AI and public health innovation", "Public health technology adoption", "AI and health policy analysis", "Trends in AI for public health", "Public health AI usage", "AI and public health research", "AI-driven public health solutions", "Health policy and AI adoption", "AI frameworks in public health", "AI governance in health policy", "Public health data analytics with AI", "Ethical AI in public health", "AI strategies in public health"	• Recent studies (last 5 years). • Research that primarily focuses on the United States. • Primary research articles, reviews, meta-analyses and policy papers. • Articles focused on AI applications specifically in public health contexts. • Articles discussing both successes and challenges in AI adoption in public health.

Figure A-1 (continued). **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	• Inclusion Criteria
2. How is bias mitigation addressed in AI policies?	To examine strategies for mitigating bias in AI outputs	"Bias mitigation in AI," "AI policy bias reduction," "Bias prevention strategies in AI," "AI fairness and policy," "Equity in AI policies," "Mitigating algorithmic bias," "AI ethical guidelines for bias," "Bias handling in AI frameworks," "Reducing discrimination in AI," "AI bias prevention measures," "Bias correction in AI models," "AI governance and bias," "Ethical AI policy on bias," "Bias in machine learning policies," "Regulatory frameworks for AI bias"	• Recent studies (last 5 years). • Research that primarily. focuses on the United States. • Articles focused on the types of biases that AI systems can include. • Articles discussing strategies to address bias in outputs.
3. How is data privacy addressed in AI policies?	To examine strategies for addressing data privacy concerns associated with the use of AI	"Data privacy in AI policies," "AI data protection strategies," "AI and user privacy," "Data security in AI frameworks," "Privacy safeguards in AI," "AI policy data confidentiality," "AI data governance," "Privacy measures in AI regulation," "AI and personal data protection," "Ethical AI data practices," "AI policy on data security," "Data privacy compliance in AI," "Protecting user data in AI," "Data handling in AI systems," "AI privacy standards."	• Recent studies (last 5 years). • Research that primarily focuses on the United States.

Figure A-1 (continued). **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	Inclusion Criteria
4. How is transparency addressed in AI policies?	To examine strategies for addressing transparency concerns associated with the use of AI	"Transparency in AI policies," "AI transparency strategies," "AI policy transparency measures," "Transparent AI frameworks," "AI explainability in policy," "Ethical transparency in AI," "AI accountability and transparency," "AI policy for clear decision-making," "Algorithm transparency in AI," "Ensuring transparency in AI systems," "Policy guidelines for AI transparency," "Transparent data practices in AI," "AI governance and transparency," "AI decision process transparency," "Enhancing transparency in AI regulations."	• Recent studies (last 5 years). • Research that primarily focuses on the United States.

Figure A-1 (continued). **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	Inclusion Criteria
5.What role do AI policies assign to human oversight and intervention in automated decision-making processes?	To examine the role and necessity of human oversight in AI-driven decision-making processes. To analyze how existing AI policies address human intervention and oversight. To discuss ethical implications and considerations for incorporating human oversight into AI policies.	"AI policies human oversight", "AI human intervention", "Automated decision-making ethics", "Ethical standards in AI", "Human oversight in AI", "AI policy human involvement", "AI ethical decision-making", "Human role in AI policies", "AI governance and human oversight", "Ethics in automated AI systems", "Human intervention in AI processes", "AI policy and ethical standards", "AI decision-making oversight", "Human oversight in AI ethics", "AI ethical governance"	• Recent studies (last 5 years). • Research focuses on the United States and globally. • Articles discussing the role of human oversight in AI decision-making. • Studies analyzing ethical implications of automated systems. • Documents proposing guidelines for human intervention in AI processes. • Case studies illustrating the impact of human oversight in AI governance. • Literature reviews on the importance of human control in AI policies.
6. What are the equity and ethical considerations of AI that should be addressed in policies?	To outline key principles and guidelines that promote ethical AI practices.	"Equity considerations in AI policies", "Ethical considerations of AI", "AI ethics in policy"	• Recent studies (last 5 years) to capture current trends. • Research primarily done in US. • Articles discussing ethical considerations in AI policy frameworks.

Figure A-1 (continued). **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	Inclusion Criteria
7. What are the impacts of AI on individuals with disabilities and how should these issues be addressed in AI policies?	To assess the impacts of AI on individuals with disabilities and identify policy measures to address these issues.	"AI impact on individuals with disabilities", "AI policies and disabilities", “AI ethics and disability considerations," "AI policy for accessibility."	• Recent studies (last 5 years) to capture current trends. • Research that primarily focuses on the United States. • Articles discussing ethical considerations in AI policy frameworks.
8. What are the impacts of AI on older adults and how should these issues be addressed in AI policies?	To evaluate the impacts of AI on older adults and recommend policy measures to address these issues.	"AI impact on older adults", "AI policies and older adults"	• Recent studies (last 5 years) to capture current trends. • Research that primarily focuses on the United States. • Articles discussing ethical considerations in AI policy frameworks.
9. What are the impacts of AI on racial and ethnic minorities and how should these issues be addressed in AI policies?	To examine the impacts of AI on racial and ethnic minorities and propose policy measures to address these issues.	"AI impact on racial minorities", "AI policies and racial minorities", "AI impact on ethnic minorities", "AI policies and ethnic minorities"	• Recent studies (last 5 years) to capture current trends. • Research that primarily focuses on the United States. • Articles discussing ethical considerations in AI policy frameworks.

Figure A-1 (continued). **Detailed Literature Review Methodology by Research Question**

Research Question	Aims of Review	Search Keywords	Inclusion Criteria
10. What are the best practices for addressing community engagement in AI policies?	To identify best practices for incorporating community engagement in AI policies to ensure inclusive and effective participation.	“Community engagement in AI policies", "best practices for AI policy community engagement", "addressing community engagement in AI policies,” "inclusive AI policy development with community input"	• Recent studies (last 5 years) to capture current trends. • Research that primarily focuses on the United States. • Articles discussing ethical considerations in AI policy frameworks.

Appendix B: Protocols and Lessons Learned: Using AI Tools for Literature Review

Protocol

This section details the procedures used by the research team for utilizing two AI tools, ChatGPT and Petal, to support the literature review process. The summary outlines detailed procedures for using AI tools to determine the type of information that should be captured for each research question, create summaries of individual articles and generate an overall summary across all articles.

Section 1: Determining Additional Columns for Comprehensive Evidence Tables: Using ChatGPT

Goal for Section 1: *Identify and incorporate additional columns in the evidence tables that provide a comprehensive analysis of the specific research questions, ensuring all relevant aspects are thoroughly addressed.* An evidence table for a literature review is a systematic tool used to organize and summarize critical information from sources reviewed in the context of a specific research question or topic.

Step 1.1. To effectively determine the need for additional columns in the evidence table (*Figure B-1*) for the specific research questions, the research team employed a two-fold approach described in Steps 1.2 and 1.3.

Step 1.2 Initial AI Column Suggestions: The research team copied and pasted each research question into ChatGPT and asked it: *"What additional columns, beyond those already included in the general evidence table, could be added to effectively address the research question (list research question) while ensuring that equity considerations are centered?"*

Figure B-1: **General Evidence Table**

List Article Title	List Type of Source & Year	List Author	List Research Question/Goal	List Study Design	List All Findings	List Recommendations	List Citation in APA format

Step 1.3 Article-Based Column Refinement: Next, the team members uploaded two sample articles related to the research question to ChatGPT and updated the prompt to ask: *"Based on the articles, what columns, in addition to those already included in the evidence table referenced below, could be added to effectively address the research question (listed research question) while ensuring that equity considerations are centered?"*

Step 1.4 Review and Confirmation: The research team reviewed the results developed through steps 1.2 and 1.3 and confirmed the ones that were reasonable based on their expertise. See *Figure B-2* as an example.

Figure B-2. **Example of the Updated Evidence Table Specific to the Research Question**

Research Question 1: What does the current landscape for the adoption and use of AI in public health look like?										
A	B	C	D	E	G	H	I	J	K	L
List Article Title	List Type of Source & Year	List Author	List Research Question/Goal	List Study Design	List All Findings	List Findings Related to Areas of AI Contribution	List Challenges	List Findings Related to Equity and Ethical Considerations	List Recommendations	List Citation in APA Format

Step 1.5: Excel Setup for Finalized Columns: Once the table columns were confirmed, the team set up an Excel file in each RQ folder on Teams containing these columns.

Section 2: Automated Article Summarization and Verification Process for AI: Using ChatGPT and Petal

Goal for Step 2. *The goal for this step was to efficiently extract, summarize, and verify key information from articles to accurately populate the evidence table, ensuring clarity and accuracy in the summarized data.*

Step 2.1. Download the Articles: The process began by downloading the articles that were to be worked with.

Step 2.2. Upload Articles to ChatGPT/Petal AI: The attachment icon in ChatGPT was used to add one article file at a time.

Step 2.3. Input the Prompt: The research team used the following prompt: “Use the attached article only and not any other information. Populate the table listed below with a detailed summary in each column, to address the research question (list the specific research question). Please include page numbers from the article for each finding. Leave columns blank if the article does not describe the issues identified in the table columns.” (Note for a user: *At the end of the prompt, include the relevant evidence table for the research question.)* It is important to note that in ChatGPT, the internet browsing function was disabled to ensure that the tool only accessed and analyzed the uploaded articles rather than browsing the web.

Step. 2.4. Human Review of the AI-Generated Summary:

- The results included in the table were read and compared to the article.
- If the AI summary did not address equity and ethical considerations, the article was reviewed to see if these aspects were overlooked. If any information was incorrect, a correction was made and it was noted that the AI produced an error.
- If any information was unclear, a follow-up prompt was used. For example: "What specific indicators in the article led ChatGPT to identify the study as exploratory?"

Step 2.5. Transfer Information: The information from ChatGPT was copied and pasted into the Excel file.

Step 2.6. Repeat the Process: The next article was attached and the process was repeated.

Step 2.7: Verification: The page numbers referenced by ChatGPT in the summaries were used to check the articles and ensure that the information was accurate.

Section 3: Summarizing Information: ChatGPT

Goal for Step 3: *The objective is to efficiently generate concise and accurate summaries of the research findings for each relevant column in the table.*

Step 3.1: Extract Column Data: The research team copied the information from each relevant column, one at a time.

Step 3.2: The research team pasted all the copied information into ChatGPT.

Step 3.3: The research team used the sample prompts below to create summaries (*Figure B-3*).

Figure B-3. **Sample Prompts for Summary Creation and QA Strategies**

Column	Summarizing Prompt	Quality Assurance (QA) Process
Type of Source & Year	"Categorize the types of sources and note the distribution of publication years for the articles reviewed."	• Verify that sources are correctly categorized. • Cross-reference years with original articles.
Research Question/Aim	"Summarize the primary research questions or aims of the articles to understand the main focus areas of the studies."	• Ensure summaries align with the research question.
Study Design	"Categorize the study designs used in the articles to assess the methodological approaches taken."	• Verify correct categorization of study designs. • Ensure consistency in summarizing study designs.
Data Sources	"Summarize the types of data sources used in the articles to identify common data collection methods."	• Cross-reference with original articles.
Key Findings	"Extract and summarize the key findings from each article to identify common themes and insights regarding AI policies."	• Ensure summaries are consistent with original findings. • Ensure key findings are relevant to the research question.
Policy Components	"List and categorize the specific components and provisions of AI policies identified in the articles."	• Verify correct categorization of policy components.
Sector	"Categorize the sectors addressed in the articles to determine which areas are most frequently discussed in AI policies."	• Verify correct categorization of sectors. • Ensure all sectors are included.

Figure B-3 (continued). **Sample Prompts for Summary Creation and QA Strategies**

Column	Summarizing Prompt	Quality Assurance (QA) Process
Ethical Principles	"Identify and summarize the ethical principles mentioned in the articles to highlight common ethical considerations. Summarize how the articles address equity considerations, if at all, in AI policies." "Highlight any gaps or inconsistences in how AI policies address bias."	• Ensure ethical principles are accurately summarized. Ensure all ethical principles are included.
Recommendations	"Summarize the recommendations provided in each article to identify suggested best practices and guidelines for AI policies."	• Ensure recommendations align with original articles.
Citation in APA format	"Compile the APA-formatted citations for all articles reviewed to ensure proper referencing."	• Ensure all articles are cited.

Human Oversight of Petal and ChatGPT

The research staff developed an oversight procedure for using ChatGPT and Petal throughout the literature review process. The typical oversight process included reviewing outputs generated based on prompts and verifying the information against the original sources. For easier cross-referencing in the literature review summaries, the prompts included a request to reference page numbers and list authors of the articles. This strategy allowed the research team to verify the results effectively. Quality assurance for several literature review questions also involved additional staff to ensure accuracy.

Lessons Learned from Using AI Tools

The information provided below is not intended to endorse any specific AI tool but is primarily meant to provide a high-level description of the lessons learned related to their usage in the conducted literature review.

In general, both tools — ChatGPT and Petal — demonstrated their utility in the literature review process. The team utilized both free and paid subscription versions of ChatGPT. Advantages included the identification of articles and the development of summaries organized by themes based on prompts with reasonable accuracy. However, in some cases, the summaries were too general and required additional editing through further prompting and additions from the

authors. An additional issue occasionally observed was that some summaries included information not sourced from the articles provided. To mitigate this, the research team implemented measures, including disabling ChatGPT's memory and internet browsing capabilities in the settings, to ensure outputs were based on the article content.

The second AI tool used was Petal. Given that the functionality of the Petal tool differs from ChatGPT — it does not browse the web and works solely with the documents provided — the research team did not need to implement any extra measures beyond standard review and quality assurance to ensure that the summaries accurately reflected the content of the provided articles. However, some limitations noted included challenges in recognizing specialized terminology and phrases used in the articles, which sometimes resulted in outputs that did not fully capture the intended meaning. The team addressed this issue by conducting manual reviews to correct AI-generated outputs, ensuring they aligned with the intended meaning and by adding additional prompts to guide the tool toward producing more accurate and relevant results.

Further testing of these and other AI tools is essential to identify the most effective ways to leverage their capabilities for literature reviews in the future, while also addressing potential ethical considerations.