TECHNOSOPHY

Nicholas Kluge Corrêa,
Julia Maria Mönig (eds.)

STANDARDIZATION OF AI ETHICS

Stakeholders, Values and Profit

[transcript]

Nicholas Kluge Corrêa, Julia Maria Mönig (eds.)
Standardization of AI Ethics

Technosophy | 3

Editorial

Today's rapid technological and scientific progress has unforeseeable epistemic and social side effects. An apt approach must take into account the complex dynamics of social conditions and realities associated with scientific practices of knowledge production and their implementation in technology.

The book series **Technosophy**, published at the "Center for Science and Thought" (CST) of the University of Bonn, presents interdisciplinary work from philosophy, the humanities and social sciences that contributes to wiser uses of technology, and to interpretations of cutting-edge science, both grounded in their very ontologies and epistemologies.

The series is edited by Markus Gabriel, Dennis Lehmkuhl, Aimee van Wynsberghe and Ana Ilievska. Managing editors are Christiane Schäfer and Jan Voosholz.

Nicholas Kluge Corrêa is a postdoc researcher at Universität Bonn. He received his PhD in philosophy (AI ethics) from the Pontifical Catholic University of Rio Grande do Sul (PUC-RS, Brazil) and Universität Bonn. He holds a master's degree in electrical engineering. His thesis "Dynamic Normativity: Necessary and Sufficient Conditions for Value Alignment", was awarded the Capes Doctoral Thesis Award 2025 by the Brazilian government. His main areas of research are AI ethics, AI safety, and AI alignment.

Julia Maria Mönig (Dr. phil.) is a philosopher and project manager of the philosophical part of the KI.NRW flagship project "Zertifizierte KI". She holds a PhD in philosophy from Universität Passau. At Hochschule der Medien Stuttgart, she developed ethics guidelines for highly automated driving. She also worked for the European Network of Research Ethics Committees (EUREC Office) on projects concerning ICT ethics and institutional ethics. She is a member of the Center for Ethics and Humanism at the Free University of Brussels (VUB). Her main research areas are ethics of AI, ethics of technology, privacy research and Hannah Arendt.

Nicholas Kluge Corrêa, Julia Maria Mönig (eds.)

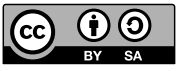
Standardization of AI Ethics

Stakeholders, Values and Profit

[transcript]

Funded by the Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available online at https://dnb.dnb.de



transcript Verlag | Hermannstraße 26 | D-33602 Bielefeld | live@transcript-verlag.de

Cover design: Maria Arndt
Cover illustration: Chris Wickenden, 2023
Printing: Elanders Waiblingen GmbH, Waiblingen
https://doi.org/10.14361/9783839478431
Print-ISBN: 978-3-8376-8241-0 | PDF-ISBN: 978-3-8394-7843-1
ISSN of series: 2943-2987 | eISSN of series: 2943-2995

Printed on permanent acid-free text paper.

Contents

Series Preface
Markus Gabriel 7

Introduction
Nicholas Kluge Corrêa, Julia Maria Mönig 11

Should AI Auditors Be Audited?
Challenges and Paths for Meta-Auditing Artificial Intelligence
Marcelo Pasetti 15

Citizens Infrastructures as a Way to Govern AI's Power to Shape our Shared Representations
How to Make Sure Auditing Institutions Are Aligned with Values and Societal Expectations?
Chiara Marcoccia 39

Algorithmic Audits of HR-AI Tools in the United States
What Are HR Professionals' and Job Seekers' Perspectives?
Tina B. Lassiter, Kenneth R. Fleischmann 67

To be Cared for or Deceived?
Operationalizing Ethics in the Case of Elderly Care Robots under the European AI Act
Gaia Contu 107

Practioner's Corner

Navigating Challenges and Opportunities of Standards Developing Organizations in Times of Rapid Technological Advancement
A DIN e.V. Perspective
Dr. Maria Mensch, Adrian Seeliger, Johannes Wellhöfer 143

Series Preface

Markus Gabriel

The book series *Technosophy* is devoted to the conceptual transformations required to engage adequately with the technological revolutions of the present. Among these, artificial intelligence occupies a singular position. AI systems do not merely extend human capacities; they reshape the frameworks within which knowledge, agency, responsibility, and normativity are understood. The question is therefore no longer whether AI can be made more efficient, but how its development and deployment transform the very standards by which efficiency, fairness, and legitimacy are judged.

This volume addresses a dimension of the AI revolution that is both urgent and philosophically profound: the standardization of AI ethics. As artificial intelligence becomes embedded in infrastructures of governance, healthcare, labor markets, finance, and communication, the demand for auditing, certification, and regulatory oversight has intensified. Yet these mechanisms are not neutral instruments. Standards codify values. Audits operationalize conceptions of fairness, transparency, and accountability. Certification regimes institutionalize particular visions of trustworthiness. In this sense, the attempt to standardize AI ethics is itself a transformative intervention into our normative landscape.

Technosophy brings together work from philosophy, the humanities, and the social sciences to examine precisely such transformations. Technological development is never purely technical. It reorganizes social relations, redistributes power, and encodes assumptions about what counts as knowledge and whose interests matter. The ethical governance of AI therefore requires more than compliance checklists or procedural safeguards. It demands reflection on the ontological and epistemological presuppositions embedded in algorithmic systems and on the values that guide their evaluation.

Rapid technological progress generates epistemic and social side effects that are often unforeseen. AI systems abstract from human practices and

reconfigure them in statistical form. When these abstractions are reintroduced into social life through automated decision-making, recommendation systems, or predictive models, they reshape the conditions under which individuals act and institutions operate. Ethical standardization must therefore grapple with a fundamental tension: how to formalize norms without reducing them to mere technical parameters, and how to ensure that quantifiable metrics do not eclipse qualitative dimensions of justice and responsibility.

The contributions assembled in this volume approach these questions from multiple perspectives. They investigate auditing practices, stakeholder participation, institutional accountability, value conflicts, and the economic dimensions of AI deployment. By examining certification schemes and regulatory frameworks, including those emerging in the European context, the volume highlights both the promise and the limits of current governance models. The standardization of AI ethics is shown to be neither a purely legal task nor a purely technical one, but a deeply philosophical undertaking concerning the institutionalization of values.

A central insight guiding this volume is that AI systems are not alien forces imposed upon humanity from outside. They are products of human practices, languages, and data. In this sense, they can be understood as complex self-portraits of contemporary societies. When we seek to audit, certify, and regulate AI, we are not merely controlling external tools; we are negotiating how our own sedimented norms and biases are to be reflected back into collective life. Standardization thus becomes a site of self-interpretation.

At the same time, the rise of certification regimes and auditing infrastructures introduces new layers of mediation between technological innovation and democratic legitimacy. Who defines the criteria of evaluation? Who audits the auditors? How can stakeholder perspectives be meaningfully integrated into formal procedures? These questions reveal that the governance of AI is inseparable from broader concerns about institutional design and public trust. Standards that lack transparency or independence risk becoming instruments of reputational management rather than genuine accountability.

The challenge, then, is to develop forms of ethical standardization that do not collapse normativity into optimization. Efficiency, robustness, and scalability are indispensable values in technical systems. Yet they cannot alone determine what counts as a good or just application of AI. Decisions about risk, fairness, and acceptable trade-offs are ultimately political and ethical judgments. The attempt to encode them into standards must remain open to contestation and revision.

The *Technosophy* series is committed to examining such tensions without succumbing either to technological determinism or to nostalgic resistance. Humanism, understood as a reflective engagement with the conditions of human agency and responsibility, remains indispensable in the age of AI. The standardization of AI ethics does not mark the end of ethical reflection; it intensifies its necessity. For as soon as norms are formalized, the question arises whether the formalization itself is adequate, inclusive, and responsive to evolving social realities.

By situating debates about auditing, certification, and regulation within broader philosophical frameworks, this volume exemplifies the ambition of *Technosophy*: to reconnect technological innovation with critical reflection on meaning, value, and shared worldhood. The governance of artificial intelligence is not only about controlling risk. It is about shaping the normative infrastructures within which future societies will deliberate, decide, and coexist.

Only by bringing philosophical analysis into sustained dialogue with technical and institutional practice can we hope to ensure that the standardization of AI ethics strengthens, rather than undermines, democratic legitimacy and human responsibility.

Introduction

Nicholas Kluge Corrêa, Julia Maria Mönig

The rapid development of artificial intelligence (AI) technologies, the potential ethical and legal issues arising from their deployment, and the legal requirements, including those under the European AI Act, have increased the need for reliable, transparent, and independent auditing structures and mechanisms. While certain technical characteristics might be easy to assess, and it might be possible to refer to existing audits, AI technologies potentially present specific challenges to human rights, ethical values, and constitutional democracy.

Hence, what a certification could look like is only one of the questions that should be asked when discussing ethical certification of AI. Furthermore, how all potentially concerned parties (stakeholders) can be included, which values might be affected, and what this means for those actors who (rightfully) need and want to profit from AI applications were central questions that the philosophy team of the KI. NRW-flagship project „Zertifizierte KI" (Certified AI) investigated. Further key questions raised in the contributions and throughout the project period include: Can ethics be standardized? Who audits the auditors? How can auditing institutions be aligned with actual societal concerns? What can an assessment look like for a specific area (e.g., human resources)? How are law and ethics intertwined? Which contribution can philosophy provide, now that there is a European regulation on Artificial Intelligence (e.g., in the discussion whether robots in elderly care should be considered as deceptive technology under the AI Act)? Finally, what role do standardization bodies play in ensuring trustworthy AI?

Answers to these questions are presented in this book, as well as in the project's central outcomes: the *Catalogue of General Ethical Requirements for AI Certification* and the *Catalogue of Field-Specific Requirements for AI Certification*. The *Catalogue of General Ethical Requirements for AI Certification* provides a structured framework for trustworthy AI development, identifying overarching ethical requirements for AI certification and outlining six ethical values: fairness,

privacy, robustness, sustainability, transparency, and truthfulness. For each value, the project team researched existing tools, metrics, and approaches to help developers, programmers, and other practitioners build trustworthy AI systems. The *Catalogue of Field-Specific Requirements for AI Certification* offers a list of requirements, an ethical stakeholder analysis, and a discussion of potential value tensions for six exemplary fields: arts, biometric profiling, industry and critical infrastructure, law enforcement, medicine and care and social media, a list of requirements, an ethical stakeholder analyses and a discussion of potential value tensions.

The volume at hand collects the contributions to the final conference of the sub-project philosophy of the KI.NRW-Flagship-Projekt "Zertifizierte KI". The topic "Standardization of AI Ethics: Stakeholders, Values, and Profit" offered the opportunity to critically discuss the above-mentioned questions and current tendencies in AI auditing and certification, and to raise ethical issues that the current schemes might entail or overlook.

In his paper "Should AI Auditors Be Audited? Challenges and Paths for Meta-Auditing Artificial Intelligence", Marcelo Pasetti argues that while AI auditing in various forms is considered necessary and desirable in the European Union and other countries, it often falls short of an independent assessment. He proposes a meta-audit to guarantee their impartiality and objective analyses.

Chiara Marcoccia examines the challenge of connecting auditing and governance in the case of AI for image generation. In her paper entitled "Citizens infrastructures as a way to govern AI's power to shape our shared representations", she proposes to leverage social sciences to address power asymmetries in genAI auditing, and their consequences across societal levels.

Gaia Contu examines in her contribution "Social Robots in Elderly Care: Ethics, Regulation, and Design" whether this is true and, if yes, whether, under the European AI Act, social robots in elderly care would need to be considered as "high risk" according to Article 6 of the Act. After discussing several other ethically issues with AI and robotics – the (in)famous COMPAS case, the often-quoted Amazon-algorithm for the selection of job applicants and the equally (in)famous Dutch childcare scandal, Contu concludes, that social robots in elderly care are not per se deceptive and suggests therefore a framework to estimate the risk posed by social robots in elderly care.

In the context of robotics for elderly care, it is being debated whether the use of social robots constitutes deception for the persons concerned. Gaia Contu examines in her contribution "To Be Cared For or Deceived? Opera-

tionalizing Ethics in the Case of Elderly Care Robots under the European AI Act" whether this claim holds and whether it can be addressed under Article 5 of the AI Act on prohibited AI practices. Drawing on considerations of the non-neutrality of technology, illustrated by well-known cases such as COMPAS and the often-cited Amazon recruitment algorithm, Contu concludes that ethical considerations must be operationalized from the earliest stages of designing robots for elderly care but that robots used in elderly care are not deceptive per se.

Tina Lassiter and Kenneth R. Fleischmann present the findings of a study in which they investigated what HR Professionals' and Jobs Seekers' perspectives on "Algorithmic Audits in Human Resources" in the US are. The study highlights substantial differences between the two stakeholder groups regarding knowledge of AI use, trust in AI and HR-AI tools, and their knowledge and opinions regarding algorithmic auditing. As specific legislation, they look into the European AI Act and the New York City Law 144.

Finally, Maria Mensch, Adrian Seeliger, and Johannes Wellhöfer offer a practitioner's perspective on how Standard Developing Organizations (SDOs) might contribute to the design, deployment, and development of ethical AI. Their view from the German Standardization Organization DIN suggests that standardization can contribute to aligning technological advancements with societal values, ensuring responsible innovation while safeguarding fundamental rights.

The papers in this volume were presented at the eponymous conference held in September 2025 at the Center for Science and Thought of the University of Bonn. The editors would like to thank those who helped to make this event a success (on stage and behind the scenes). Thanks are also due to Chris Wickenden, who not only enriched our conference with his art works, but also granted us permission to use his image "Suspense of Identified Companionship", created in exchange with the AI character Preti O'Sum, for the cover design. The project including the publication of this volume was made possible through the funding by the Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen.

Should AI Auditors Be Audited?

Challenges and Paths for Meta-Auditing Artificial Intelligence

Marcelo Pasetti[1]

Abstract: *The growing use of artificial intelligence (AI) in public administration, corporate governance, and high-risk sectors such as health, justice, public safety, and credit, which are recognized as domains that affect fundamental rights, human dignity, and access to essential services, has intensified debates about the need for reliable, transparent, and independent algorithmic audits. Although these audits are often presented as instruments of transparency and accountability, the article demonstrates that the current audit ecosystem remains marked by technical deficiencies, conflicts of interest, and insufficient oversight of the auditors themselves. Based on a qualitative and normative approach, grounded in a documentary analysis of recent international regulations, institutional technical guidelines, and specialized scholarly literature, the study examines the role of audits as governance mechanisms and the risks arising from regulatory capture, private self-regulation, and institutional opacity. Finally, it advocates for the creation of multilevel audit ecosystems, anchored in public or hybrid meta-audit structures, social participation, and shared standards of integrity, transparency, and democratic legitimacy.*

Keywords: *Algorithmic auditing; Artificial intelligence; Meta-auditing; Epistemic justice; Accountability*

1 Acknowledgements – This research is part of the RAIES (Rede de Inteligência Artificial Ética e Segura), supported by FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul).

Introduction

The rapid expansion of the use of Artificial Intelligence (AI) systems in public and private sectors has generated growing concerns about the risks of algorithmic bias, decision-making opacity, and structural inequality in the distribution of their impacts. AI tools are already used in sensitive domains such as recommendation systems, health, credit, and criminal justice, often without adequate transparency regarding the operational criteria and responsibilities involved (Bandy 2021; Mittelstadt 2016). In this context, algorithmic audits have been consolidated as technical and regulatory mechanisms for verifying automated systems' compliance, fairness, and legality. However, as Jack Stilgoe (2023) warns, public trust in AI is not based solely on technical tests; it requires a reflexive approach to the role of auditors who hold epistemic power in evaluation processes. This raises broader political concerns in liberal-democratic systems, where insufficient oversight may generate adverse effects on citizens' rights and public accountability.

As auditing practices become institutionalized, a central question arises: **should AI auditors be audited?** The growing professionalization of the sector is illustrated by initiatives such as the Ada Lovelace Institute's *Code & Conduct: How to Create Third-Party Auditing Regimes for AI* and the development of technical standards by the IEEE Standards Association. Even so, the scenario remains in formation, marked by asymmetries of epistemic and political power and the absence of structured mechanisms for supervising the auditors of algorithmic systems themselves. As demonstrated in the empirical mapping conducted by Costanza-Chock et al. (2022), many audits are carried out by internal teams or contracted directly by the companies responsible for the audited systems, compromising the results' impartiality. The lack of consolidated standards and robust regulation can turn audits into mere reputational validation instruments rather than effective accountability and transparency mechanisms.

This research investigates the ecosystem of algorithmic audits with a specific focus on the lack of oversight over AI auditors, to understand the main epistemic, regulatory, and political risks arising from the lack of institutional accountability. Although the field of auditing is developing, this study adopts the central hypothesis that the effectiveness of audits requires the construction of permanent meta-audit structures capable of ensuring the integrity of the assessment process through independent oversight, systematic review, and public transparency. Without such mechanisms, audits risk reinforcing informa-

tional inequalities and legitimizing technocratic practices without promoting true accountability.

This debate takes on relevance in the European context, with the entry into force of Regulation (EU) 2024/1689 (AI Act), which establishes a comprehensive regulatory architecture for AI's ethical and safe use. The regulation classifies AI systems by risk level and imposes proportionate obligations, providing for pre-market compliance assessments and post-market monitoring mechanisms. Although it represents an important milestone, the AI Act still lacks clear guidelines on the supervision of auditors and on ensuring the institutional integrity of notified bodies (Mökander et al. 2022). Its emphasis on internal self-assessment procedures by providers, further reduces the effectiveness of enforcement and limits the regulatory model's ability to prevent systemic risks.

In this scenario, the proposal for dynamic audits is gaining strength as a viable alternative to traditional static verification. Maughan et al. (2022) introduce the concept of prediction sensitivity, aimed at continuously auditing counterfactual fairness in deployed classifiers and detecting degradations in fairness over time. This approach reinforces the importance of iterative and adaptive audits, especially in systems subject to operational changes. The effectiveness of this proposal, however, depends on robust institutional structures and public policies that support regular and transparent inspection practices, as discussed in the Ada Lovelace Institute, AI Now Institute, and Open Government Partnership report *Algorithmic Accountability for the Public Sector: Learning from the First Wave of Policy Implementation* (2021).

The methodology of this study is qualitative and analytical, based on a bibliographic and documentary review of recent regulatory, institutional, and scientific sources, highlighting regulatory and institutional developments in the European Union and the United States. The theoretical framework adopts the work of Vecchione, Levy, and Barocas (2021), who interpret auditing as a historical tool of social justice, aimed at accountability and the inclusion of affected groups. This perspective is combined with the approach of Schiff, Kelley, and Camacho Ibáñez (2024), whose analysis of regulatory capture, delegation, and informational asymmetry highlights the institutional and political limits of ethical AI auditing. The integration of these two strands, auditing as a tool of social justice and auditing as an institution vulnerable to capture, provides the basis for assessing both the promises and shortcomings of current audit practices. In doing so, the article situates the discussion within broader debates on accountability, epistemic justice, and democratic governance.

In this article, epistemic justice is understood operationally as the equitable distribution of informational and interpretive power in algorithmic evaluation processes. It is about ensuring that affected groups have real opportunities to participate, challenge, and influence the criteria and results of audits, avoiding asymmetries of voice and authority that produce systematic disadvantages (Costanza-Chock et al., 2022; Vecchione, Levy & Barocas, 2021; Stilgoe, 2023).

The article is divided into two sections. The first addresses the structural and epistemological challenges of the current AI audit ecosystem and exposes technical, institutional, and regulatory flaws. The second section presents normative and institutional pathways for strengthening audit governance, focusing on implementing meta-audits and public accountability mechanisms. The conclusion summarizes the findings and indicates guidelines for building a more ethical, trustworthy, and socially inclusive audit ecosystem.

The Algorithmic Audit Ecosystem and Its Structural Challenges

The growing institutionalization of algorithmic auditing as a governance tool has revealed not only its technical potential but also the institutional and regulatory challenges that undermine its effectiveness. When proposing that AI systems be auditable, the need for stable criteria, independence of evaluators, and effective accountability mechanisms is assumed. As Costanza-Chock et al. (2022) point out, although algorithmic audits have gained prominence as accountability instruments, the field still faces the absence of consolidated methodological standards, transparency in procedures, and effective participation by civil society organizations. They argue that a truly robust audit ecosystem requires technical independence, external oversight, and the active involvement of communities potentially impacted by AI systems. Bandy (2021) notes that, despite the growing importance of algorithmic audits, there is still little clarity about the methods used and the criteria that determine an effective audit, highlighting the need for greater systematization and standardization in the field.

Costanza-Chock et al. (2022) state that the algorithmic audit ecosystem remains marked by institutional fragmentation, self-regulation, and a lack of defined criteria regarding the qualifications of auditors and the quality standards of the audits carried out. This framework forms an asymmetrical environment in which accountability is limited and unequal. The disparities in epistemic

power between private auditors and the individuals directly affected by automated decisions aggravate the lack of systemic accountability. As Vecchione, Levy, and Barocas (2021) argue, audits can cease to fulfill their critical control function and become instruments for legitimizing technically validated but socially unjust practices.

Faced with these challenges, this section examines the technical and political functions of audits (Section 1.1) and the institutional and regulatory obstacles that undermine their legitimacy (Section 1.2). Together, these analyses clarify why the question "should AI auditors be audited?" has become central to contemporary algorithmic governance.

Auditing as an Algorithmic Governance Tool

The rapid advance of AI technologies has created an increasing demand for institutional mechanisms that ensure their deployment is ethical, fair, and lawful. In this context, algorithmic auditing has been consolidated as a strategic tool for the governance of these systems. Although there is no universally accepted definition, recent studies advocate for broadening the methodological scope through the introduction of sociotechnical audits. According to Lam et al. (2023), these audits evaluate algorithmic systems considering both the technical components and the impacts on users, broadening the traditional lens to consider the continuous interaction between algorithms and people.

As described in the report *Code & Conduct: How to create third-party auditing regimes for AI,* by the Ada Lovelace Institute (2024), auditing is understood as a systematic process to evaluate algorithmic systems based on explicit criteria of compliance, performance, and responsibility. The sociotechnical approach proposes expanding algorithmic auditing beyond the technical analysis of computational inputs and outputs, incorporating the social, institutional, and political contexts in which systems operate. According to Lam et al. (2023), this audit model recognizes that algorithms do not exist in isolation but interact with broad sociotechnical ecosystems, in which different forms of power shape systems' functioning and their distributive effects.

Among the main functions attributed to AI audits are: ensuring compliance with legal norms, promoting public trust in automated systems, and preventing algorithmic discrimination. However, there are clear limits to their effectiveness. As Schiff et al. (2024) point out, the challenges arise not from the auditors themselves but from the structural conditions under which audits are conducted. Because many audits are commissioned and financed directly by AI

developers, the process becomes vulnerable to conflicts of interest, which undermines professional independence and weakens the credibility of the results in fragile regulatory contexts. Similarly, Costanza-Chock et al. (2022) point out that many algorithmic audits occur without sufficient transparency, making external verification and reproducibility of results difficult. The authors note the absence of a consolidated consensus on the criteria that define a rigorous audit, a gap that contributes to methodological fragmentation and weakens the reliability of auditing as an effective instrument of accountability.

These limits are intensified in the public sector, where outsourcing automated systems to private providers has become a recurring practice, especially in sensitive areas such as policing, social benefits screening, and administrative decisions. The Electronic Privacy Information Center (EPIC) report documents how delegating public functions to contractors can weaken democratic transparency and hinder institutional accountability, especially when systems fail or produce unfair decisions (Fergusson, 2023).

A notable example is the COMPAS recidivism algorithm in the United States, whose risk assessment model revealed systematic racial disparities in criminal sentencing outcomes (Angwin et al., 2016; Angwin et al., 2016b). Similarly, the Dutch Child Benefits Scandal, examined by researchers from the University of Groningen (Bouwmeester, 2023), exposed how an automated fraud detection system unjustly flagged thousands of families, disproportionately those with dual nationality, resulting in severe financial and social harm. Both cases illustrate how opaque or poorly supervised algorithmic systems, when embedded in public administration, can perpetuate structural discrimination and erode public trust in governmental institutions.

In this sense, algorithmic auditing transcends the technical sphere, becoming a political mechanism that directly influences the legitimacy of automated decisions. The absence of regulations establishing minimum independence standards, methodological transparency, and external supervision can turn the auditing process into an opaque formality, potentially legitimizing technically verified injustices. Stilgoe (2023) emphasizes that evaluating AI technologies requires not only assessing their technical effectiveness but also the public value they generate, underscoring the need for institutional structures that safeguard democratic accountability.

Diagnosis of Systemic Failures and Risks

The consolidation of algorithmic audits as governance instruments encounters persistent obstacles that undermine their effectiveness, impartiality, and democratic legitimacy. These obstacles, widely documented in recent literature, can be categorized into technical, institutional, and regulatory failures. Schiff et al. (2024) show that information asymmetries, conflicts of interest, and a lack of methodological standardization among auditors mark the audit ecosystem. Bandy (2021) points out that most studies on algorithmic audits focus exclusively on quantifiable metrics, neglecting social and structural impacts. Mökander et al. (2022) and Wörsdörfer (2023) point out that the European Union's regulatory model lacks clear and effective mechanisms for supervising auditors, which weakens *institutional accountability.* Diagnosing these structural flaws is fundamental to supporting the debate on the need for independent and continuous meta-audit structures capable of conferring public legitimacy on the regulatory process and preventing institutional capture phenomena.

On a technical level, most audits focus on metrics such as accuracy, recall, and statistical precision of AI systems, but neglect critical aspects related to algorithmic fairness, social impact, and indirect discrimination. As Bandy (2021) further observes, many studies on algorithmic auditing prioritize quantitative approaches, focusing on metrics such as accuracy and unequal impact, while neglecting broader issues of social justice and institutional context, which can limit the identification of systematic biases.

Maughan et al. (2022) propose the concept of *prediction sensitivity* as a metric aimed at continuously auditing counterfactual fairness in classifiers that have already been implemented. This approach seeks to measure whether and how the output of a system would change if the individual being evaluated belonged to a different social group, with all other characteristics remaining constant. By allowing the detection of subtle asymmetries in treatment, even in the explicit absence of sensitive attributes, this metric highlights the inadequacy of one-off audits carried out only during system development. Based on this limitation, the authors argue that algorithmic fairness should be monitored over time, in real contexts of use, which requires an iterative and continuous audit model. In this context, the proposal for so-called dynamic audits gains relevance, as it prioritizes constant monitoring of the performance of systems in operation as the most appropriate strategy for dealing with the distributive and epistemic risks associated with adaptive classifiers.

At the institutional level, AI audits are often conducted by companies hired or financed by the developers or operators of the audited systems, generating a structural conflict of interest. This economic dependence compromises the auditors' independence and undermines their conclusions' credibility. Schiff et al. (2024) warn that, without robust criteria and independent supervision, ethical AI audits can be appropriated by organizational dynamics, serving more as symbolic instruments than effective accountability mechanisms. The lack of methodological standardization among auditing firms aggravates this dynamic. As the Holistic AI report (2023) points out, the absence of internationally consolidated standards for audits of large language models (LLMs) results in divergent approaches between organizations in terms of methodology and evaluation criteria. This disparity can jeopardize audit comparability and make it difficult to form a reliable and interoperable algorithmic enforcement ecosystem.

On a regulatory level, the AI Act represents a significant milestone by establishing requirements proportionate to the risk of AI systems, including obligations for conformity assessment, post-market monitoring, and the designation of notified bodies. These bodies must demonstrate technical competence, independence, and impartiality. However, the regulation gives member states a broad discretion regarding these bodies' designation, notification, and supervision, without providing detailed and uniform mechanisms for external auditing or ongoing public accountability. This gap can compromise the robustness of oversight, especially in sensitive areas. Mökander et al. (2022) have already warned that the proposed European regulation lacks operational guidelines on the supervision of auditors, especially about their institutional independence, the integrity of compliance processes, and the responsibility of notified bodies.

Even after the publication of Delegated Regulation (EU) 2024/1773 by the European Commission, which details criteria on the organizational independence of notified bodies, management of conflicts of interest, and documentation requirements, gaps remain about external oversight and the effective accountability of these entities. The regulatory text does not establish concrete meta-audit mechanisms or impose clear obligations of public transparency on the audit procedures carried out, allowing these bodies to act with a wide margin of autonomy and low exposure to institutional scrutiny. This weakness compromises the effectiveness and legitimacy of the European Union's conformity assessment model, especially in sensitive AI applications.

These technical, institutional, and regulatory failures can be interpreted through the lens of regulatory capture theory, as formulated by Stigler (1971), which posits that regulatory bodies tend to be gradually appropriated by the sectors they are intended to oversee. In domains characterized by high technical complexity and low transparency, such as AI auditing, this capture occurs in diffuse and incremental ways, aligning auditors over time with the values and narratives of the organizations they evaluate. This process undermines critical independence and diminishes the transformative potential of audits, as also noted by Schiff et al. (2024), who warn of the institutional absorption of corporate priorities by the evaluators themselves.

In addition to these distortions, there are the classic problems of the principal–agent relationship, in which the public (as principal) delegates supervisory functions to technically specialized agents (auditors), whose objectives do not always converge with the collective interest, and whose actions usually occur with low visibility and little accountability (Eisenhardt, 1989). As the report *Algorithmic Accountability for the Public Sector* (2021) summarizes, implementing effective algorithmic accountability policies faces structural obstacles, such as institutional fragmentation, the absence of common terminology between public bodies, and the lack of clear legal incentives to support robust and continuous supervision.

This set of factors reveals a fragile algorithmic audit ecosystem, marked by low standardization, conflicts of interest, and the absence of independent oversight mechanisms. The prevalence of symbolic audits and the lack of accountability on the part of auditors undermine the legitimacy of the entire process. These flaws highlight the need to audit the auditors themselves, thereby clarifying the central question of this study: should AI auditors be audited?? Additionally, these challenges underscore the urgent need for a broader normative and institutional reconfiguration of AI auditing practices, grounded in independence, transparency, and democratic accountability.

Institutional, Regulatory and Technical Pathways Towards Trustworthy Audits

Considering the technical, institutional, and regulatory flaws identified in the current algorithmic audit ecosystem, advancing beyond isolated adjustments has become imperative. Consolidating trustworthy audits requires a comprehensive reconfiguration grounded in multi-level governance, articulating legal

guidelines, internationally recognized technical standards, effective accountability mechanisms, and channels for informed public participation. As noted in the *Code & Conduct* report by the Ada Lovelace Institute (2024), the creation of auditable ecosystems requires not only consistent assessment methods, but also legitimized institutional structures that are open to public scrutiny. From this perspective, this section examines the main regulatory and institutional frameworks currently in place, emphasizing the European context, but with an eye on converging global initiatives.

The AI Act establishes a risk-based regulatory framework for AI systems. The regulation classifies AI systems according to risk and imposes corresponding obligations, such as conformity assessments, notifications, technical record keeping, and post-market monitoring. However, as discussed in the previous section, criticism persists regarding the lack of specific mechanisms for supervising notified bodies and the limited external monitoring instruments. Studies indicate that the model proposed by the AI Act still lacks operational guidelines on how to guarantee the integrity and independence of auditors responsible for verifying compliance (Mökander et al., 2022). Although the regulatory text outlines technical and administrative obligations for AI providers, it does not define specific parameters for auditing the bodies responsible for certification, which raises doubts about the effectiveness of regulatory control.

The literature also indicates that institutional reforms should follow a sequenced logic. Priorities include: (i) establishing documentation and traceability infrastructures (Ada Lovelace Institute, 2024; ISO/IEC 42001); (ii) strengthening the independence and regulatory capacity of supervisory authorities (Mökander et al., 2022); and (iii) developing participatory and contestability mechanisms that redistribute epistemic power (Vecchione, Levy & Barocas, 2021). This sequence emphasizes that governance capacity must precede technical expansion to prevent capture and ensure meaningful accountability.

In parallel with the European Union's regulatory initiative, international efforts have expanded to define technical and institutional parameters for auditing AI systems. Notable among these efforts is the Ada Lovelace Institute's *Code & Conduct* report (2024), the Standards Council of Canada (SCC) accreditation program, the ISO/IEC 42001:2023 standard, and the initiatives of the International Association of Algorithmic Auditors (IAAA). These initiatives converge to overcome the limitations of the self-regulatory model, proposing to adopt more robust criteria for traceability, documentation, and independent

verification. Although at different stages of maturity, these instruments seek to structure an audit ecosystem based on multiple layers of control and guided by principles of public legitimacy, regulatory interoperability, and procedural transparency.

The following sections examine the main regulatory, institutional, and technical pathways currently available, or under development, for improving AI system audits. Section 2.1 provides a critical analysis of the AI Act and its delegated regulations, emphasizing the structure and limitations of independent audit mechanisms and the performance of notified bodies. Section 2.2 discusses the concept of dynamic auditing, aimed at the continuous monitoring of systems after implementation, and evaluates the technical and institutional requirements necessary for its effectiveness. Finally, Section 2.3 explores the construction of collaborative ecosystems and the proposal of meta-audit structures, drawing on international experiences involving public participation, independent scientific research, and socio-technical accountability mechanisms.

Regulatory Frameworks and Compliance Mechanisms

The AI Act represents the most comprehensive legislative initiative to regulate AI within a democratic legal framework. Its structure adopts a risk-based approach, classifying AI systems into four categories: prohibited, high-risk, limited-risk, and minimal-risk. Systems considered high-risk include applications in sensitive sectors such as critical infrastructure, employment relations, administration of justice, and law enforcement, including policing, migration control, and border management, as well as applications related to access to essential services and benefits, including health (see Arts. 6 and 8 and Annex III of Regulation (EU) 2024/1689).

For systems classified as high-risk, the AI Act requires suppliers to undergo one of three conformity assessment procedures: (i) self-assessment based on criteria defined in the Regulation and harmonized standards; (ii) evaluation of technical documentation with the participation of a notified body; or (iii) full certification by a notified body (Art. 43). The Regulation also sets out strict requirements for the designation, independence, and accountability of notified bodies (Arts. 31–34), including periodic audits and ongoing supervision by the notifying authorities of Member States. However, as Mökander et al. (2022) observe, gaps remain regarding independent structures capable of exercising continuous oversight over authorized auditors,

particularly concerning methodological transparency and systematic public accountability. The absence of meta-audit mechanisms, social participation, or cross-audits conducted by external entities limits the capacity of the European model to prevent information asymmetries and institutional capture in socially sensitive contexts.

Within the broader European Union regulatory ecosystem, the European Commission adopted Delegated Regulation (EU) 2024/1773, which establishes minimum criteria for the activities of notified bodies, including periodic audits, impartiality requirements, and documentation obligations. Although this Delegated Regulation represents progress in consolidating operational and internal governance parameters, its mechanisms remain focused on indirect state supervision and do not provide for independent external audits or public meta-audit structures. Consequently, transparency and systematic public scrutiny of algorithmic certification processes remain limited (European Commission, 2024). The AI Act (Regulation (EU) 2024/1689), adopted on 13 June 2024, still lacks operational mechanisms to ensure the independence and transparency of notified bodies, particularly regarding continuous supervision, methodological disclosure, and external accountability mechanisms (Mökander et al., 2022; European Union, 2024).While the European framework remains the most advanced legislative experiment in AI regulation, a broader transnational field is emerging to define technical, ethical, and institutional standards for algorithmic audits. These international initiatives seek to fill the gaps left by national and supranational legislation, providing more detailed assessment parameters compatible with legal and organizational contexts. A key example is ISO/IEC 42001:2023, published in December 2023, which defines requirements for AI management systems focused on traceability, continuous improvement, and the integration of governance principles across the algorithmic life cycle.

In the field of ethical standardization, IEEE has promoted guidelines aimed at incorporating human values into the development and application of autonomous and intelligent systems. The document *Ethically Aligned Design* (2023) establishes a normative vision oriented towards human well-being, arguing that AI design should prioritize human dignity, agency, and fundamental rights. This approach provides the foundation for developing the IEEE P7000 series of standards, which seeks to operationalize ethical principles through technical requirements, including traceability, interpretability, and explainability of systems. Still evolving, these standards aim to expand the

scope of algorithmic audits by integrating social and moral requirements into technical evaluation processes.

This perspective is corroborated by the *Code & Conduct* report (Ada Lovelace Institute, 2024), which proposes that AI audits be understood as sociotechnical and relational processes, involving not only technical verifications, but also ethical commitments, transparency obligations, and effective public accountability mechanisms.

Among the certification models under development, the program launched in 2024 by the SCC stands out, establishing formal criteria for the accreditation of entities responsible for auditing AI systems. The proposal seeks to align local regulatory governance with international interoperability standards to ensure greater institutional reliability. This initiative is complemented by the work of the Canadian Standards Association (CSA), which proposes peer auditing and cross-certification mechanisms between independent institutions to reduce the risks associated with self-regulation among private companies in the sector. In both cases, the importance of subjecting certifying entities to external and regular oversight is recognized as a condition for the legitimacy and effectiveness of algorithmic evaluation processes.

Building on the consolidation of regulatory and technical initiatives such as those led by the Ada Lovelace Institute, ISO, and IEEE, the International Association of Algorithmic Auditors (IAAA) has assumed a particularly relevant role in shaping the ethical and operational guidelines of algorithmic auditing at the global level. Adopting a sociotechnical approach, the association emphasizes that AI systems are not merely technical artifacts, but rather the outcome of institutional, social, and political arrangements. From this perspective, the IAAA Code of Conduct (2024) proposes that auditors act with independence, fairness, and responsibility, explicitly considering the distributive effects of automated decisions and the risks of discrimination, exclusion, and opacity. Complementarily, the IAAA AI Auditing Definitions (2024) underline that effective audits must integrate multiple dimensions of analysis—including transparency, accountability, interpretability, and epistemic justice—while ensuring the active participation of communities potentially affected by algorithmic systems. In doing so, the IAAA advances a comprehensive vision of auditing as a process that transcends the verification of codes and data, demanding the scrutiny of institutional and organizational structures that sustain algorithmic practices, and reinforcing the normative commitment to integrity, contestability, and democratic governance.

Algorithmic audits have been widely promoted as a technical solution to the risks of artificial intelligence, but they are insufficient to ensure effective accountability. As proposed by the AI Now Institute (2023), it is necessary to go beyond the limits of the traditional audit model and build robust institutional structures, with enforcement power and the ability to impose corrective measures, suspensions, or sanctions when systems cause harm. Meaningful accountability requires technical verification of compliance and the insertion of audits into broader regulatory ecosystems, involving the active participation of civil society and state mechanisms of public control.

This criticism converges with the proposal for epistemic justice defended by Costanza-Chock et al. (2022), according to which audits should incorporate participatory and community approaches. For the authors, the groups most affected by technologies need an active voice in defining the audit criteria, methods, and objectives, as a condition for their social legitimacy and contextual sensitivity. Both perspectives point to the need to transform audits into public accountability instruments, not just technical processes conducted under corporate logic.

In this sense, sociotechnical audits are based on continuous evaluation processes, with the involvement of multiple actors and attention to the concrete effects of systems on specific groups and contexts. Institutional structures capable of ensuring independence, adaptability, and openness to public scrutiny are essential for such audits to be viable. Although still incipient in most national legislations, these practices point toward a necessary transformation of algorithmic governance, one guided by democratic legitimacy and social justice.

Dynamic Auditing and Post-Market Monitoring

While traditional audits focus on discrete and static evaluations, typically conducted during the development phase or immediately prior to system deployment, the advancement of AI in dynamic and adaptive contexts requires continuous verification models. In this scenario, dynamic auditing presents itself as a more appropriate response to new complexities by proposing iterative inspections that remain sensitive to transformations occurring throughout the entire lifecycle of automated systems. As Maughan et al. (2022) argue, compliance should not be treated as a final state, but as a permanent condition, which requires constant monitoring, periodic reassessment, and updated testing in the face of changes in input data, application contexts and the AI -models

themselves, especially in the case of adaptive systems or systems based on *online machine learning*, understood here as algorithms that update their parameters with the arrival of new data, rather than "online learning" in the sense of remote or distance education.

Dynamic auditing can be understood as the inspection that accompanies AI systems over time. It is based on three main ideas. The first is the need to periodically review systems, through continuous assessment cycles and updates with new data. The second is the flexibility of audit methods, which must adapt to different practical contexts and legal requirements. The third is maintaining communication channels between auditors, developers, and affected users, ensuring that the assessment considers diverse perspectives and real user experiences (Maughanet al,. 2022).

Practical examples reinforce the urgency of this approach. In automated systems used by large corporations and public agencies, **neutral adjustments to decision criteria can inadvertently generate significant distortions along sensitive attributes such as race, gender, or geographic origin.** The report by the *Electronic Privacy Information Center* (Fergusson, 2023) shows that, even after declared adjustments, third-party automated systems continued to present structural flaws, compromising equity and making it difficult to hold those involved accountable.

The problem becomes particularly visible in recommendation and personalization systems, where rapid adaptation to user behavior amplifies informational asymmetries. Mittelstadt (2016) argues that such systems can compromise informational diversity and, consequently, the cognitive autonomy of users. By filtering content based on profiles and past behavior, these systems may create self-reinforcing informational environments, often described as "algorithmic bubbles", that hinder exposure to diverse perspectives and constrain pluralistic public debate. Although influential, this notion has also been criticized as overly reductive, since empirical evidence for fully isolated "filter bubbles" is mixed. Still, the underlying concern remains relevant: opaque personalization mechanisms can weaken the conditions for informed democratic participation.

The European regulatory framework for high-risk AI systems requires vendors to establish ongoing processes for monitoring system performance after implementation, as described in the delegated regulations that complement the AI Act. These processes include data collection, risk analysis, independent audits, and systematic documentation. **However, primary responsibility for these activities remains with vendors, which undermines the effectiveness of**

the model in the absence of explicit meta-audit mechanisms or mandatory external oversight (European Union, 2024).

This regulatory framework has been criticized for emphasizing self-regulation, especially in technically complex and socially sensitive sectors. As Mökander et al. (2022) point out, the reliance on self-diagnosis by developers themselves undermines the credibility of the governance model and perpetuates structural asymmetries between the companies responsible for the systems and the regulatory bodies. This dynamic weakens effective institutional control and makes it challenging to detect failures promptly, generating additional risks in applications with high social impact.

Institutional and technical challenges stand out among the main barriers to the effective implementation of dynamic audits. On the one hand, many regulatory authorities lack infrastructure, qualified personnel, and timely access to data and models, making it challenging to exercise autonomous and timely oversight powers. On the other hand, there are structural challenges in the design of AI systems themselves, which are not always designed to be auditable from the outset. Rastogi et al. (2023) demonstrate that, even in technically sophisticated environments involving large language models (LLMs), auditability depends on providing specific tools, contextual organization of failures, and active interaction with human auditors. The study highlights that limitations such as the lack of interpretable logs, standardized documentation, and adequate interfaces for exploring algorithmic behaviors hinder the effectiveness of audits, especially those of a dynamic nature.

Despite recent regulatory advances, continuous audit enforcement remains incipient in most jurisdictions. The Ada Lovelace Institute's *Code & Conduct* report (2024) and initiatives such as the SCC (2024) program outline essential guidelines for the institutional structuring of independent audits. However, both lack mandatory legal mechanisms to ensure their systematic implementation. As the Holistic AI report (2023) warns, the lack of regulatory standardization and mandatory requirements for dynamic audits favors an asymmetric scenario, in which only companies with a greater reputation or resources voluntarily adhere to good practices, compromising the regulatory ecosystem's fairness and effectiveness.

In short, dynamic auditing represents a necessary evolution of the algorithmic verification paradigm, aligned with the principles of adaptive justice, continuous surveillance, and iterative correction. Its consolidation, however, will depend on profound institutional reforms, strengthening oversight capabilities and enforcing technical standards that make auditability a legal obli-

gation rather than an optional choice for AI systems with significant societal impact.

Collaborative Ecosystems and Meta-Audit Infrastructure

The evolution of the algorithmic auditing debate underscores that technical procedures alone are insufficient to ensure legitimacy, effectiveness, and fairness in the evaluation of AI systems. In contexts characterized by epistemic asymmetries, institutional inequality, and technological opacity, audits require broader sociotechnical infrastructures capable of integrating diverse social actors, including civil society organizations, universities, regulatory bodies, and independent technical communities.

In this debate, sociotechnical auditing has been discussed as an alternative approach that emphasizes the need to examine not only the technical performance of AI systems but also the institutional, political, and social conditions under which they operate (Lam et al., 2023).

As argued by Lam et al. (2023), sociotechnical audits seek to reveal how algorithmic decisions are shaped by power dynamics and social relations, which requires an interdisciplinary approach, with the active participation of experts in ethics, law, engineering, and social sciences. Such an approach also poses the challenge of including, in the audit processes, the voices and experiences of the communities most directly affected by automated systems, especially those historically marginalized in decision-making arenas. The research by Rastogi et al. (2023) documents the development of *AdaTest++*, an audit tool that combines the strengths of human auditors and generative language models. The study demonstrates that human–LLM collaboration can enhance audit effectiveness by enabling the generation of more diverse tests, the contextual organization of failures, and the iterative inspection of algorithmic behavior.

Among the most advanced proposals to institutionalize this approach, the public or hybrid meta-audit model, advocated by the Ada Lovelace Institute (2024), stands out. The concept of meta-audit refers to the creation of independent mechanisms to monitor the auditors themselves, to ensure that verification processes are conducted ethically, transparently, and technically valid. The *Code &Conduct* report (Ada Lovelace Institute 2024) recommends that AI audit regimes establish independent and transparent structures subject to public scrutiny, involving external actors such as civil society organizations, independent experts, and public institutions to ensure verifiable ethical and technical accountability.

The European Union, through the AI Act, establishes that the conformity assessment function falls to notified bodies, independent conformity assessment organizations designated by Member States and authorized to evaluate whether high-risk AI systems meet the requirements of the Regulation. These entities may, under their responsibility and with the supplier's consent, subcontract parts of the process or use subsidiaries, as provided for in Article 33 (European Union, 2024). This subcontracting, however, is subject to strict criteria and does not constitute an institutionalized form of cooperation with external independent or public entities. On the other hand, Canada has advanced toward a more distributed governance model through the SCC accreditation program, which imposes criteria of accountability, independence, and transparency on certifying entities (SCC, 2024) and proposes the creation of technical public audit centers. In turn, the CSA has promoted peer audit projects in partnership with universities and applied research centers (CSA, 2024).

Although there is no comprehensive federal regulation on artificial intelligence in the United States, several decentralized initiatives have been implemented. The National Institute of Standards and Technology (NIST) developed the *AI -Risk Management Framework* (AI RMF 1.0), a voluntary guide released in January 2023, which aims to help organizations manage risks associated with AI by promoting trustworthy and responsible systems. This framework is structured into four main functions: Govern, Map, Measure, and Manage (NIST, 2023). In parallel, the Federal Trade Commission (FTC) has taken steps to ensure the responsible use of AI. Per the *Office of Management and Budget Memorandum M-24-10*, the FTC has developed a compliance plan emphasizing transparency, accountability, and a focus on public benefit in using AI tools (FTC, 2024).

Among the emerging actors, the IAAA stands out, seeking to consolidate a global ethical-normative repertoire for AI audits, promote the training of auditors, and create an international public certification registry (IAAA, 2024). Holistic AI has also contributed technical proposals on audits of large language models, emphasizing good operational practices, the documentation of procedures, and the responsible use of metrics in complex environments (Holistic AI, 2023).

Establishing a global meta-audit infrastructure is therefore both a technical and a political imperative. It requires strengthening the institutional capacity of states and independent organizations, creating transparent and interoperable protocols, and institutionalizing mechanisms for social participation in algorithmic governance. As Vecchione, Levy, and Barocas (2021) argue,

algorithmic audits should be evaluated not only for their technical robustness but also for their capacity to empower those affected by automated systems to contest their outcomes, an endeavor that presupposes structures open to public criticism and continuous review. Overcoming the closed self-regulation of private audits is essential to building and maintaining public trust, ensuring that AI systems operate in accordance with democratic values and fundamental rights.

Limitations

This study is based on a normative and documentary analysis and does not include primary empirical data. The literature reviewed highlights structural constraints that limit deeper empirical investigation, such as the lack of standardized documentation and comparable audit records (Costanza-Chock et al., 2022), the technical opacity of systems subject to audit (Mittelstadt, 2016), and restricted public access to models, logs, and evidence required for detailed analysis (Fergusson, 2023). In line with Schiff, Kelley, and Camacho (2024), this study acknowledges that the absence of interviews or fieldwork limits the examination of real audit practices and potential dynamics of institutional capture. These constraints reflect broader challenges in the emerging field of AI audit research and should be addressed in future studies.

Conclusion

This study addressed the central question, "Should AI auditors be audited?" Rather than a merely technical or regulatory issue, this study argues that the question reflects deeper concerns about the power structures shaping contemporary algorithmic governance. The analysis has shown that, although conceived as accountability instruments, AI audits often operate in ecosystems marked by epistemic asymmetries, conflicts of interest, and opaque self-regulation.

The discussion identified structural flaws across multiple dimensions: technical, when audits privilege quantitative metrics and neglect social impacts; institutional, when auditors depend economically on those being audited; and regulatory, due to weak external oversight mechanisms, including the European model centered on notified bodies. Without ongoing public

supervision, reliance on self-assessment and certification risks undermining the legitimacy and effectiveness of audits. These findings should also be interpreted considering the methodological constraints discussed earlier, particularly the limited availability of standardized documentation and the opacity of systems subject to audit, which restrict the empirical visibility of real auditing practices.

To confront these limitations, the study proposes the creation of permanent public or hybrid meta-audit structures grounded in transparency, integrity, and institutional independence. Such a reconfiguration entails inter-institutional cooperation among regulatory agencies, universities, civil society organizations, and technical communities, fostering democratic forms of supervision and contestation.

Ultimately, auditing should be understood not merely as a technical verification mechanism but as a political and epistemic practice that determines who defines standards and interprets social effects. Ensuring independent audits of auditors is therefore both a demand of epistemic justice and a democratic safeguard. Meta-audits, by institutionalizing independent and participatory oversight, are essential to protect fundamental rights and reinforce the democratic legitimacy of algorithmic governance.

References

Ada Lovelace Institute. *Code & Conduct: How to Create Third-Party Auditing Regimes for AI*. Londres, 2024. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2024/06/Ada-Lovelace-Institute-Code-and-conduct-FINAL-1906.pdf.

Ada Lovelace Institute; AI Now Institute; Open Government Partnership. *Algorithmic accountability for the public sector: learning from the first wave of policy implementation*. London: Ada Lovelace Institute, 2021. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

AI Now Institute. *Algorithmic Accountability: Moving Beyond Audits*. 2023. Available at: https://ainowinstitute.org/publications/algorithmic-accountability.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. ProPublica*. New York, 2016a.

Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. ProPublica*. New York, 2016b. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Bandy, Jack. *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits*. arXiv, 2021. Available at: https://arxiv.org/abs/2102.04256.

Bouwmeester, M. *System Failure in the Digital Welfare State: Exploring Parliamentary and Judicial Control in the Dutch Childcare Benefits Scandal*. University of Groningen, 2023. Available at: https://research.rug.nl/en/publications/system-failure-in-the-digital-welfare-state-exploring-parliamentar.

Costanza-Chock, S.; Raji, I. D.; Buolamwini, J. *Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem*. In: ACM Conference on Fairness, Accountability and Transparency – FAccT '22, 2022, New York. *Proceedings*. New York: Association for Computing Machinery (ACM), 2022. p. 1571–1583. Available at: http://dx.doi.org/10.1145/3531146.3533213.

Canadian Standards Association (CSA). *Advancing responsible AI: SCC launches accreditation program for AI management systems*. 2025. Available at: https://www.scc.ca/en/news-events/news/launch-ai-governance-accreditation.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review,* 14(1), 57–74. Available: https://doi.org/10.2307/258191.

European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Official Journal of the European Union, L 168, 12.7.2024, p. 1–138. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

European Commission. *Commission Delegated Regulation (EU) 2024/1773 of 1 March 2024 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council as regards the request for exemption from the obligation to create an advertisement repository by providers of very large online platforms*. Official Journal of the European Union, L 177, 26.6.2024. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401773.

Federal Trade Commission (FTC). *Artificial intelligence compliance plan*. Oct. 2024. Available at: https://www.ftc.gov/ai.

Fergusson, G. *Outsourced automated: how AI companies have taken over government decision-making*. Electronic Privacy Information Center (EPIC), 2023. Available at: https://epic.org/wp-content/uploads/2023/09/FINAL-EPIC-Outsourced-Automated-Report-w-Appendix-Updated-9.26.23.pdf.

Holistic AI. *LLM auditing guide: what it is, why it's necessary, and how to execute it.* Oct. 11, 2023. Available at: https://www.holisticai.com/papers/llm-auditing-guide.

Institute of Electrical and Electronics Engineers (IEEE). *Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems.* New York: IEEE, 2023. Available at: https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf.

International Association of Algorithmic Auditors. *IAAA AI auditing definitions.* [S.l.]: IAAA, Oct. 2024. Available at: https://iaaa-algorithmicauditors.org/wp-content/uploads/2024/10/IAAA-AI-Auditing-Definitions-1.pdf.

International Association of Algorithmic Auditors. *IAAA code of conduct.* [S.l.]: IAAA, Oct. 2024. Available at: https://iaaa-algorithmicauditors.org/wp-content/uploads/2024/10/IAAA_Code_of_Conduct-2.pdf.

ISO/IEC. *ISO/IEC 42001:2023 – Artificial intelligence — management system*. International Organization for Standardization, 2023. Available at: https://www.iso.org/standard/81230.html.

Lam, M. S. et al. *Sociotechnical audits: broadening the algorithm auditing lens to investigate targeted advertising*. *Proceedings of the ACM on Human-Computer Interaction*, v. 7, n. CSCW2, Article 360, p. 1–37, Oct. 2023. DOI: https://doi.org/10.1145/3610209.

Maughan, K.; Ngong, I. C.; Near, J. P. *Prediction sensitivity: continual audit of counterfactual fairness in deployed classifiers.* 2022. Available at: https://arxiv.org/pdf/2202.04504.

Mittelstadt, B. *Auditing for transparency in content personalization systems. International Journal of Communication*, 2016. Available: https://www.researchgate.net/publication/309136069.

Mökander, J.; Axente, M.; Casolari, F.; Floridi, L. *Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Minds & Machines*, v. 32, p. 241–268, 2022. Available: https://doi.org/10.1007/s11023-021-09577-4.

National Institute of Standards and Technology (NIST). *AI risk management framework (AI RMF 1.0)*. Jan. 2023. Available at: https://www.nist.gov/itl/ai-risk-management-framework.

Rastogi, C.; Sugimoto, C. R.; Maddox, T. et al. *Supporting human-AI collaboration in auditing LLMs with LLMs.* 2023. Available at: https://arxiv.org/pdf/2304.09991.

Standards Council of Canada (SCC). *AI governance accreditation program*. 2024. Available at: https://www.scc.ca/en/news-events/news/launch-ai-governance-accreditation.

Schiff, D. S.; Kelley, S.; Camacho Ibáñez, J. *The emergence of artificial intelligence ethics auditing. Big Data & Society*, 2024. Available at: https://journals.sagepub.com/doi/10.1177/20539517241299732.

Stigler, G. J. *The theory of economic regulation. The Bell Journal of Economics and Management Science*, v. 2, n. 1, p. 3–21, 1971. Available: https://doi.org/10.2307/3003160.

Stilgoe, J. *We need a Weizenbaum test for AI. Science*, v. 381, n. 6658, p. 770–771, 2023. Available: https://www.science.org/doi/10.1126/science.adk0176.

Treleaven, P.; Galpin, V.; Zardousti, P. *Algorithmic auditing: AI accountability in public sector applications.* SSRN, 2021. Available at: https://ssrn.com/abstract=3889612.

Vecchione, B.; Levy, K.; Barocas, S. *Algorithmic auditing and social justice: lessons from the history of audit studies.* In: *Equity and Access in Algorithms, Mechanisms, and Optimization – EAAMO '21*, In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). New York: ACM, 2021. p. 1–9. Available: http://dx.doi.org/10.1145/3465416.3483294.

Wörsdörfer, M. *The E.U.'s Artificial Intelligence Act: an ordoliberal assessment.* Forthcoming in: *AI and Ethics*, 2023. Available at: https://ssrn.com/abstract=4544276 or http://dx.doi.org/10.2139/ssrn.4544276.

Citizens Infrastructures as a Way to Govern AI's Power to Shape our Shared Representations

How to Make Sure Auditing Institutions Are Aligned with Values and Societal Expectations?

Chiara Marcoccia

Abstract: *The example of generative AI models (genAI) for the production of images points to a serious gap in AI auditing. While genAI models deployed on digital platforms for information and communication (ICTs) have an increasing and particularly persuasive influence on users' representations of social events, groups and dynamics; users themselves have no effective way of reclaiming power over their shared representations, governing when and how AI can shape their vision of the world. I argue that auditing institutions, which assess the alignment of AI systems with a set of previously defined societal expectations, are not able to keep up with the challenge of aligning with societal values and expectations genAI models and recommender systems that are evolving at an incredible speed and through a non-disclosed frequency of retrainings.*

In this paper the author argues that, to fill this gap, the auditing pipeline needs to leverage users themselves, in the form of an intermediary structure that will enable a dynamic feedback loop to keep up with the evolving impact of recommender systems and genAI models on ICTs, across the individual, the community and the societal level. I propose the form of this structure by drawing from the social sciences and the concept of citizens infrastructures. These are an enduring structure to leverage existing citizen engagement into the socio-technical networks that interact with and are impacted by the deployment of a given AI system. In this paper two examples of such infrastructures are being discussed and an argument is made for their potential to govern the power of AI over our shared representations, before drawing the implications for how the very concept of AI auditing needs to evolve.

Keywords: *AI governance; Algorithmic auditing; Digital ecosystems; Artificial Intelligence; Generative AI; Recommender systems*

Introduction

AI-generated images are one of the most glaring failures of auditing institutions in AI governance. Companies have been cultivating artificial image synthesis since the development of Generative Adversarial Networks (GAN) in 2014 (Elgammal et al. 2017), but since OpenAI's image generator DALL-E, that generates images from text, was opened to the public in 2022 (OpenAI 2022), AI-generated images have flooded all public information and communication platforms (ICP) (Bond 2024). Presently, any internet image search displays some kind of AI-generated images, not all of which are immediately recognizable as such (Gangadharbatla 2022).

Yet, the public response to the large diffusion of AI-generated images is not a positive one[1]. Ratings of comfortableness towards and estimated capability of AI producing artistic images is very low compared to other areas of application of AI technology (Schepman et al. 2020), and the appreciation of AI-generated images is much lower than that of human crafted ones (Bellaiche et al. 2023; Hong et al. 2019; Chamberlain et al. 2017). This points out a misalignment between the industry and the public, but one that doesn't currently have a suitable reflection upon effective governance, as there is no adapted institution to assess, support and address this misalignment.

However, what is at stake behind this representation, are our very own shared representations. Indeed, ICPs where AI-generated images are taking up more and more space actually convey social and cultural cues that influence the construction of our shared beliefs and behaviours (Glickman et al. 2025; Ashkinaze et al. 2024; Pappalardo et al. 2024; Hoang et al. 2023; Sîrbu et al. 2019; Bail et al. 2018). Consequently, AI-generated contents themselves are already becoming a non-neglectable factor in our perception and appreciation

1 A *caveat* to be introduced here, as most data available on the public's reaction to AI-generated images is culturally situated, and, while significant in its proportion for the purpose of calling out on current auditing processes in AI governance, doesn't necessarily represent all attitudes around the globe equally. See for example Wu et al. (2019) about the different perception of AI-generated art among American test subjects as compared to Chinese.

of reality, influencing our opinions (Floridi 2024), our tastes (Wu et al. 2019) and our values (Brinkmann 2023), on a personal, interpersonal and societal level, with the potential to shape our perception of the self, our attitude towards others and our shared representations of the future, of democracy and of social identities (Floridi et al. 2024; Burrell et al. 2023; Ovadya et al. 2023). For such high stakes and wide-ranging consequences, we have no matching institutional levers: we lack infrastructures to connect between the big-tech industry, national and international regulatory instances, and the public. An issue for which I try to suggest a workable solution in this paper.

In fact, in the first part of my paper I show that the auditing institutions that have failed to govern the development of AI-generated images in alignment with societal values and expectations, have done so because of the lack of citizen infrastructures. These can take the shape of digital platforms, community forums, shared skills, information networks, or lived environments (AbdouMaliq 2004) and aim at supporting and enabling citizen-led and enduring public engagement (Gabrys 2021). Their specificity is that they often originate from everyday practices and in the very spaces where citizens interact (Livingstone et al. 2010).

I make the claim that such infrastructures can be particularly well-suited and effective for the purpose of enabling and supporting citizen engagement in auditing institutions involved in AI governance, effectively and durably aligning those institutions with social values and expectations. Indeed, a successful infrastructure connecting citizens and auditing institutions must allow governance issues to be formulated and discussed by citizens *contextually*, in a way that is coherent and organic with the live perception of these issues and with the experience-led evaluation of their consequences.

For this reason, citizens infrastructures appear as an interesting solution: they emerge from and in turn provide structure to the living environment or the everyday activities that are the very reason for citizens' perceptions, valuing and expectations regarding the issue at hand. For AI regulation, this embedding of the infrastructure into the living experience would mean that the infrastructure is part of the citizens' consistent interaction with AI-generated content, which mostly happens through ICPs. This naturally points toward a kind of digital infrastructure, whose characteristics I outline in the last part of my paper.

In fact, digital infrastructures have become more and more promising as technology advances (Barns 2016), but they have an especially interesting potential today to address the issue of aligning social values and expectations

towards AI with auditing institutions. Indeed, the live context in which citizens encounter AI-generated content are ICPs, and these already host community forums, shared skills and information networks, which are the channels through which citizens infrastructures are commonly organized and maintained.

Hence, a digital citizens infrastructure appears to be an interesting option to govern AI's power to shape shared representations in alignment with social values and societal expectations. In fact, such a digital infrastructure would close the gap between consuming and governing AI-generated content, insofar as the space where AI-generated content has an impact on citizens' life and perception — ICPs — would also be the space where citizens can have an impact on the generation and diffusion of artificial content. This brings me to conclude my paper on the idea that it is possible to counter the trend that makes out people to be more passive as AI becomes more pervasive, and stress on the contrary how human-AI interaction can be a chance to enhance governance, interaction and decision-making.

The Misalignment Between AI Deployment, Societal Values and Expectations

Over-Promising and Under-Delivering

When we look at the expectations surrounding AI, we are faced with a deep contrast between high hopes and actual user experiences (Kinney et al. 2024). Indeed, the computing power unlocked by AI technology opens up the potential for many promising applications in the fields of healthcare (Lamberti et al. 2019; Aung et al. 2021; Nguyen et al. 2023), business (Pathak et al. 2010; Chen et al. 2004), government (Pi 2021), education (Tlili et al. 2023; Zawacki-Richter 2019), and justice (Westermann et al. 2024).

In fact, the machine learning technology that powers AI systems enables them to extract patterns from huge amount of data (the so called big data) and use them to interpret new data (LeCun et al. 2015), potentially offering a tool to reduce human work-load (Zysman et al. 2024), support decision-making (Turki et al. 2024; Shin et al. 2024), democratize access to knowledge (Mehandry et al. 2025; Conryut et al. 2015) and enhance task efficiency (Autor et al. 2024; Crafts 2021; Ernst et al. 2019).

As a result, people expect AI technology to take up an important societal role, improving institutions and enhancing citizens' quality of life (Kinney et al. 2024: 1). Indeed, a global survey published in the 2025 United Nations' Human Development Report indicates that two thirds of people across countries expect to use AI in education, health and work within one year (UNDP 2025: 4). In fact, in the public sector, AI technology could be an opportunity to fuel innovative public services improving services to citizen queries and enhancing predictive capability for decision-making thanks to its capacity to analyze high-dimensional as well as unstructured data (Eggers et al. 2017).

However, when we look at current AI uses, not only do they not live up to expectations, but they point to a deeper problem in the overall direction of the development of this technology. On the one hand, current AI systems don't offer enough transparency in their reasoning (Burrell 2016) and trustworthiness in their sources (Castelvecchi 2016; Kaur et al. 2022) to support widespread applications in healthcare, business, government, education, and justice (Bender et al. 2021). On the other hand, current AI deployment actually causes at least as many (if not more) harms as benefits for citizens.

First of all, the capacity of AI to reduce human work-load is dependent on increasing hidden (and sometimes exploitative: see Capraro et al. 2024) human labor (Muldoon et al. 2024), such as data labelling and annotation (Acemoğlu et al. 2019). Secondly, evidence shows that current use of AI to democratize knowledge and support decision-making actually also has a negative social impact, especially on marginalized groups (Sartori et al. 2022), insofar as it reproduces harmful biases (Mehrabi et al. 2021), spreads misinformation (Allcott et al. 2019) and leads to unfair decision-taking (Crawford 2021; Angwin et al. 2016; Dastin 2018).

As a consequence, there is a mismatch between society's high expectations for AI's beneficial applications and the actual possibilities offered by current AI systems. Crucially, the problem seems to be more than just a question of technical advancement (UNDP 2025: 48): it seems to be part of a broader misalignment between societal expectations and the current direction of AI development.

In fact, while the techno-solutionist narrative would have biased AI outputs be just a sign of a still underdeveloped technology to be fixed by future refinement (Altman 2024), research shows that it's the very development strategy of AI systems that produces biased outcomes. Indeed, the datasets used to train AI systems such as Large Language Models (LLMs) are built from data coming with overwhelming majority from Western countries (Rahman et al.

2024), and consequently carry a certain culturally situated value system (Atari et al. 2025).

Moreover, training data are taken from the internet, and subsequently reflect the over-representation of English speakers (McIntosh et al. 2024), young people and people from developed countries among internet users (World Bank 2018), but also the over-representation of men among the writers of online forums used as training data, like for instance Reddit (Pew 2018) and Wikipedia (Barera 2020).

For this reason, if we want AI systems to reflect societal values such as inclusiveness, diversity and fairness, we can't rely on technical advancement (Winner 2017), but we have to purposefully include such values in the development and deployment of AI technology (UNDP 2025: 103–104; 119–121). This paper suggests a way how. First, I propose to identify the difficulties for the inclusion of societal values in AI development.

The Alignment Problem

The difficulty to design AI systems to reflect societal values has come to be known as the alignment problem (Russell 2020). This consists on the one hand in a normative and on the other hand in a technical difficulty: how to frame the set of values that AI should be aligned with, and how to successfully incorporate them in AI systems (Gabriel 2020).

The latter is only partially a task for AI engineers to solve, for it requires we identify what the goals of AI are, in order to identify how they can meet our own (Kinney et al. 2024). This is not an anthropomorphization of AI, which would attribute goals to the AI system and question its intrinsic moral values. On the contrary, it's about understanding the technology behind the AI system, what it was developed for and what it can do (McCoy et al. 2023).

Indeed, the two questions are interconnected, and to successfully incorporate societal values into AI systems, we have to identify the values that an AI system should be aligned with, *given the tasks that the system has been developed for* (McCoy et al. 2023). For instance, when addressing the alignment problem in the case of LLMs, if we take into account that the technology on which they are based is designed to predict the next word given a context (Bender et al. 2021), we can identify *what* values are concerned by their computing strategies and *how* they are mobilized in their behaviors (Millière 2023), allowing us to determine the relevant discussions and measures to be undertaken.

This requires an understanding of AI technology that is often confined to techno-economical *niches* that are not accessible to the public and sometimes not even to regulating institutions (Qi et al. 2024; Wachter et al. 2024), which concentrates the power over the social transformations associated with AI in a few hands and away from citizens, fueling the misalignment (UNDP 2025: 137–139).

But solving the alignment problem also requires an assessment of citizens' attitudes and positioning towards specific societal values and their concrete applications in AI development (UNDP 2025: 3–4). Indeed, the alignment problem is also a problem of operationalizing human agency within the development and deployment of AI technology: the values with which AI is aligned must emerge by democratic and deliberate expression of values and expectations in order for AI to reflect societal values (Innerarity 2024).

Hence, to address the misalignment between AI and societal values and expectations we need infrastructures that are apt to bring together societal auditing with technical knowledge about AI development. For this reason we will now focus on the infrastructures currently available for citizens to participate in AI governance, to identify the gap between citizens and auditing institutions and map the way for the development of new infrastructures.

The Problem with Current Auditing Institutions

Defining AI Auditing

Auditing AI means assessing the alignment of given AI systems with a set of previously defined expectations (Birhane et al. 2024). Who defines these expectations and whether they reflect societal values varies among different AI auditing institutions. These include the media, law firms, regulators, consulting agencies and academia (Bandy 2021). The goals for which each of these stakeholders assess AI systems differ, and, consequently, so do the standards they define as expectations for the evaluation of the AI system (Costanza-Chock et al. 2022).

Indeed, current auditing institutions either look to comply with mandatory requirements and minimize corporate liability (internal auditing), or to identify and minimize the harms impacting users (Birhane et al. 2024). Hence the auditing process, which could offer a means for actors developing and regulating AI to align expectations, actually offers no common ground for differ-

ent stakeholders to exchange assessments, only entertaining the distinction in their roles.

The implication is that the policy applications of these audits are far from systematic (Birhane et al. 2024): AI assessments often don't lead to proactive re-design, products recall, voluntary corporate action or broader governmental regulation; firstly, because of their heterogeneous nature (ALI 2021), and, secondly, because of their lack of a structured articulation into the social fabric around AI development (Vecchione et al. 2021).

Indeed, firstly, socio-political repercussions of AI audit are impeded by the lack of clearly defined auditing standards (Costanza-Chock et al. 2022), for it's those standards that drive not only the assessment of AI systems, but also their development. In fact, standards are 'consensus-based and agreed-upon ways of doing things' (Dignum 2022), that auditing helps to define and that in AI development translate into the minimum specification on how to carry out a process in alignment with socially-acknowledged expectations. This means that the standards that should be defined by auditing institutions are the same standards that are needed to guide AI development (Raji et al. 2020).

Secondly, transformative actions don't come out of AI audit because the auditing process lacks a direct involvement of the stakeholders that are most likely to be harmed by AI systems (Costanza-Chock et al. 2022). Indeed, institutions that carry out AI auditing lack a structured articulation into the social ramifications of AI development and deployment: the channel between tech-developing companies and users is severed (UNDP 2025: 137–139), preventing auditing standards to be informed by practical expectations and societal values, and limiting the active and democratic determination of these standards, which is an essential requirement for meaningful and impactful auditing (Innerarity 2024).

One example of these stalling effects of auditing institutions in their current form is the massive deployment of image-generating AI systems (Gangadharbatla 2022). Because impacted parties are not directly present in the elaboration of the auditing process, auditing institutions lack the focus and the pervasiveness to identify and address necessary changes in the regulation and the development of image-generating technology and translate them into auditing standards (Luccioni et al. 2023). Indeed, issues regarding privacy, copyrights and discrimination are central to a fair and trustworthy development of image-generating AI technology (Kaur et al. 2022), but auditing institutions are not sufficiently integrated into the societal adoption of and response to this

technology to meaningfully and effectively mediate the dialogue between the different stakeholders (Costanza-Chock et al. 2024).

For this reason, governing AI in alignment with societal values and expectations requires a new kind of auditing institution, apt to build meaningful and impactful infrastructures between auditing institutions and the different stakeholders affected by AI development. To define such a structure, we will first identify the role that it needs to have in AI governance.

Why Add Infrastructures?

The need for AI audit is widely advocated for in the name of AI (or algorithmic) accountability (Raji et al. 2020; ALI 2021; Bandy 2021; Costanza-Chock et al. 2022; Raji et al. 2022; Birhane et al. 2024). Indeed, the goal of AI audit is to provide an account of the way that a given AI system exercises its power in society (Bandy 2021: 74:3), meaning to give an account of the AI system's performance as well as of its social impact.

For this reason Wieringa (2020) talks about a 'networked account' for what is essentially a 'socio-technical system' (Wieringa 2020): AI systems are deployed within social networks, as for instance social media (Knott et al. 2021) and online retail platforms (Lee et al. 2014), and have network effects, like the spreading of information and the success of products (Pappalardo et al. 2024); therefore, the identification of the network effects of AI deployment and the understanding of their relation to specific technical features of the given AI system (Fabbri et al. 2022; Donnelly et al. 2021) is essential for the design of action-informative AI audit (Birhane 2024).

This is an added requirement to those successfully covered by already developed benchmarking systems (Liu et al. 2024) because it takes into account the unprecedented characteristic of AI systems: namely the fact that they evolve through their deployment (Martínez et al. 2023). Indeed, AI systems update their behaviors in response to their interaction with users: thus recommenders (Isinkaye et al. 2015) update their recommendations to align with users' preferences (Piao et al. 2023; Mansoury et al. 2020; Jiang et al. 2019), and Large Language Models (LLMs) integrate the new data generated by the interaction with users into their following outputs (Hataya et al. 2023). It's the process commonly known as feedback loop (Jiang et al. 2019; Sun et al. 2019; Pedreschi et al 2024).

Furthermore, this process is augmented by the fact that AI systems are re-trained over time, to integrate new sets of data (Shumailov et al. 2024; Pe-

dreschi et al. 2024). As a consequence, benchmarks that are defined on the data fed into the AI system at time t_1 won't necessarily match the characteristics of the data added at time t_2.

For instance, let's look at the benchmarks defined to account for the trustworthiness of a LLM system (Liu et al. 2024; Huang et al. 2024), namely reliability (the information is accurate and the system does not hallucinate), safety (the outputs are not violent, guarantee privacy and do not cause danger for the user), fairness (the reasoning is not biased towards certain communities and the outputs are not discriminatory), resistance to misuse (the system cannot be used to cause harm and protects minors), explainability (the reasoning is transparent and the outputs interpretable), alignment with social norms (it does not go against universally acknowledged social values) and robustness (typos, grammatical errors and repetitions in the prompts do not cause inaccurate outputs).

The specific data that need to be labeled as problematic in accordance with these benchmarks are defined from the dataset on which the LLM was built, and they will change with the usage that users make of the LLM. Bender et al. (2021) already noted that the list of words related to sex that was chosen to filter out offensive, violent and pornographic data from the training data of ChatGPT-2 actually reduced the influence of many forums built by or for the LGBTQ+ community, because it contained slurs reclaimed by certain marginalized groups (Bender et al. 2021).

This is where infrastructures come into play to align auditing institutions with societal values and expectations. If benchmarks, and especially the concrete application of benchmarks, evolve with use, then users need to be integrated in the process of continually keeping those benchmarks up-to-date. To be more specific, it's the very structure of the auditing system that has to be updated to match its role in ensuring accountability, given the specificities of AI technology (Pasquale 2019). Indeed, AI technology itself makes traditional auditing difficult: technology advances with considerable speed (Wu et al. 2025), models are retrained on new data (Zhu et al. 2024) and each system evolves with usage, by learning from the feedback of its users (Glickman et al. 2025; Mansoury et al. 2020), but also by tracking and analysing their behaviors (Boeker et al. 2022). For these reasons, audits can no longer be confined to their role of evaluating the compliance with previously established standards and benchmarks, as they did with softwares that didn't learn from interaction (Vecchione et al. 2021; Birhane et al. 2024). On the contrary, they need to take on an active

role in informing standardization and benchmarking institutions as well as the mitigation practices from system developers (Birhane et al. 2024).

It's to meet this new requirement that infrastructures are needed to bridge the gap between auditing institutions and expectations arising through actual AI usage. In the next part of this paper, I will present the characteristics of these infrastructures, and argue in favor of developing citizen infrastructures to support AI auditing.

Proposing a Solution: Citizens Infrastructures to Support AI Audit

A Proposition for Governance

The aim of this paper is to propose a workable proposition to build the missing link between citizens and auditing infrastructures that is essential to leverage the potential of AI audits to fuel policy, research and industry initiatives. The idea is to leverage existing citizen engagement (Gabrys 2021) into the socio-technical networks (Wieringa 2020) that interact with and are impacted by the deployment of a given AI system (Pedreschi et al. 2024, Pappalardo et al. 2024).

It is not about distributing data-collection tasks (Gabrys 2021: 88) and exploiting consumers for technology development, but rather about leveraging what is currently the most viable and transformative way of incorporating, on the one hand, democratic engagement into the auditing processes, and, on the other, auditing institutions into the social networks impacted by AI (Gabrys 2021: 89).

In fact, citizens infrastructures are citizen-led network practices that take place in everyday spaces and activities (AbdouMaliq 2004), such as interaction with AI powered systems on Very Large Platforms (VLOPs), but that have a rare enduring quality among traditional participatory infrastructures (Gabrys 2021: 88). For this reason, they make a particularly good candidate for bridging the divide between auditing institutions and the socially impacting and time-transformative nature of the AI systems they aim at assessing.

The concept of citizens infrastructures comes from data collection practices in climate studies (Gabrys 2021). It describes when communities that are impacted by and have an impact on a climate-related phenomenon, rather than passively providing data to institutions in the form of surveys for instance, take an active and network-organised role in the data collection, organising feedback, participating in the measurements, monitoring the territory.

Examples of these sustainable practices and of their role in making AI audits actually impactful for the industry and informative for regulation practices can be found in the Open Data Science Initiative (Lawrence 2014), which consists in a users-organized space for open-access development of new analysis methodologies available 'as widely and rapidly as possible with as few conditions on their use as possible' (Lawrence 2014), addressed to the data science community, as well as to commercial, scientific and medical institutions, with the aim of achieving 'a balance between data sharing for societal benefit and the right of an individual to own their data' (Lawrence 2014).

This initiative leverages spaces that are already invested by its users: GitHub, Jupyter Notebook and Reddit; and directs initiatives towards defining AI accountability in the domain of data usage. If relayed by auditing institutions, such a space could become not only an informative but also a directive structure for the design of audits that are fit to support policy and industry action (Capraro et al. 2024). Auditing AI is about assessing the alignment of AI systems with a set of previously defined expectations, however, to serve their purpose, these need to be defined, firstly, relatively to what AI is used for (which means through consultation of the public), and, secondly, relatively to what AI has been developed for (which means with some technical knowledge of how AI is designed and how it works), which indicatives like Open Data Science allow.

Notably, this prevents the characteristic weaknesses of traditional attempts to involve citizens in AI governance, which often don't encounter sufficient participation or continuous engagement (Lahdili et al. 2024; Sieber et al. 2024). Indeed, citizen infrastructures emerge from already rooted practices (Livingstone et al. 2010), which gives them a unique strength within the landscape of available means of citizen engagement with AI development to support the alignment with societal values and expectations (Wilson 2022).

This rootedness in existing practices gives citizen infrastructures a continuity and legitimacy that top-down participatory initiatives often lack. Rather than requiring the creation of new frameworks for engagement, these infrastructures build on collective habits of knowledge production, discussion, and critique that are already active within technical communities and user networks. As such, they are not only more likely to sustain participation over time, but also better positioned to translate between technical discourse and social concerns. This embedded quality can enable them to support the development of auditing standards that are both technically informed and socially respon-

sive, reinforcing the role of auditing as a site of alignment between AI systems and evolving public expectations.

Moreover, a citizen infrastructure modelled after the Open Data Science Initiative could ensure continuity between mitigation actions and technology developments, effectively tackling the challenge of keeping up with the evolution of AI systems, not only following industry developments (Shin et al. 2023), but also as a consequence of the feedback loop between users and AI (Pedreschi et al. 2024), and following possible re-trainings of already available generative AI and recommender systems (Shumailov et al. 2024).

As such, structures of this kind could act as successful infrastructures bridging between citizens and auditing institutions, however, there is one more reason why citizens infrastructures are an attractive solution to the problem of aligning auditing institutions with societal values and expectations. By guaranteeing accountability, AI audits support our governance over AI's power in society (Bandy 2021: 74:3), and integrating citizens infrastructures to the process can elevate this governance.

A Way to Claim Agency over the Power of AI to Shape Our Shared Representations, Values and Behaviors

AI auditing assesses the power of AI technology and of its deployment over society (UNDP 2025: 136–147), which governance strives to align with values and expectations that citizens actively decide upon (UNDP 1997). This is what powers the idea of creating an infrastructure through which AI auditing can be leveraged for governance. On the one hand, citizens infrastructures can potentially support the integration of auditing institutions into the circular evolution of AI with users, but, on the other, they can also add to auditing processes the space for active deliberation that is necessary to make auditing a means of exercising governance (Kuziemski et al. 2020).

Indeed, interaction with AI systems has an impact on the evolution over time of human behaviors (Pappalardo et al. 2024), as can be seen in the way that recommender systems influence the tone of social interactions on digital platforms (Ovadya et al. 2023), and chatbots can improve cooperation within a human team (Shirado and Christakis 2020). But, in the longer term, interaction with AI also has an influence on the evolution of cultural products (Fogarty et al. 2023) and values (Brinkmann 2023) as well as on the evolution of social norms (UNDP 2025: 61), ultimately contributing to defining human values, behaviors and representations.

For these reasons, it is essential that we, as citizens, are able to exercise governance over the power of AI to shape our very own behaviors, representations and values. And, in fact, the integration of citizens infrastructures to institutions auditing AI accountability can successfully be a step in the direction of such a governance (Stilgoe 2024), by providing a space for citizens to contribute to the assessment of AI, the mitigation of its risks and the fostering of its potential, but also by bridging the gap between regulating institutions and AI users (Dave 2019; McQuillan 2018).

Crucially, citizens infrastructures could bridge the gap between usage and development of AI without going through the established lobbies to reach the AI industry (Livingstone et al. 2010). Indeed, if auditing institutions can incorporate citizen-led governance initiatives such as the Open Data Science Initiative, then the auditing process can become a channel through which citizens, by informing the standards set and assessed by auditing institutions, can have a real space in the AI industry, and develop an alternative kind of AI auditing, in contrast with profit-driven certifications (Kuziemski et al. 2020).

Indeed, the Open Data Science initiative is a kind of citizens infrastructure that emerges from the spaces and activities that are characteristic of digital platforms in general: however, one other option is to incorporate as citizens institutions, spaces and activities that emerge from the integration of AI itself into digital platforms, leveraging not only user networks but also user interaction with AI and the way it resonates within these networks. An example of this is something that has recently started to be explored in the literature (see Hayashi et al. 2024 for discussion of the potentials of decentralized platforms; but also Dotan et al. 2023 for discussion of their weaknesses) under the name of decentralized platforms. In this type of platform, each AI system evolves within the community of interest that it interacts with, and, if this structure had to be leveraged for auditing, the AI system could be audited within that same community, for the purpose that it was built for and for the usage that it evolves with. Because of this, auditing standards would be closely tailored and constantly evolving, matching the pace of user-system feedback loops. Moreover, on such a platform, AI systems could also be audited across different communities, insofar as, on decentralized platforms, whenever a user from one community wants to connect with something in another community (item, person, opinion, information) and it is up to the AI system to find a so-called bridge for that connection (Ovadya et al. 2023; Stray et al. 2023). This way, both the expectations of the users and the actual evolution of the network become variables that can be taken into account, enriching the auditing process with standards

for the network effects of user-AI interaction (Fabbri et al. 2022; Donnelly et al. 2021), rather than remaining limited to the individual effects of a given AI system on a given user.

This is particularly important because the power of AI over our shared representations comes mostly from network effects (Alcott et al. 2019). Indeed, although our individual interaction with AI-generated content or with a recommender system may transmit ideas, knowledge or bias to us (Ashery et al. 2025; Colther et al. 2024; Peralta et al. 2021; Mansoury et al. 2020; Sîrbu et al. 2019; Sun et al. 2019; Zhao et al. 2019), what makes AI so increasingly powerful is the fact that that interaction and its effects have so to say ripple effects onto all of the networks that we are a part of (Pansanella et al. 2022a; 2022b). First, because of the network and sharing based nature of the digital platforms where human-AI interactions take place (Wang et al. 2019). Second, because of what Coté et al. (2025) call the recursive nature of AI technology, in which every interaction is always compared to and reinvested in another one. Sometimes explicitly, like in the case of comparative filtering (Sun et al. 2019), but more largely because of the traceability of human-AI interaction (Pedreschi et al. 2024), which allows AI to keep track of all interactions and to turn them into data, but also to perform the next interaction as a kind of projection (LeCun 2019). Indeed, an AI assistant makes an estimation of what the reaction of the user will be to its suggestions or notifications, that it generates accordingly, and it registers the actual reaction of the user as a confirmation or a contradiction of its estimation. It then uses this information to generate the next estimation, making it act as a kind of prediction (Coté et al. 2025). As a consequence, what individual interaction with AI tends to do, is to isolate the user into an autoreferential bubble (*confer* the literature on filter bubbles, echo chambers, etc. See: Del Yang et al. 2023; Srba et al. 2023; Tomlein et al. 2021; Cinus et al. 2022; Aridor et al. 2020; Del Vicario et al. 2016; Pariser 2011), where every interaction appears self-referential and as closely matched as possible to the respective user. While, in fact, the effects of that bubble are always network effects, as shows a very interesting study on music streaming with AI-powered recommenders (Porcaro et al. 2024): each individual is listening to more diverse music thanks to AI recommendations, but the overall audience is all listening to the same tracks, which shows that even when it's all about an individual user, it is always about everyone else too (see also Doshi et al. 2024). For this reason, it is so important to operationalize networks as an intentional infrastructure to govern the power of AI, countering this isolating trend.

In fact, what I propose is to update the auditing system to keep up with the originality of AI technology. To do so, I argue that we should leverage expertise from the social sciences, to identify and integrate into the auditing process networks that could work as citizens infrastructures, solving the problem of the speed and complexity of AI evolution, to make sure auditing institutions are aligned with societal expectations. Indeed, the development of citizens infrastructures to support auditing institutions would not only transform citizen engagement in AI governance, but could also transform the structure of the auditing systems around AI and reinforce auditing institutions themselves, redefining their role and curating the channel that connects them to other stakeholders in society, industry and in regulation.

Citizens infrastructures, by embedding deliberation into the very process of AI auditing, would make it possible for societal values to emerge through continual collective reflection. Rather than auditing AI against static standards, this model allows values to be negotiated and redefined over time, in response to technological developments and shifting social contexts. Auditing, then, becomes a recursive and situated process—one that not only aims to hold AI accountable, but also enables the articulation and iterative reconfiguration of normative frameworks in dialogue with the systems that affect them. We may look back here on the case of AI-generated images, and of the necessity to govern their influence over socially shared representations. An active and iterative confrontation between the stakeholders impacted by this issue is the key here to achieve governance over the power of AI to shape shared representations.

Indeed, by grounding oversight in use and experience, citizens infrastructures respond to the fragmentation between AI development, deployment, and evaluation. For this reason, it would be a strong framework to support further research about how such infrastructures could contribute to establishing a circular model of accountability in which the feedback of users informs the criteria that auditing institutions apply, and in turn shapes how systems are designed and governed. Achieving this continuous flow is of crucial importance, for it would reduce the dependency on intermediaries such as industry lobbies, and instead support the institutionalization of citizen input within audit bodies themselves. In doing so, citizens infrastructures would not merely complement existing mechanisms, but strengthen them, transforming auditing institutions into agents of governance embedded in the social environments they are meant to serve.

Conclusion

In this paper I relayed the criticism that current AI auditing institutions do not or do not often produce an impact on AI industry and regulators, and suggested a solution to leverage AI audit for algorithmic governance: the integration of citizens infrastructures.

I argued in favor of this solution firstly by describing the misalignment between the current AI deployment system and societal values and expectations. I then proceeded to point out how the current auditing system is unable to respond to this misalignment, and what the reasons are for this *impasse*. Finally, I argued in favor of a solution, advocating for citizen infrastructures as action-oriented, enduring and democratic structures, with the potential to transform the role of auditing institutions and their integration into the social ecosystems affected by AI development and deployment.

To conclude, the integration of citizens infrastructures into AI auditing not only offers a response to the limitations of current institutions, but also redefines the epistemic and political conditions under which algorithmic governance takes place. By anchoring auditing processes within the communities affected by algorithmic systems, this model grounds accountability in situated knowledge and collective deliberation. It enables a redistribution of authority, bringing into the governance process a broader range of actors—public interest organizations, local collectives, and individual users—whose perspectives are often marginalized in existing audit frameworks.

Further research is needed to operationalize this model into AI governance: among the foreseeable future steps, there is a need to define the scalability of citizens infrastructures, and to identify and address limitations and failure modes.

What is at stake here is not simply the improvement of technical procedures, but the reconfiguration of auditing as a democratic practice. The proposal to embed citizens infrastructures into audit institutions entails a shift away from compliance-driven, external assessments, towards forms of oversight that are participatory, reflexive and entangled with the lived realities of AI deployment. It opens the possibility for a mode of governance in which standards are co-produced by those whom they actually affect —one in which the evaluation of AI systems evolves with the values, concerns and expectations of the societies they shape and inhabit.

References

AbdouMaliq, S. (2004): "People as Infrastructure: Intersecting Fragments in Johannesburg", in: Public Culture 16 (3), 407–429.

Acemoğlu, D., Restrepo, P. (2019): "Automation and New Tasks: How Technology Displaces and Reinstates Labor", in: Journal of Economic Perspectives 33(2), 3–30.

Allcott, H., Gentzkow, M., Yu, C. (2019): "Trends in the diffusion of misinformation on social media", in: Res. Polit., 6 (2).

Altman, S. (2024): "The Intelligence Age". Available at: https://ia.samaltman.com/. Accessed 28 April 2025.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016): "Machine bias", in: ProPublica, May 23, 2016.

Aridor, G., Goncalves, D., Sikdar, S. (2020): "Deconstructing the filter bubble: user decision-making and recommender systems", in: Proceedings of the 14th ACM Conference on Recommender Systems, pp. 82–91.

Ashery, A.F., Aiello, L.M., Baronchelli, A. (2025): "Emergent social conventions and collective bias in LLM populations", in: Science Advances 11, eadu9368, doi: 10.1126/sciadv.adu9368.

Ashkinaze, J., Mendelsohn, J., Qiwei, L., Budak, C., and Gilbert, E. (2024): "How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence from a Large, Dynamic Experiment", in: arXiv preprint. arXiv:2401.13481.

Aung, Y.Y.M., Wong, D.C.S., Ting D.S.W. (2021): "The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare", in: Br. Med. Bull., 139 (2021), pp. 4–15.

Autor, D., Chin, C., Salomons, A., Seegmiller, B. (2024): "New Frontiers: The Origins and Content of New Work, 1940–2018", in: The quarterly journal of economics: qjae008.

Bail, C., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Fallin Hunzaker, M.B., Lee, J., Mann, M., Merhout, F., Volfovsky A. et al. (2018): "Exposure to opposing views on social media can increase political polarization", in: Proceedings of the National Academy of Sciences 115.37, pp. 9216–9221.

Bandy, J. (2021): "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits", in: Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 74.

Barera, M. (2020): "Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia". Accessible at http://hdl.handle.net/10106/29572.

Barns, S., Cosgrave, E., Acuto, M., Mcneill, D. (2016): "Digital Infrastructures and Urban Governance", in: Urban Policy and Research, 35(1), 20–31.

Bellaiche, L., Shahi, R., Turpin, M.H. et al. (2023): "Humans versus AI: whether and why we prefer human-created compared to AI-created artwork", in: Cogn. Research 8, 42.

Bender, E.M., Friedman, B., (2018): "Data statements for natural language processing: Toward mitigating system bias and enabling better science", in: Transactions of the Association for Computational Linguistics 6, 587–604.

Birhane, A., Steer, R., Ojewale, V., Vecchione, B., Raji, I. D. (2024): "AI Auditing: The broken bus on the road to AI accountability", in: arXiv:2401.14462v1. Accessed 14/05/25. https://doi.org/10.48550/arXiv.2401.14462.

Boeker, M., Urman, A. (2022): "An empirical investigation of personalization factors on TikTok", in: Proceedings of the ACM Web Conference 2022, pp. 2298–2309.

Bond, S. (2004): "AI-generated spam is starting to fill social media. Here's why", in: *NPR Untangling disinformation*. Accessed 27/02/2025. Availavle at: https://www.npr.org/2024/05/14/1251072726/ai-spam-images-facebook-linkedin-threads-meta.

Brinkmann, L., Baumann, F., Bonnefon, JF. et al. (2023): "Machine culture" in: *Nat Hum Behav* 7, 1855–1868.

Burrell, J. (2016): "How the machine 'thinks': Understanding opacity in machine learning algorithms", in: *Big Data & Society*.

Burtell, M., Woodside, T. (2023): "Artificial influence: An analysis of AI-driven persuasion", in: arXiv preprint. arXiv:2303.08721.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F. et al. (2024): "The Impact of Generative Artificial Intelligence on Socioeconomic Inequalities and Policy Making", in: PNAS Nexus 3(6).

Castelvecchi, D. (2016): "Can we open the black box of AI?", in: *Nature* Oct 6 538 (7623): 20–23.

Chamberlain, R., Mullin, C., Scheerlinck, B., Wagemans, J. (2017): "Putting the art in artificial: Aesthetic responses to computer-generated art", in: *Psychol. Aesthet. Creat. Arts.*

Chen, P.-Y., Wu, S.-Y., Yoon, J. (2004): "The impact of online recommendations and consumer feedback on sales", in: International Conference on Information Systems (ICIS).

Cinus, F., Minici, M., Monti, C., Bonchi, F. (2022): "The effect of people recommenders on echo chambers and polarization", in: Proceedings of the In-

ternational AAAI Conference on Web and Social Media, vol. 16, Association for the Advancement of Artificial Intelligence, Washington, DC USA, pp. 90–101.

Colther, C., Doussoulin, J.P. (2024): "Artificial intelligence: Driving force in the evolution of human knowledge", in: Journal of Innovation & Knowledge, 9 (4) 100625, ISSN 2444–569X, https://doi.org/10.1016/j.jik.2024.100625.

Crawford, K. (2021): *The Atlas of AI*. Yale University Press.

Dastin, J. (2018): "Amazon scraps secret AI recruiting tool that showed bias against women", in: *Reuters.* Available at https://www.reuters.com.

Dave, K. (2019): "Systemic algorithmic harms", in: Data & society, May 31st 2019.

Del Vicario, M. Vivaldo, G. Bessi, A. Zollo, F. Scala, A. Cardelli, G. Quattrocchi, W. (2016): "Echo Chambers: Emotional Contagion and Group Polarization on Facebook", in: Scientific Reports, Open Access, Volume 61, December 2016, Article number 37825, https://doi.org/10.1038/srep37825.

Donnelly, R. Kanodia, A. Morozov I. (2021): "The long tail effect of personalized rankings". Available at SSRN 3649342.

Doshi, A.R., Hauser, O.P. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances 10, eadn5290. https://doi.org/10.1126/sciadv.adn5290.

Dotan, M., Yaish, A., Yin, H-C., Tsytkin, E., Zohar, A. (2023): "The Vulnerable Nature of Decentralized Governance", in: DeFi 25–31, https://doi.org/10.1145/3605768.3623539.

Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. (2017): "CAN: Creative adversarial networks, generating 'Art' by learning about styles and deviating from stylenorms", in: arXiv preprint. arXiv:1706.07068.

Ernst, E., Merola, R., Samaan, D. (2019): "Economics of Artificial Intelligence: Implications for the Future of Work", in: IZA Journal of Labor Policy 9(1).

Fabbri, F., Croci, M.L., Bonchi, F., Castillo, C. (2022): "Exposure inequality in people recommender systems: the long-term effects", in: Proceedings of the International AAAI Conferenceon Web and Social Media, vol. 16, Association for the Advancement of Artificial Intelligence, Washington, DC USA, pp. 194–204.

Floridi, L. (2024): "Hypersuasion – On AI's Persuasive Power and How to Deal with It", in: Philos. Technol. 37, 64.

Gabriel, I. (2020): "Artificial Intelligence, Values, and Alignment", in: Minds and Machines 30(3), 411–437.

Gabrys, J. (2021): "Citizen Infrastructures and Public Policy: Activating the Democratic Potential of Infrastructures", in: Cohen, K. and Doubleday, R. (eds): Future Directions for Citizen Science and Public Policy, Cambridge: Centre for Science and Policy.

Gangadharbatla, H. (2022): "The Role of AI Attribution Knowledge in the Evaluation of Artwork", in: Empirical Studies of the Arts, 40(2), 125–142.

Glickman, M., Sharot, T. (2025): "How human–AI feedback loops alter human perceptual, emotional and social judgements", in: Nature Human Behavior 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2.

Hagerty, A. Rubinov, I. (2019) Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint. arXiv:1907.07892.

Hataya, R., Bao, H., Arai, H. (2023): "Will large-scale generative models corrupt future datasets?", in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20555–20565.

Hayashi, S., Caron, B. A., Heinsfeld, A. S. et al. (2024): "brainlife.io: a decentralized and open-source cloud platform to support neuroscience research", in: Nature Methods 21, 809–813, doi: https://doi.org/10.1038/s41592-024-02237-2.

Hong, J.-W., Curran M. N. (2019): "Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence", in: ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15(2):58, 1 – 16.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q. et al. (2024): "TRUSTLLM: trustworthiness in large language models", in: Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 813, 20166–20270.

Innerarity, D. (2024): "Defensa y crítica de la gobernanza algorítmica", in: Revista CIDOB d'Afers Internacionals, 138, pp. 11–25.

Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A. (2023): "Recommendation systems: principles, methods and evaluation", in: Egypt. Inform. J., 16 (3), pp. 261–273.

Jiang, R., Chiappa, S., Lattimore, T., György, A., Kohli, P. (2019): "Degenerate feedback loops in recommender systems", in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19, Association for Computing Machinery, New York, NY, USA, pp. 383–390.

Kimutai, L. (2024): "Is AI killing Pinterest's magic?", in: *Medium*. Accessed 27/02/2025. Available at: https://medium.com/@kimutail/is-ai-killing- pi nterests-magic-2f84c395fd98

Knott, A., Hannah, K., Pedreschi, D., Chakraborti, T., Hattotuwa, S., Trotman, A., Baeza-Yates, R., Roy, R., Eyers, D., Morini, V., Pansanella, V. (2021): "Responsible AI for social media guidance: a proposed collaborative method for studying the effects of social media recommender systems on users", in: Technical report. Global Partnership on AI (GPAI).

Kuziemski, M., Misuraca, G. (2020): "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings", in: Telecommunications Policy, 44(6).

Lahdili, N., Onder, M., Nyadera, I. (2024): "Artificial Intelligence and Citizen Participation in Governance: Opportunities and Threats", in: Amme Idaresi Dergisi. 57. 202–229.

Lamberti, M.J., Wilkinson, M., Donzanti, B.A., Wohlhieter, G.E., Parikh, S., Wilkins, R.G., Getz, K. (2019): "A study on the application and use of artificial intelligence to support drug development", in: Clin. Therapeut., 41 (2019), pp. 1414–1426.

Lawrence, N. (2014): "Open Data Science", in: inverseprobability.com: Neil La wrence's Homepage, accessed on: 15/05/25, available at: https://inversepro bability.com/2014/07/01/open-data-science.

LeCun, Y., Bengio, Y., Hinton, G. (2015): "Deep learning", in: Nature, 521 (7553), pp. 436–444.

Lee, D., Hosanagar, K. (2014): "Impact of recommender systems on sales volume and diversity", in: International Conference on Interaction Sciences.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M.F., Li, H. (2024): "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment", in: arxiv.org/abs/2308.05374.

Livingstone, S., Lievrouw, L. A. (2010): "How to infrastructure", in: Star, S., Bowker, G. (Eds.): *Handbook of New Media: Social Shaping and Social Consequences of ICT*, pp. 230–245. SAGE Publications Ltd.

Martínez, G. Watson, L. Reviriego, P. Hernández, J. A., Juarez, M. Sarkar, R. (2023): "Towards understanding the interplay of generative artificial intelligence and the Internet", in: arXiv:2306.06130.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., Burke, R. (2020): "Feedback loop and bias amplification in recommender systems", in: Proceedings of the 29th ACM International Conference on Information

& Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, pp. 2145–2148.

Mazzone M., Elgammal, A. (2019): "Art, Creativity, and the Potential of Artificial Intelligence", in: *Arts* 8(1):26.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., Griffiths, T. L. (2023): "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve", in: arxiv.org/abs/2309.13638v1.

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., and Halgamuge, M. N. (2024): "A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination", in: IEEE Transactions on Artificial Intelligence 5(6): 2739–2751.

McQuillan D. (2018): "Rethinking AI through the politics of 1968", in: Open Knowledge, October 13th 2018.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021): "A survey on bias and fairness in machine learning", in: ACM Comput. Surv., 54 (6) (2021), pp. 1–35.

Muldoon, J., Graham, M., Cant, C. (2024): Feeding the Machine: The Hidden Human Labour Powering AI. Canongate Books.

Nguyen, M.H., Tran, N.D., Le, N.Q.K. (2023): "Big data and artificial intelligence in drug discovery for gastric cancer: current applications and future perspectives", in: Curr. Med. Chem., 31.

OpenAI (2022): DALL·E 2. Accessed 27/02/2025. https://openai.com/index/dall-e-2/.

Ovadya, A., Thorburn, L. (2023): "Bridging Systems. Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance", in: Knight First Amend. Inst., 23–11.

Pansanella, V., Rossetti, G., Milli, L. (2022a): "Modeling algorithmic bias: simplicial complexes and evolving network topologies", in: Appl. Netw. Sci., 7 (1), p. 57.

Pansanella, V., Rossetti, G., Milli, L. (2022b): "From mean-field to complex topologies: network effects on the algorithmic bias model", in: R.M. Benito, C. Cherifi, H. Cherifi, E. Moro, L.M. Rocha, M. Sales-Pardo (Eds.), Complex Networks & Their Applications X, Springer, Cham, pp. 329–340.

Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., Gezici, G., Giannotti, F., Lalli, M., Gambetta, D., Mauro, G., Morini, V., Pansanella, V., Pedreschi, D. (2024): "A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions", in: Arxiv 2407.01630. https://arxiv.org/abs/2407.01630.

Pariser, E. (2011): The Filter Bubble: What the Internet is Hiding from You, New York: Penguin Press. 294 pp.

Pathak, B., Garfinkel, R., Gopal, R.D., Venkatesan, R. Yin, F. (2010): "Empirical analysis of the impact of recommender systems on sales", in: J. Manag. Inf. Syst., 27 (2) (2010), pp. 159–188.

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A-L., Dignum, F., Dignum, V ., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y ., Lukowicz, P., Passarella, A., Pentland, A.S., Shawe-Taylor, J., Vespignani, A. (2025): "Human-AI coevolution", in: Artificial Intelligence, V olume 339, 104244, ISSN 0004–3702, https://doi.org/10.1016/j.artint.2024.104244.

Peralta, A.F., Kertész, J., Iñiguez, G. (2021): "Opinion formation on social networks with algorithmic bias: dynamics and bias imbalance", in: J. Phys. Complex., 2 (4).

Piao, J., Liu, J., Zhang, F. et al. (2023): "Human–AI adaptive dynamics drives the emergence of information cocoon", in: Nature Machine Intelligence 5, 1214–1224. https://doi.org/10.1038/s42256-023-00731-4.

Pew (2018): Internet/Broadband Fact Sheet. https://www.pewinternet.org/fact-sheet/internet-broadband/.

Pi, Y. (2021): "Machine learning in Governments: Benefits, Challenges and Future Directions", in: JeDEM – EJournal of EDemocracy and Open Government, 13(1), 203–219.

Porcaro, L., Gómez, E., and Castillo, C. (2024): "Assessing the Impact of Music Recommendation Diversity on Listeners: A Longitudinal Study", in: ACM Trans. Recomm. Syst. 2, 1, Article 3 (March 2024), 47 pages. https://doi.org/10.1145/3608487.

Rahman, R., Owen, D., You, J. (2024): "Tracking Large-Scale AI Models", in: EpochAI. Availavle at: https://epoch.ai/blog/tracking-large-scale-ai-models.

Russell, S. (2020): Human Compatible: Artificial Intelligence and the Problem of Control, Penguin Publishing Group.

Sartori, L., Theodorou, A. (2022): "A sociotechnical perspective for the future of AI: narratives, inequalities, and human control", in: Ethics Inf. Technol., 24 (1), p. 4.

Schepman, A., Rodway, P. (2020): "Initial validation of the general attitudes towards Artificial Intelligence Scale", in: Computers in Human Behavior Reports 2020(1).

Shin, M., Kim, J., van Opheusden, B., Griffiths, T. L. (2023): "Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty", in: Proceedings of the National Academy of Sciences 120(12): e2214840120.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y. (2024): "AI models collapse when trained on recursively generated data", in: Nature, 631 (8022), pp. 755–759.

Sieber, R., Brandusescu, A., Sangiambut, S., Adu-Daako, A. (2024): "What is civic participation in artificial intelligence?", in: Environment and Planning B, 0(0).

Sîrbu, A., Pedreschi, D., Giannotti, F. and Kertész, J. (2019): "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model", in: PloS one, 14(3), p.e0213246.

Srba, I. Moro, R. Tomlein, M. Pecher, B. Simko, J. Stefancova, E. Kompan, M. Hrckova, A. Podrouzek, J. Gavornik, A. et al. (2023): "Auditing YouTube's recommendation algorithm for misinformation filter bubbles", in: ACM Trans. Recommend. Syst., 1 (1), pp. 1–33.

Stilgoe, J. (2024): "AI has a democracy problem. Citizens' assemblies can help", in: *Science* 385, eadr6713.

Stray, J., Iyer, R., Puig Larrauri, H. (2023): "The Algorithmic Management of Polarization and Violence on Social Media", in: Knight First Amendment Institute. KnightColumbia.Org, forthcoming, Available at: http://dx.doi.org/10.2139/ssrn.4429558.

Sun, W., Khenissi, S., Nasraoui, O., Shafto, P. (2019): "Debiasing the human-recommender system feedback loop in collaborative filtering", in: Companion Proceedings of the 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, pp. 645–651.

Tlili, A., Shehata, B., Adarkwah, M.A., Bozkurt, A., Hickey, D. T, Huang, R., Agyemang, B. (2023): "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education", in: Smart Learning Environments, 10.

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Bielikova, M. (2021): "An audit of misinformation filter bubbles on YouTube: bubble bursting and recent behaviours changes", in: Proceedings of the 15th ACM Conference on Recommender Systems, pp. 1–11.

Turki, A. T., Engelke, M., Sobas, M. (2024): "Advances in Decision Support for Diagnosis and Early Management of Acute Leukaemia." The Lancet Digital Health 6(5): e300–e301.

UNDP (United Nations Development Program) (1997): Governance for Sustainable Human Development, New York: UNDP Policy Document.

UNDP (United Nations Development Programme) (2025): Human Development Report 2025: A matter of choice: People and possibilities in the age of AI, New York: UNDP Report.

Vecchione, B., Levy, K., Barocas, S. (2021): "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies", in: Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, NY, USA. ACM, New York, NY, USA.

Wagner, C., Strohmaier, M., Olteanu, A. et al. (2021): "Measuring algorithmically infused societies", in: Nature 595, 197–204. https://doi.org/10.1038/s41586 -021-03666-1.

Wang, W., Liu, Q.-H., Liang, J., Hu, Y., Zhou, T. (2019): "Coevolution spreading in complex networks", in: Physics Reports, Volume 820, Pages 1–51, ISSN 0370–1573, https://doi.org/10.1016/j.physrep.2019.07.001.

Westermann, H., Savelka, J. (2024): "Analyzing Images of Legal Documents: Toward Multi-Modal LLMs for Access to Justice", in: *Arxiv* 2412.15260. Accessed on 15/05/25. https://doi.org/10.48550/arXiv.2412.15260.

Wieringa, M. (2020): "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability", in: FAT* 2020 – Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, Inc, pp. 1–18.

Wilson, C. (2022): "Public engagement and AI: A values analysis of national strategies", in: Government Information Quarterly, 39(1).

Winner, L. (2017): Do Artifacts Have Politics? Computer Ethics. London: Routledge.

World Bank (2018): "Indiviuals Using the Internet". https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2017amp;locations=USamp;start=2015.

Wu, Y., Yi Mou, Zhipeng Li, Kun Xu (2019): "Investigating American and Chinese Subjects' explicit and implicit perceptions of AI-Generated artistic work" in: Computers in Human Behavior 2019(104).

Wu, J., You, H., Du, J. (2025): "AI generations: from AI 1.0 to AI 4.0", in: arXiv preprint. arXiv:2502.11312.

Yang, C., Xu, X., Nunes, P., Siqueira, S.W.M. (2023): "Bubbles bursting: investigating and measuring the personalisation of social media searches", in: Telemat. Inform., 82, Article 101999.

Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F., (2019): "Systematic review of research on artificial intelligence applications in higher education – where are the educators?", in: International Journal of Educational Technology in Higher Education, 16.

Zhao, J., Wang, T., Yatskar, M., Cotterell, Ordonez, V., and Chang, K-W. (2019): "Gender Bias in Contextualized Word Embeddings", in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 629–634. https://doi.org/10.18653/v1/N19-1064.

Zhu L, Rong Y, McGee L. A., Rwigema J. M., Patel S. H. (2024): "Testing and Validation of a Custom Retrained Large Language Model for the Supportive Care of HN Patients with External Knowledge Base", in: Cancers (Basel). 16(13):2311. https://doi.org/10.3390/cancers16132311. PMID: 39001375; PMCID: PMC11240646.

Zysman, J., Nitzberg, M. (2024): "Generative AI and the Future of Work: Augmentation or Automation?". Available at SSRN 4811728.

Algorithmic Audits of HR-AI Tools in the United States

What Are HR Professionals' and Job Seekers' Perspectives?

Tina B. Lassiter, Kenneth R. Fleischmann

Abstract: *Algorithmic audits are gradually becoming more common, especially in the human resources (HR) sector. The use of Artificial Intelligence (AI) in HR and particularly regarding employment decisions is considered a high risk and can have long-term consequences for the affected stakeholders. In response to the risks involved, recent regulation in New York City and in the European Union requires audits of AI-based tools used in employment decisions. The market is responding by developing and offering a wide range of types of audits. While there is a growing amount of research centered on algorithmic audits in general and how they could be conducted and regulated, less research has focused on the perspectives on algorithmic audits of HR professionals and job seekers. This study aims to fill the gap in the literature by presenting research based on a survey conducted with 436 users of human resource AI tools (HR-AI tools), including 329 job seekers and 107 human resource managers and recruiters in the United States (U.S.) to learn about their perspectives. The survey shows substantial differences between the two stakeholder groups regarding knowledge of AI usage, trust in AI and HR-AI tools, and their knowledge and opinions with respect to algorithmic auditing. It also demonstrates the current general lack of information about algorithmic auditing and calls for more inclusion of all stakeholder groups in the future development of algorithmic auditing in the HR sector.*

Keywords: *AI; Human Resources; artificial intelligence; HR-AI tools*

Introduction/Background

Use of Artificial Intelligence in the Human Resource Sector

For our study, we understand human resource AI tools (HR-AI tools) as AI tools that are used to manage any part of an employee's life cycle (including recruiting, hiring, onboarding, training, evaluation, or promotional decisions).

Most of the current research in the field has focused on hiring decisions. The use of Artificial Intelligence (AI) tools for filtering and identifying suitable applicants has become increasingly common. Hilke Schellmann reports in her recent book "The Algorithm" that already 99% of Fortune 500 companies use algorithms and artificial intelligence for hiring (Schellmann 2024). AI tools influence who sees job offers, can be used to evaluate resumes, sometimes conduct job interviews, and may interact with job seekers in various other ways (Jarrahri et al. 2021; Li et al. 2021). Resume screeners, such as the one offered by Eightfold.AI[1], find candidates and provide ranked lists. They are often integrated into Applicant Tracking Systems (ATS) (Chavan et al. 2024). Some companies, like Hirevue[2], offer one-way interviews analyzed by AI. As a recent review on the use of AI in HR-management demonstrates, there is a wide range of different AI applications in the field (Bujold et al. 2024). While the use of AI in HR-management might lead to efficiencies and cost savings (Li et al. 2021), it can also lead to harmful results. Examples of potential harms are plenty and can be found in repositories of AI gone wrong.[3] Job seekers can be systematically excluded by algorithms because of age, sex, or other demographics (Chen 2023). Resume screeners might filter out candidates through variables that directly discriminate or by proxy variables (Raghavan 2020), and job interviews recorded and analyzed by AI have shown discrimination regarding race, religion, or other attributes, e.g., demonstrated in an experiment by a German TV station that tested AI used in automated interviews (Harlan & Schnuck 2021). In a famous case from 2023, a hiring algorithm from iTutor automatically excluded female applicants over 55 and male applicants over 60 and iTutor paid $ 365.000 to settle a hiring discrimination suit initiated by the Equal

1 https://eightfold.ai/use-case/candidate-screening/.

2 https://www.hirevue.com/.

3 https://incidentdatabase.ai/; https://www.aiaaic.org/home; https://airisk.mit.edu/ai-incident-tracker.

Employment Opportunity Commission (EEOC)[4]. The question of fairness regarding the use of AI in HR-decisions is raised by multiple researchers (Harlan & Schnuck 2021; Hunkenschroer et al. 2024; Rigotti & Fosch-Villaronga 2024). Though less researched, the use of AI-tools in other areas of the employment cycle can have similarly negative impacts. As a result of these potential negative outcomes, the use of AI for employment has been on the forefront of AI regulation and AI guidance.

Algorithmic Audits of HR-AI Tools

One way to ensure HR-AI tools provide what they promise to achieve and are trustworthy could be to have them audited. AI audits have been suggested increasingly in the literature as a useful way to create informed trust in AI (Mökander 2023, Mökander & Floridi 2021, Selinger 2023). Algorithmic audits are relatively new, even though Sandvig (2014) proposed the idea over a decade ago. Currently, many consider the situation in the algorithmic auditing ecosystem as the "Wild West" (Costanza Shock 2022; Lassiter & Fleischmann 2024). There is no one definition of an algorithmic audit, there are various ways to conduct algorithmic audits, and currently there is little regulation or standardization (Bandy 2021, Costanza-Shock et al. 2022, Goodman & Trehu 2023, Lam 2024, Raji et al. 2020, Selinger 2023). Recently, there were movements to change this, such as the formation of the International Association of Algorithmic auditors,[5] and there have been calls for common standards regarding algorithmic audits, for example from the National Telecommunications and Information Administration (NTIA 2024).

In our study, we aim to explore algorithmic audits in the widest sense. For this work, we follow the approach of the UK government, using it as an umbrella term describing any assessment of an algorithmic system which can be "*technical and non-technical measures that range from assessing organizational algorithmic governance policies to the specific data and models being used*" (DRCF 2022). This allows us to include a wide range of understandings regarding algorithmic audits. Auditing algorithms used in the human resource (HR) sector are particularly important as the decisions are highly consequential (Ajunwa 2020). Interest in algorithmic audits in HR is increasing (Dawson 2022, Lassiter & Fleis-

4 https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit.

5 https://iaaa-algorithmicauditors.org/.

chmann 2024). Past experiences of voluntarily published audit reports regarding HR-tools have had mixed successes (Geiger et al. 2023), in particular the controversial audits from Pymetrics (Wilson 2021) and HireVue (Engler 2021). Recently, regulation has been introduced to require algorithmic audits in HR. The European AI Act (European Commission 2021) includes the recommendation to audit high-risk algorithms which include HR-AI tools. It also requires increased transparency requirements for high-risk AI systems such as employment decisions. If natural persons are directly interacting with an AI system, they need to be notified that they are interacting with an AI system (Art. 50 EU AI Act). The Digital Service Act (DSA) also requires assessments/audits (European Parliament 2022). In the US, the New York City (NYC) has taken further action specifically in the HR sector and has introduced the first law worldwide (Groves et al. 2024, Wright et al. 2024) to mandate AI bias audits for AI tools used for employment decisions (New York City Council 2022). The New York City Law 144 took effect in July 2023. It requires NYC-based employers who use automated employment decision-making tools (AEDTs) to hire a third-party auditor to conduct an annual bias audit and to make the audit report public (New York City Council 2022, Wright et al. 2024). In Section § 20–871 2 b of Law 144 it also requires a notice of the usage of an automated employment decision tool. The law went through several iterations before it was finalized. It is criticized because of its lack of defining key aspects leading to a failure to protect job applicants as intended (Groves et al. 2024, Fuchs 2023). Early research focusing on the implementation of Law 144 also found that only a small fraction of employers has conducted the required audits, most likely because of possible loopholes in the law, and additionally it was observed that records of the bias audits were of limited value for job seekers because of lack of accessibility and usability (Wright et al. 2024).

Research Focus

Our research is focused on algorithmic audits of HR-AI tools in the United States (U.S.) and how they could create calibrated trust of stakeholders who use the tools and/or are affected by them. Employment AI tools have been one of the first areas that have been explored by algorithmic audits, mainly due to the regulation by the NYC law. In anticipation of this regulation, algorithmic auditing companies prepared for the need for algorithmic audits in this sector and were most active in this area.

While there is a growing body of literature studying algorithmic audits in general and how they should be designed (e.g., Bandy 2021, Costanza-Shock et al. 2022, Goodman & Trehu 2022, Lassiter & Fleischmann 2024, Raji 2020), there is less research on the various stakeholders' perspectives on algorithmic audits. This lies in the nature of the matter. Algorithmic audits are not yet common, which explains the lack of experience with such audits. AI auditing professionals have also expressed that they have very limited knowledge about how their audits are received by end-users as they mostly focus on their relationship with the client who hires them to conduct an audit (Lassiter & Fleischmann 2024).

With this study, we aim to fill part of this gap by looking closer at the perspectives of users regarding algorithmic audits of HR-AI tools. We surveyed users of HR-AI tools and stakeholders affected by them to find out how much they know about algorithmic audits and how they feel about them. We chose to focus on the two stakeholder groups most impacted by HR decisions: human resource managers and recruiters (HR professionals) who use the tool to find employees and for other employment decisions, and job seekers, who are impacted by the tools by being the subject of them. It is important to look at these two stakeholder groups separately, as there is a notable power asymmetry between the two groups (Hunkenschroer 2023). While HR-professionals mostly have knowledge of the use of HR-AI tools, most job seekers will not know whether an HR-AI tool is used or not. Furthermore, several HR-professionals might have a choice to use AI for HR-decisions. Job seekers on the other hand, will most likely not be able to have a say in the use of HR-AI tools regarding their application. At the most, they might have an option to opt out, after being informed about the use of AI, for example in an automated job interview where the use of AI is obvious. In most circumstances, for example with resume screeners, they will not even know that there will be an AI decision. Furthermore, even if they knew that AI is used and would rather not have it used, they will not always opt out if they need to find a job.

Our research is guided by the following research questions (RQ):

1. How much do HR professionals and job seekers trust AI and HR-AI tools?
2. What do HR professionals and job seekers know about algorithmic audits?
3. What are HR professionals' and job seekers' perspectives on algorithmic audits of HR-AI tools?

Methods

Data Collection

The research results are part of a mixed method study in the U.S., exploring the opinions of users of human resource AI tools regarding algorithmic audits of such tools and how the latter can create trust in these tools. In previous research, we examined through an interview study with 19 AI auditing professionals how they aim to create calibrated trust in AI by users (Lassiter & Fleischmann 2024). This research was followed by a study surveying users of HR-AI tools on how algorithmic audits potentially influence their trust in AI in the HR sector.

Sampling Strategy

We conducted an online survey with 436 job seekers and human resource managers and recruiters (HR professionals) in the U.S. The survey was run in the fall of 2023 with purchased participant panels from Qualtrics[6] (requesting HR professionals with experience using HR-AI tools, and job seekers, defined as participants who are currently seeking a job or have recently been on the job market). Recruitment was undertaken by Qualtrics which sent out invitations to all potential participants. The survey resulted in 329 valid responses from job seekers and 107 valid responses from HR professionals.

Demographics

The participants on the job seekers side were overwhelmingly female (77% vs. 22% male and 2% non-binary/other). Thus, the sample includes a larger proportion of female participants than the population being sampled. According to 2023 data from the Bureau of Labor Statistics, around 55% of job seekers are male, and around 45% are female (BLS, 2023). An explanation for this discrepancy is that there is a general trend observed in online surveys where females are more responsive than males (Rourke & Lakner 1989, Smith 2008). As there were no significant differences between males and females in the responses seen when we separated the results according to gender, the gender distribution of participants does not present a challenge.

6 https://www.qualtrics.com/.

The detailed demographics of the participants are shown in table 2. About two thirds of the participants were between 25–54 years old, about 5% between 18 and 24, and about 29% were 55 and older. The participants were predominantly white/Caucasian (77%). Black/African American participants made up 11% of the sample, Hispanic 6%, Asian 3%, and Indian American/Native American or Alaska natives another 3% and others 1%. The educational background of the participating job seekers varied, with the majority having a 4-year or 2-year degree. All HR professionals have had experience with HR-AI tools, and we only included participants in this category who had at least a 4-year degree. Most job seekers were currently looking for a job (67%). The others have recently looked for a job (33%). We did not include anyone in the job seeker category who was not currently seeking a job or has not recently been on the job market.

Table 1: Demographics of participants

Demographics		**Job Seekers (329 total)**	**HR professionals (109 total)**	**Total**
Age	18–24	24	0	24
	25–34	55	27	82
	35–44	69	49	118
	45–54	65	19	84
	55–64	64	12	76
	Over 65	51	0	51
	Prefer not to say	1	0	1
Gender	Male	70	44	114
	Female	253	63	316
	Non-binary/third gender	5	0	5
	Prefer to self-identify	1	0	1
	Prefer not to say	0	0	0

Demographics		Job Seekers (329 total)	HR professionals (109 total)	Total
Race	African American or Black	37	19	56
	American Indian/ Native American, or Alasca Native	9	4	13
	Asian	12	4	16
	Hispanic/Latino	19	11	30
	Native Hawaiian, or Other Pacific Islander	1	2	3
	White or Caucasian	264	79	343
	Other	3	2	5
	Prefer not to say	1	0	1
Education	Some high school or less	6	0	6
	High school graduate or GED	54	0	54
	Some College	77	73	150
	2 year degree	52	34	86
	4 year degree	94	0	94
	Professional degree or graduate degree	42	0	42
	Other	3	0	3
	Prefer not to say	1	0	1

Survey Design

The survey contains 37 questions. The most relevant ones for this research are shown in table 1. The complete list can be found in the appendix. Most of the questions use a Likert scale of 1–5. Of these, 4 are attention check questions for validation. Following questions determining whether the participants are qualified for the study and demographic questions (Q1-Q8), we asked the participants about their experience and opinions regarding AI and HR-AI-tools. There are many different definitions of AI with no one settled standard. We provided a brief definition for the participants to have a common understanding of the concept for the purpose of this survey. In the survey, we explained to the participants that we understand Artificial Intelligence (AI) 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*", using the definition of the National Artificial Intelligence Initiative Act from 2020 (H.R. 6216). We explained HR-AI tools as "*AI tools managing any part of an employee's life cycle (including recruiting and hiring)*". We then asked the participants to rate their trust in AI in general (Q 9) and later their trust in HR-AI tools in particular (Q 15). Trust in HR-AI tools can be influenced by knowledge about use of AI and previous experience of AI. Therefore, the survey wanted to explore first the participants' knowledge of AI and then their previous experience with AI. All HR professionals had previous experience using HR-AI tools. This was a requirement for participation, as we wanted to focus on their experience as users knowingly using AI for their work. The same requirement was not made for job seekers, as we assume that many are not aware of the use of AI in HR. Therefore, Q11 specifically asks job seekers whether they were ever told about the use of AI regarding an HR-AI tool and how important it is for them to know AI is used (Q12). Building trust in HR-AI tools is likely based on the lived experience of the users with the tools. We therefore also wanted to explore the stakeholders' overall experience using the tools or being affected by them (Q 13, 14). To explore the perspectives of both stakeholder groups on algorithmic audits, we first wanted to know how much they know about these audits and what their experience with audits were (Q20, 21). We only asked HR professionals about their experience with algorithmic audits, as job seekers so far have very limited experience with algorithmic auditing.

The survey went through multiple iterations. In a pilot run, we tested a survey version without any explanations of AI auditing. In this first sample, most participants seemed to have not fully understood the concept of AI au-

dits. Their answers showed several inconsistencies and were not useful. In addition, participants also seemed to just have randomly answered questions to finish the survey quick. We adapted the survey accordingly and excluded data of low quality. Following a second successful pilot run, the survey questions were finalized. The most significant changes were adding an explanation of AI audits (Q19), and the addition of several validation questions (Q10, 16, 30 – see annex). After initially asking about the knowledge about algorithmic audits, we added a brief explanation of the basic concept and then continued with more detailed questions. We gave an operational definition of the term AI audit in Q19 by explaining: "*Since AI does not always work as expected or intended, in some cases, organizations audit AI, which checks to ensure that they comply with legal and ethical standards*." This description aims to not lead the answers of the participants too much but still gives them a very basic understanding of the concept. The question intends to elicit spontaneous reactions once the participants understand the basic concept. Validation questions have been shown to help improve the quality of surveys (Abbey & Meloy 2017, Kung 2018). The validation questions led to a significantly improved set of data.

Table 2: Survey Questions used in this study

Q9	In this survey we understand Artificial Intelligence (AI) as … How much do you trust Artificial Intelligence in general?	Slider question (scale of 1–5)
Q11	In our study we are examining AI tools used in Human Resources. We understand Human Resource AI Tools (HR-AI tools) as AI tools managing any part of an employee's life cycle (including recruiting and hiring). Were you told that AI was used in the application and hiring process when you applied for jobs?	Only for job seekers Slider question (scale of 1–5)
Q12	How important is it for you to know if AI is used in the application or hiring process?	Only job seekers Slider question (scale of 1–5
Q13	How was your overall experience using HR-AI tools as a Human Resource Manager or recruiter?	Only HR Slider question (scale of 1–5)

Q14	How was your overall experience with HR-AI tools as a jobseeker	Only job seekers Slider question (scale of 1–5)
Q15	How much do you trust Human resource AI tools?	
Q17	Have you ever heard of Algorithmic auditing?	Yes – No
Q18	Have you heard of the New York City Law 144?	Multiple Choice
Q19	Since AI does not always work as expected or intended, in some cases, organizations audit AI, which checks to ensure that they comply with legal and ethical standards. How important is it to audit AI, and when and how should this be done?	Open text
Q20	Have you ever worked with an HR-AI tool that has been audited?	Only HR Yes – Unsure – No
Q21	What kind of audit was performed?	Open text
Q22	How important are each of these aspects of audits of AI for you? (Safety and reliability of the AI tool, Legal compliance, Discovering of bias, Potential harm to the public)	Slider question for each (scale of 1–5)
Q23	Are there other important aspects of audits of AI for you that were not mentioned above? (Please explain)	Open text
Q24	What level of trust would you have in HR-AI tolls in these different situations (HR-AI tool is not audited, HR-AI tool is self-assessed by the company, HR-AI tool is audited by an external auditor)	Slider question for each (scale of 1–5
Q35	Please share any additional comments you have regarding audits of AI of Human Resource AI tools	Open text

Data Analysis

The data was analyzed by the first author, using Qualtrics, Excel, and ATLAS.ti[7].

Statistical calculations

An initial data analysis was performed with Qualtrics. Some values were transformed into percentages, and Excel was applied for the final charts and statistic

7 https://Atlasti.com.

calculations. For most statistical calculations, 2-tailed t-tests were performed to determine the difference between the means of the two analyzed groups (job seekers and HR-professionals). In section 3.2.1 (figures 9, 10), a Chi square test was chosen as most appropriate as the data is categorical (Yes, No). There has been a considerable debate among scholars whether Likert scales should only be analyzed with non-parametric statistics, or whether they could also be analyzed with parametric statistics, such as the t-test (DeWinter & Dodou 2010, Huh & Gim 2025). Non-parametric tests seem the appropriate choice for ordinal values, such as the Likert scale. Parametric statistics have, however, often been used in social science research when the intervals have been deemed equal and the sample size is large enough. Parametric statistics is seen as a powerful and comprehensive method (Norman, 2010). Research has also shown that both kinds of test mostly lead to similar results (DeWinter & Dodou 2010). Many scholars therefore consider the use of parametric tests for Likert scales as a valid choice (Norman 2010, Vieria, 2016). Following this approach, we apply parametric statistics like the t-test for our data, as our sample size is large enough and the intervals in the Likert scale can be considered equal.

Qualitative analysis

ATLAS.ti was used for the qualitative analysis of the open-ended text questions. The method was inductive, using a reflective thematic analysis according to Braun and Clarke (2021). The answers of both job seekers and HR professionals were coded by the first author. In a next step, codes were merged and sorted. The data analysis by atlas.ai was then transformed into charts through Excel. The second author supervised the data analysis and provided constructive feedback on it throughout the process.

Findings

RQ 1: Trust in AI and Trust in HR-AI Tools

Trust in AI

The overall trust in AI for all participants (figure 1) appears to be distributed like a bell curve, indicating an almost normal distribution with a slight tendency towards little or no trust. However, isolating the results for HR and for

job seekers (figure 2) shows a stark difference regarding trust in AI tools for the two groups.

Figure 1: Trust in AI of all participants

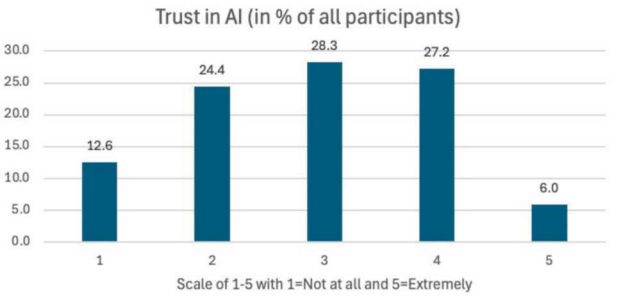
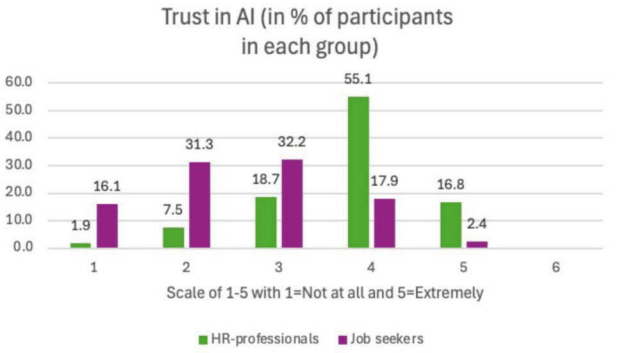


Figure 2: Trust in AI of different stakeholders



Job seekers seem to have much more distrust in AI than HR managers and recruiters. Job seekers echoed a bad feeling towards AI in some of their answers to text questions in the survey.

Some stated, "*I don't agree with using AI at all*", "*AI is a tool that is going to ruin our world*", or "*AI is too new to be trusted*". Another job seeker described in more detail:

> *I haven't been that impressed with AI. I've only seen it when I've asked questions on Google and the answers it generates are very generic. I don't find it particularly helpful. I imagine the same would carry over into human resource AI tools. For example, I applied to a job and was rejected in the same time it took me to create the application. I have to assume it was screened by AI. Some machine just decided based on arbitrary information that I wasn't qualified without considering my overall application. I tend to stay away from these systems if I can.*

HR professionals did not add any additional general comments about AI in the open-ended questions.

Overall, HR professionals have much more trust in AI in general than job seekers. While more than 70% of HR trust AI in general, only about 21% of job seekers do, while 47% have little or no trust in AI. This is a remarkable difference in the attitude towards AI. The average value for HR professionals regarding their overall experience using HR-AI tools on a scale of 1–5 was 3.77 and only 2.58 for job seekers. A t-test for figure 2 was performed, showing a 2-tailed significance of $p<0.001$.

Trust in HR-AI Tools

Knowledge of Job Seekers About Use of AI in Human Resources

The following graphs (figure 3 and figure 4) show how much knowledge job seekers have about the use of AI in the application and hiring process and whether they would like to know when AI is used in the process.

Figure 3: Knowledge of the use of HR-AI tool (job seekers)

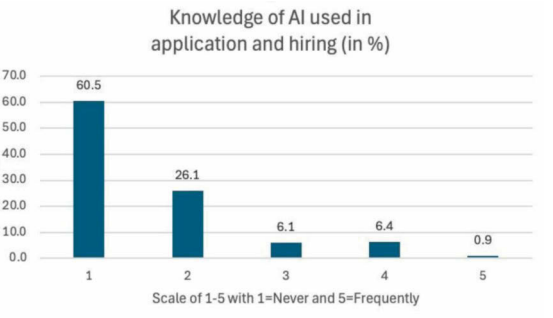
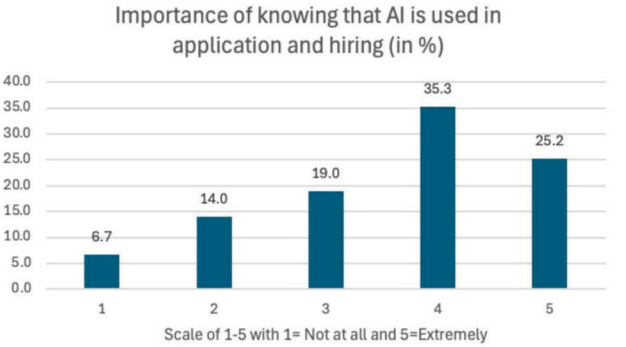


Figure 4: Desire to know when AI is used (job seekers)



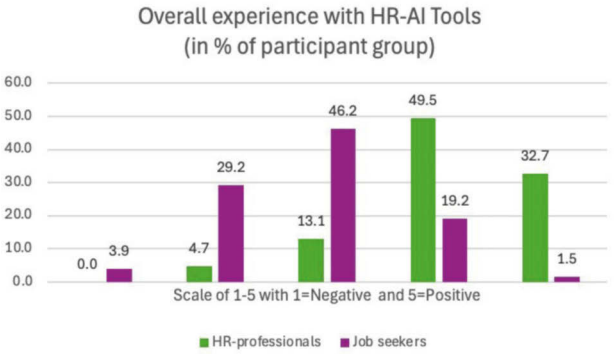
The responses of the job seekers in figures 3 and 4 show that most job seekers do not know or have very little knowledge about when and to what extent AI is used in the application and hiring process but would very much like to know when it is used. Around 60% found it important or extremely important. This shows a large gap between the current situation and the desired information job seekers would like to have.

Overall Experience Using HR-AI Tools

To understand the following graph (figure 5), it must be noted that, as explained above, it was a requirement for HR professionals to participate in the survey to have used AI tools for HR, so the responses reflect the answers from all participating HR professionals. Most job seekers had no knowledge of whether AI tools were used or not, as shown above. The results shown in figure 5 reflect the proportion of job seekers who have used AI tools or having knowingly been subjected to them, which was about 40% of the participating job seekers (130 of 329 job seekers).

The average value for HR professionals regarding their overall experience using HR-AI tools on a scale of 1–5 was 4.1 and only 2.85 for job seekers. A t-test was performed for figure 5, showing a 2-tailed significance of $p<0.001$.

Figure 5: Experience of stakeholders with HR-AI tools (HR professionals: n=107, Job seekers with experience: n=199)



The experience of HR professionals was thus overall, much better than the experience of job seekers who had knowledge about the use of AI tools in the hiring process, with around 82% having a good or very good experience as opposed to only around 21% of the job seekers with similar good or very good experiences. Job seekers who have used HR-AI tools are mostly neutral (46%) with a tendency to a negative experience, though not an extremely bad one.

Trust of HR Managers and Recruiters and Job Seekers in HR-AI Tools

The trust in HR-AI tools of all participants (figure 6) is overall lower than in AI in general (figure 1), with about half of the participants not trusting the tools or only trusting them a bit. There is again a big difference between the trust HR professionals place in HR-AI tools and the trust job seekers place in these tools. Looking at both, the experience of users with HR-AI tools and their responses on how much they trust these tools (figure 7), shows that there is a high mistrust among job seekers but not so among HR professionals. While HR professionals generally trust HR-AI tools with an average of 3.83 points on the scale of 1–5, the job seekers generally distrust the tools with an average of 2.27. A t-test was performed for figure 7, showing a 2-tailed significance of $p<0.001$.

Figure 6: Trust in HR-AI tools (all participants)

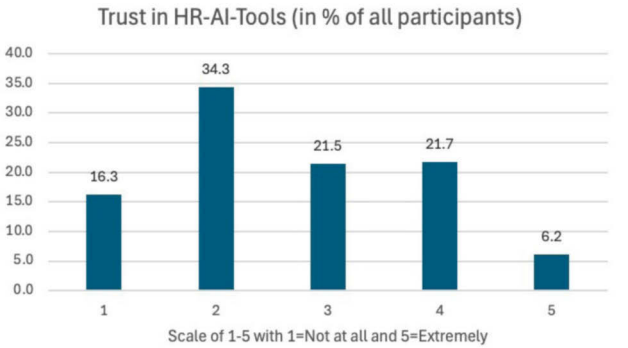
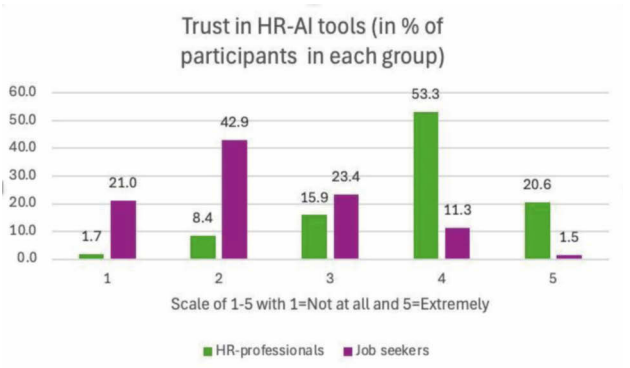


Figure 7: Trust in HR-AI tools of different stakeholders

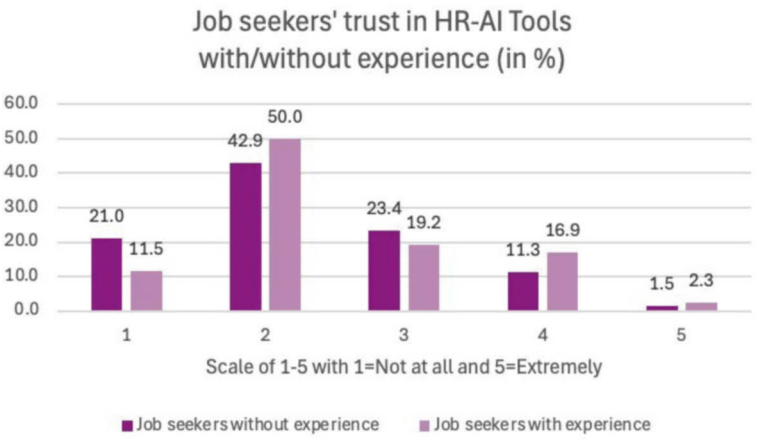


Comparing the trust of both stakeholder groups in AI in general versus the trust of AI in HR, there is an even more surprising difference notable. HR professionals trust HR-AI tools slightly more than AI in general, while job seekers trust it considerably less. Thus, the stakeholders using the tool consciously generally trust it, while the stakeholders who are not in control of using it but who are impacted by it tend to distrust it.

To find out whether the level of experience influences the level of trust, we also looked at the difference between job seekers who have been aware of the use of AI versus job seekers without such knowledge (figure 8). There is some

difference noticeable. Job seekers with experience tend to trust HR-AI tools more than the ones without. It is noteworthy though, that the percentage of job seekers not trusting the tools at all goes down significantly with increased experience. The average trust level of job seekers without experience is 2.29 and the average trust level of job seekers with experience is 2.48. The t-test with regard to figure 8 showed a 2-tailed significance of p=0.003.

Figure 8: Trust in HR-AI tools of job seekers



RQ 2: Knowledge about Algorithmic Audits, Experience with Audits

Knowledge about Algorithmic Audits and the New York City Law

Our survey shows that the participants of our survey overall know little about algorithmic auditing (figure 9) (69% have never heard of it) and even less about the New York City law bias audits (figure 10) (84% have never heard of it). There is, however, a significant difference between job seekers and HR professionals. While most job seekers have never heard of algorithmic audits (84%), most HR professionals who use AI tools are aware of algorithmic audits (87%). The difference regarding the NYC law 144 is even more distinct. Of the job seekers 98% have never heard of it, while 63% of the HR-mangers/recruiters have some or good knowledge of it. Chi-square tests of independence showed that there was a significant relation between stakeholder role and knowledge of algorithmic audits and of the NYC law. The p-value is <0.001 for both tests. For the NYC

Law results, the two columns indicating knowledge were aggregated, so the comparison was between Yes and No responses.

Figure 9: Have you ever heard of Algorithmic auditing?

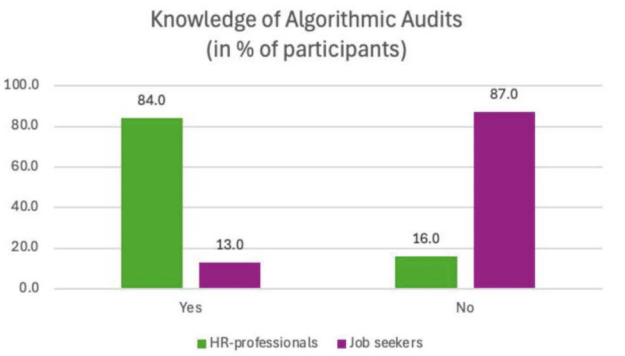


Figure 10: Have you ever heard of the New York City Law 144?

The results imply that there is a lack of information about algorithmic audits in the public but that most HR professionals who use AI have at least heard about such audits.

Experience with Algorithmic Audits

About half (51%) of the HR professionals indicated that they have worked with HR-AI tools that have been audited, while 24% were unsure and 24% indicated they had not. Of the HR professionals who have confirmed that they have experience with algorithmic audits about half (55%) answered our optional question about what kind of audit was performed. The rest did not explain what kind of audit was performed. The answers showed a full variety of audits (safety, security, privacy, bias, compliance, ethics, tax, accounting, functionality, accuracy and more) with no common definition. Many of them did not go into any detail ("*general run-of-the-mill audit*", algorithmic audit, general audit, "*a thorough one*") or said they were not sure what kind of audit it was, indicating some lack of understanding of what the audit conducted was about. Some concentrated on who conducted the audit (human, third party, naming a specific company). Few provided more detailed descriptions. One HR professionals described the audit in a vague way as:

> *We performed tasks and tests to see what the AI was scanning and basing its analyzation on, making sure it complies with legal compliance, and making sure it wasn't discriminating anyone based on different factors,*

Another provides more details but still lacks specifics:

> *We perform audits of our learning models, our datasets, and our results versus manual and more historic/traditional models.*

Only a couple of participants named classic algorithmic audits like bias audits (3 participants) or compliance audits (3 participants).

The results suggest that HR professionals notice the occurrence but not the type of audit. It also is confirming that there are a variety of audits out in the field. The concept of an algorithmic audit covers multiple aspects for our participants and can vary from one another substantially. Overall, the results indicate that there is no common understanding of what an algorithmic audit is among HR professionals. Our findings show that job seekers in general have almost no knowledge regarding algorithmic auditing, and HR professionals have a basic knowledge and some experience with them but mostly lack a deeper understanding.

RQ3: Stakeholders' Perspectives on Algorithmic Audits of HR-AI Tools

Importance of Algorithmic Audits

We asked the participants in question 19 how important they found it to audit AI. We expected that most respondents would find algorithmic audits important but were interested in what different nuances their answers would show. Several participants just echoed part of the definition, simply answered with 'Yes' or a similar answer, and did not come up with additional thoughts. Many answered very vague. Others just simply indicated that it is important. However, there was an overwhelming majority of job seekers and HR professionals who found audits very or extremely important.

Figure 11: Importance of Audits (n-261)

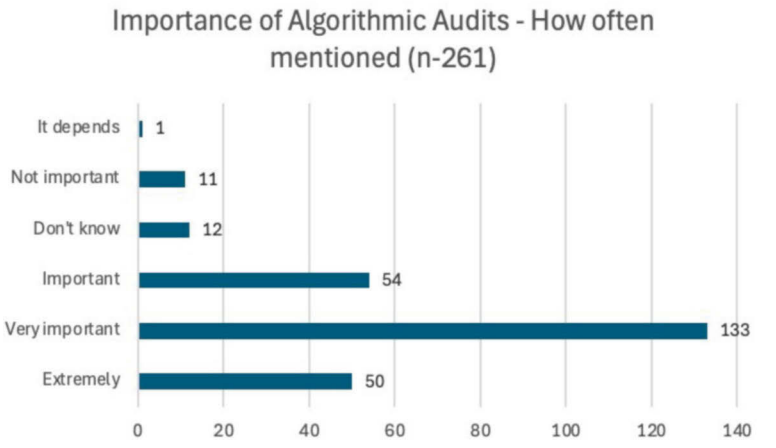
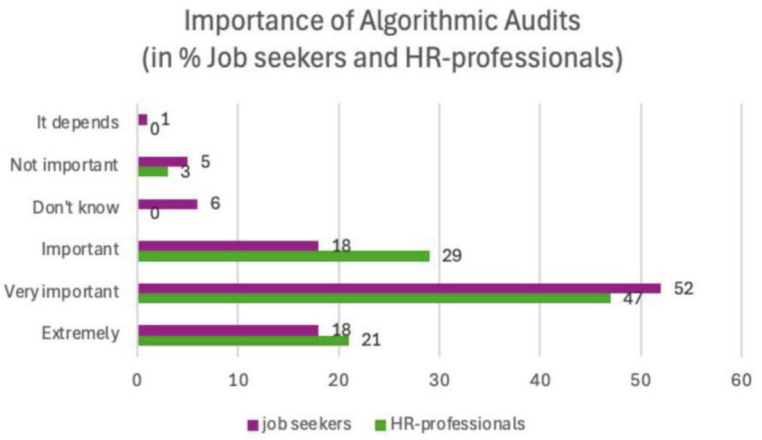


Figure 12: Importance of Audits for different stakeholders (in %)



Even though there are some slight differences in the degree of importance and there are more job seekers who do not know what to say, there is no remarkable difference between the job seekers and HR professionals. Both groups find the audits mostly very or extremely (including mentions like crucial, vital, essential etc.) important (job seekers 70%, HR professionals 68%).

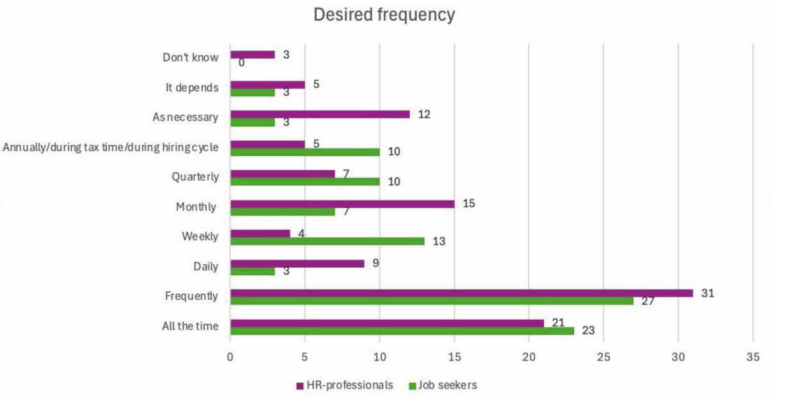
An additional comment emphasized the importance of audits in view of the novelty of AI.

> *It is extremely important, especially considering how early on we are with AI in the grand scheme of things. It being used for business, particularly with human interests in mind, it should be audited constantly. Used more as an assistant rather than a tool to use on its own. (jobseeker quote)*

Desired Frequency of Algorithmic Audits

In the same question, we asked how often they should be performed. The majority wanted them surprisingly frequently.

Figure 13: Desired frequency of audits (147 mentions)



Like the importance of algorithmic audits, there is no substantial difference regarding the frequency of algorithmic audits. About half of both groups want the audits all the time or frequently (Job seekers 52%, HR professionals 50%). The variations in the different categories are probably due to the low number of mentions and therefore less remarkable. The overall trend is similar.

A couple of participants suggested adapting the frequency over time and depending on needs.

> *With AI being a relatively new tool for HR, I think audits are very important and should be done at least once a hiring cycle. As the AI becomes more 'experienced' the audits probably need to be done less frequently but since each business has its own unique hiring practices and every position is different, audits should be required for every position for several hiring cycles before the company can choose to decrease to once or twice a year. (jobseeker quote)*

Other Thoughts about Algorithmic Auditing

Human-in-the-loop

It was striking that many participants particularly mentioned that it should be performed by humans, even though this was not included in the definition. It came up 17 times by job seekers and once by HR professionals. Having a human in the loop seems to be important.

Performance of Audits

In addition, there were some quite thoughtful suggestions on how they could be done. One jobseeker suggested that the AI decisions should be checked manually to allow for some more flexible decisions, indicating that a human touch is sometimes needed. Some suggest that people working in human resource should check whether they would reach the same output, and this should be done more frequently in the beginning and over time less frequently. One participant recommends doing this as a 'blind' test. HR professionals mentioned the independence of the auditor a bit more than job seekers. They also had some more concrete ideas in what phases of the AI lifecycle the audit should happen. One participant pointed out the auditing is particularly important for human resources. The responses indicate that the users asked in this survey find algorithmic audits in HR very to extremely important, would appreciate frequent checks of algorithms through audits and have some interesting ideas on how this could be done. Direct quotes underlining this can be found in the annex.

Independence of Auditor

Another striking difference in our survey between HR professionals and job seekers is the difference between the trust level in HR-AI tools depending on whether they have been not audited (figure 14), self-assessed (figure 15) or audited by an external auditor (figure 16).

HR professionals trust non-audited tools with an average of 2.68 while job seekers trust them with an average of only 1.55. Self-assessed tools are trusted with an average of 3.51 by HR professionals and with an average of only 2.39 by job seekers, and tools audited by an external auditor are trusted with an average of 4.9 by HR professionals and 3.78 by job seekers. T-tests performed for all three graphs showed a 2-tailed significance of $p<0.001$.

The trust HR professionals would have in self-assessment as opposed to job seekers is surprising. It is also noticeable that HR have more trust in non-audited tools than job seekers. Both groups have most trust in audits conducted by an external auditor.

The findings underline that most of the stakeholders deem algorithmic audits as quite important, stress values such as humans in the loop and independency of the auditor and have concrete and thoughtful suggestions on how audits could be conducted.

Figure 14: Trust with no audits

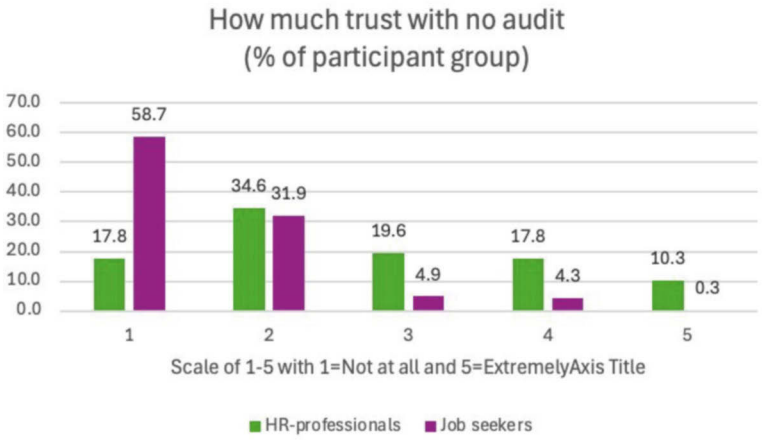


Figure 15: Trust with self-assessment

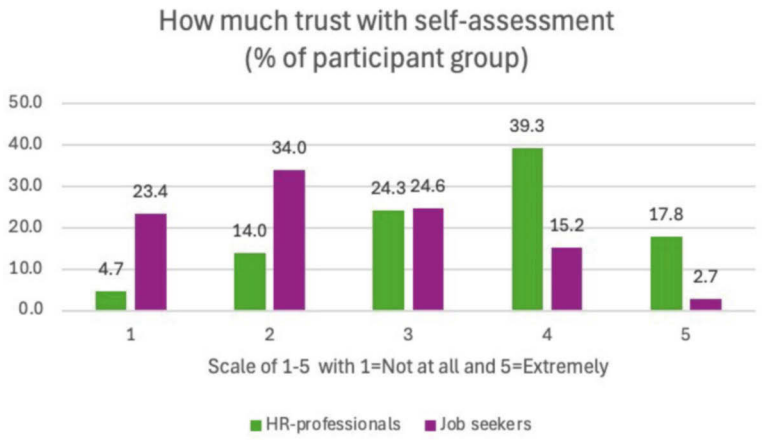
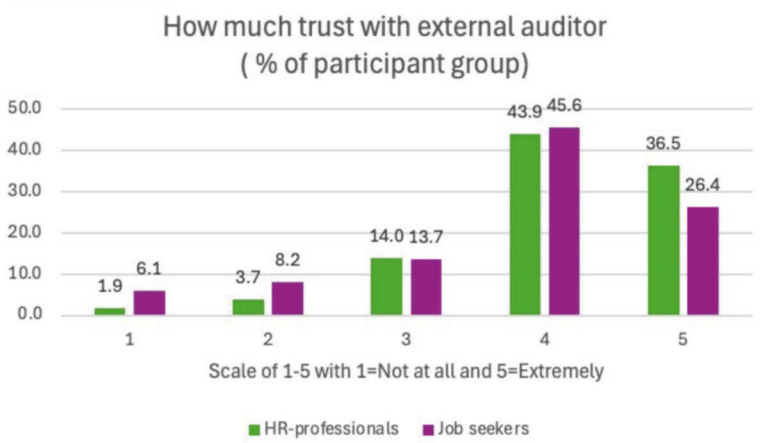


Figure 16: Trust with external auditor



Discussion

Trust of Stakeholders in AI and in HR-AI tools

Our findings demonstrate that HR professionals trust AI in general and HR-AI tools in particular substantially more than job seekers do. Job seekers show a high amount of distrust towards the tools. This attitude shows some connection to their knowledge of and experience with AI.

Trust requires knowledge and experience of the technology that should be trusted. While HR professionals are using the AI tools directly and therefore know when the tools are used, our survey shows that most job seekers do not know when and how these tools are used even though they are directly affected by the use. Current regulation like the New York City Law and the European AI Act aim to fill the knowledge gap by requiring that the use of AI in the hiring process should be made public, as detailed above. Such regulatory requirements seem to fill a need on the side of the stakeholders who would like to be more informed. Due to the limited scope of the NYC law these requirements will, however, not affect most job seekers in the US. The literature has also shown that currently the NYC law is rarely applied and that published audit reports lack accessibility and usability (Groves et al. 2024; Wright et al. 2024). It would be desirable to provide more and widespread

transparency for job seekers, similarly to the European AI Act. There were some movements towards more transparency following the Executive Order of the Biden administration 14110 regarding federal offices requiring them to be more transparent (DOL 2024). As a suggested Promising Practice, federal contractors should *"provide advance notice and appropriate disclosure to applicants, employees, and their representatives if the contractor intends to use AI in the hiring process or employment decisions that allows individuals to understand how they are being evaluated"* (DOL 2024). This Executive Order is, however, not in effect anymore.

More concrete and widespread regulation and standards expanding to the whole nation and including the private sector would help the desire of users of HR-AI tools to be better informed about the use of AI in employment decisions and help protect them from adverse impacts.

There is a striking difference between the trust level of job seekers and HR professionals regarding HR-AI tools. Of the two stakeholder groups job seekers are much more vulnerable than the HR professionals as they are affected on a more personal level. There is also a striking power imbalance in the relationships between job seekers and HR professionals which is exacerbated using advanced algorithms. While the workers are subjected to algorithmic management, the managers are implementing the decisions (Jarrahi et al. 2021). Furthermore, potential algorithmic discrimination can extend beyond traditional power imbalances and create new unprotected, vulnerable groups, such as single mothers, homeless people, or users who share a browser history (Rigotti & Fosch-Villaronga 2024). This power imbalance and the vulnerability could explain a higher level of mistrust by job seekers as they face a higher risk, such as the risk to be not hired or let go from their position.

As a first step, the use of AI should be made more transparent to allow all stakeholders a better understanding and more experience with the use of AI to allow for more calibrated trust in the future. As it is unlikely that there will be appropriate regulation requiring more transparency in the near future, the government, academia, and civil society should foster more education about the use of AI to improve AI literacy. Improved AI literacy could lead towards more calibrated trust.

Knowledge about Algorithmic Audits and in particular Algorithmic Audits in HR

The findings show that the knowledge about algorithmic audits is very sparse, even though audits are aimed to help inform the public. Job seekers have almost no knowledge about them, and even though many HR professionals claim to know what they are, they seem to have little actual knowledge about how they are conducted.

So far, not much research has been conducted regarding the communication of auditing results to users (Scharowski et al. 2023). As the literature points out, there are some significant issues in making the results of audits public. Wright et al. (2024) show the difficulties for users to find the publications of conducted bias audits in New York. Our study also found that the knowledge about the existing New York Law 144 is very limited. Regular job seekers have hardly heard about algorithmic audits and even less about the NYC law. Hiring managers and recruiters have some knowledge, but it is rudimentary. It seems that there is a significant lack of communication regarding algorithmic audits, even though job seekers and HR professionals expressed that they find them important. To have users and the public benefit from AI audits they need to be easily accessible and usable as Wright et al. (2024) point out. How this could be achieved has so far been understudied. A study by Scharowski et al. (2023) investigates the use of certification labels in an empirical mixed-method study and concludes that they could lead to increased trust and willingness to use AI, especially in high-risk scenarios. As shown above, employment decisions are considered high risk. The Responsible AI Institute has developed a model for certificates for responsible AI (Casovan 2022) using a comparable approach. Similar ideas have been around regarding the use of datasets, such as the research regarding data nutrition labels that provide information to dataset users to mitigate potential harm (Chmielinski 2022).

The ideas above can serve as raw models for the communication of AI audits. Further studies are needed to see how these could be implemented and how the acceptance would be.

Besides improving the way information about algorithmic audits is communicated, the government, academia, and civil society should also aim to improve general information about the nature of algorithmic audits. Once the general public has more knowledge about audits and what they can and cannot do, they can ask for more algorithmic audits, and they will also know better how to evaluate individual audits.

Perspectives of Stakeholders and Importance of Algorithmic Audits

The comments to our open text questions clearly showed that users across the board find algorithmic audits very if not extremely important and would support frequent audits conducted by humans.

So far, the perspective of HR professionals and job seekers regarding algorithmic audits in human resource has not been explored in more depth and has not been considered much. Some research regarding algorithmic audits introduces the idea of including users in audits. This is mostly suggested in AI content moderation and social media (Lam 2022, Shen 2021). A comparable direct input of users for the use of AI in employment decisions seems less applicable. The concerns and worries of users in such decisions should, however, be part of the discussion. As major stakeholders and the most vulnerable group, their opinion and ideas should be more included in designing proper algorithmic audits. This could happen in form of participatory design sessions where various stakeholders are actively included in the design process of audits. Another approach would be to use a value based design approach and inform the design of algorithmic audits by qualitative and quantitative research such as this study. Our research shows that our participants had very concrete and creative ideas. Furthermore, it seems that once users learn more about the concept of audits, they find them useful. Our survey demonstrated that job seekers have mixed experience with AI used in HR and would like to know more about the use. Established algorithmic audits communicated to them in a proper way could potentially improve their experience and increase trust. Many participants also stressed the importance of the human factor in the process. This should be another consideration for designing audits in the future.

Limitations and Future Work

A main limitation of our research is that users of HR-AI tool do not know much about algorithmic audits, and on the job seekers side they are not even always aware of the use of AI in HR. It is difficult to receive substantial and useful information about an area the respondent is unfamiliar with. This research is meant as a first step towards exploring what users might feel about algorithmic audits. It can so far only provide trends and first ideas to a very limited extent, as audits of HR-AI tools are not common yet.

Once algorithmic audits are more established and common, new studies should explore what stakeholders think about algorithmic audits, and what designs of such audits work, and which do not. Further studies could be done in the form of participatory design, including all stakeholders in the discussion. Future research should also extend beyond the field of HR to explore how different fields might need different approaches regarding algorithmic audits.

Conclusion

One key takeaway from our study is that job seekers have little to no knowledge about the use of AI during the application and hiring process or throughout their employment but would overwhelmingly like to know when AI is used. HR professionals who consciously use AI tools and have some experience with the tools have a higher trust overall in the tool. This could indicate that knowledge and experience with AI tools could increase trust in the tools. Current regulation aims to provide stakeholders with information about the use of AI in HR but so far is only very rarely applicable and not widely used in the field. We recommend further regulation, standards, and/or business best practices to make the use of AI in HR transparent for all stakeholders. Once job seekers have more information on the use of AI in HR and more experience with it, it should be further researched whether this helps them trust the tools and whether this trust is calibrated to the trustworthiness of the tool.

The surveyed stakeholders overall have a limited understanding of algorithmic audits and have not heard much about the New York City Law 144 requiring bias audits for automated employment decisions. While HR professionals have more knowledge about algorithmic audits, their understanding of the concept is rudimentary and there is no common understanding of what an algorithmic audit constitutes. More well communicated information about algorithmic audits and the results of conducted audits is needed to provide users with information that could influence their experience with and trust in AI. Our survey results indicate that further knowledge about what an audit is and how it is conducted is important to stakeholders and has the potential to affect their calibrated trust in the tools. Such information needs to be accessible, usable, and well communicated.

Once HR professionals and job seekers learn more about the concept of algorithmic audits, they find them very if not extremely important. While most are not sure how they should be conducted, they find that they should be per-

formed frequently and thoroughly and should have a human element. Our research shows that there is a wide interest in algorithmic audits once more information is available and that HR professionals and job seekers have creative and thoughtful opinions regarding algorithmic audits. The voices of the stakeholders should be heard to improve algorithmic audits in the future. It should be further explored how these stakeholders' opinions and concerns could be included in the design of algorithmic audits to give them a voice in the process.

Acknowledgements

The research was supported by a grant from the Notre Dame-IBM Technology Ethics Lab.

Annex

Complete questionnaire

	Question	**Kind of question**
Q1	Do you have prior experience working as a Human-Resource manager or as a recruiter?	Multiple choice
Q2	How many years of experience do you have as a Human-Resource manager and/or as a recruiter?	Multiple choice
Q3	In our study we are examining Artificial Intelligence tools used in Human Resources. We understand Human Resource AI Tools (HR-AI tools) as AI tools managing any part of an employee's life-cycle (including recruiting, hiring, onboarding, training, or promotional decisions). Have you in your professional life used HR-AI tools?	Yes – No
Q4	Are you actively seeking a job, or have you recently been on the job market as a job seeker?	Multiple choice
Q5	What is your age?	Multiple choice
Q6	What is your gender?	Multiple choice

	Question	Kind of question
Q7	What is your race/ethnicity? Please select as many as apply!	Multiple choice
Q8	What is the highest level of degree or school that you have completed?	Multiple choice
Q9	In this survey we understand Artificial Intelligence (AI) as ... How much do you trust Artificial Intelligence in general?	Slider question (scale of 1–5)
Q10	How often have you used HR-AI tools in your position as a Human Resource Manager or a recruiter?	Slider question (scale of 1–5)
Q11	In our study we are examining AI tools used in Human Resources. We understand Human Resource AI Tools (HR-AI tools) as AI tools managing any part of an employee's life cycle (including recruiting and hiring). Were you told that AI was used in the application and hiring process when you applied for jobs?	Only for job seekers Slider question (scale of 1–5)
Q12	How important is it for you to know if AI is used in the application or hiring process?	Only job seekers Slider question (scale of 1–5
Q13	How was your overall experience using HR-AI tools as a Human Resource Manager or recruiter?	Only HR Slider question (scale of 1–5)
Q14	How was your overall experience with HR-AI tools as a jobseeker	Only job seekers Slider question (scale of 1–5)
Q15	How much do you trust Human resource AI tools?	Slider question (scale of 1–5)
Q16	Please mark position four with the slider!	Slider question (scale of 1–5)
Q17	Have you ever heard of Algorithmic auditing?	Yes – No
Q18	Have you heard of the New York City Law 144?	Multiple Choice
Q19	Since AI does not always work as expected or intended, in some cases, organizations audit AI, which checks to ensure that they comply with legal and ethical standards. How important is it to audit AI, and when and how should this be done?	Open text
Q20	Have you ever worked with an HR-AI tool that has been audited?	Only HR Yes – Unsure – No

	Question	**Kind of question**
Q21	What kind of audit was performed?	Open text
Q22	How important are each of these aspects of audits of AI for you? (Safety and reliability of the AI tool, Legal compliance, Discovering of bias, Potential harm to the public)	Slider question for each (scale of 1–5)
Q23	Are there other important aspects of audits of AI for you that were not mentioned above? (Please explain)	Open text
Q24	What level of trust would you have in HR-AI tolls in these different situations (HR-AI tool is not audited, HR-AI tool is self-assessed by the company, HR-AI tool is audited by an external auditor)	Slider question for each (scale of 1–5)
Q25	How important is it to you that audits are mandatory?	Slider question for each (scale of 1–5)
Q26	How important is it to you that audits of AI are conducted by organizations that are required to meet established standards?	Slider question for each (scale of 1–5)
Q27	What is most important for you regarding the AI auditor?	Multiple choice
Q28	How important is it for you to understand how the audit of AI is conducted?	Slider question for each (scale of 1–5)
Q29	How important is it for you that the audit explains the AI and its consequences?	Slider question for each (scale of 1–5)
Q30	What do audits of AI evaluate?	Multiple choice
Q31	How would you like to see the results of an audit/certification? (You can choose more than one)	Multiple choice
Q32	How do you think the audit report should be shared with the public?	Multiple choice
Q33	How important is it to you that it will be required that the audit results are shared?	Slider question for each (scale of 1–5)
Q34	How important is it for you to know what actions the company took in response to the audit?	Slider question for each (scale of 1–5)
Q35	Please share any additional comments you have regarding audits of AI of Human Resource AI tools	Open text

Quotes regarding the performance of audits

Quote 1
It should be checked for reasonableness at least 50% of the time, by reviewing the same material manually and see if the answers are the same as those arrived at by AI. There could be some points that may have a "gray area" that should not be interpreted strictly by the book. AI cannot determine these "iffy" points. (jobseeker quote)

Quote 2
It is extremely important to audit AI, Computer(s) can and do fail and oftentimes need to be slightly reprogrammed for the original intent. This should be done by a person who is looking to have the same output the AI is intended for. As an example, if AI is working for HR, then someone who works in HR should check frequently to ensure the AI is working as intended. I would see this should be done around 50% of the time at the beginning, and then less frequently if the AI appears to be working as intended. (jobseeker quote)

Quote 3
I think it is very important because the decisions or advice given may not comply with standards. I believe this should be done every time AI is used to ensure its efficacy and areas of weakness. This will help evolve the technology until it is up to par with standard both legal and ethical. This should be done with a group of individuals who do not know they are analyzing the work of AI. A blind examination such as this can help prevent bias towards technology of this kind. (jobseeker quote)

Quote 4
It needs to be constantly audited and by third party sources. AI can be extremely dangerous, especially if it is connected to external sources like the internet or if it is taught coding. When these influences are introduced it's imperative that assessments be made as often as possible to verify input and output from the system as well as internal monitoring for processes being done. (HR professional quote)

Quote 5
Auditing artificial intelligence is of great importance because as AI is widely applied, reasonable regulations are needed to ensure its compliance. Auditing should be conducted during the development stage of AI, and an in-

dependent auditing team should be established to maintain independence from the developers and carry out objective audits. (HR professional quote)

Quote 6

I think it is incredibly important to audit AI, especially when it has to do with HR. Human resources can carry vital information for individuals and if AI is not frequently audited and kept working properly, important data could be lost or leaked, and people can be put in danger. (jobseeker quote)

References

Abbey, James D./Meloy, Margaret G. (2017): "Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality", in: Journal of Operations Management, 53–56, pp. 63–70. https://doi.org/10.1016/j.jom.2017.06.001.

Ajunwa, Ifeoma (2020): "An auditing imperative for automated hiring systems", in: Harv. JL & Tech., 34, pp. 621–699.

Bandy, Jack (2021): "Problematic machine behavior: A systematic literature review of algorithm audits", in: Proceedings of the ACM on Human-Computer Interaction 5, No. CSCW1, pp. 74:1-74:34.

Braun, Virginia/Clarke, Victoria (2021): Thematic analysis: A practical guide, London: Sage.

Bujold, Antoine/Roberge-Maltais, Isabelle/Parent-Rocheleau, Xavier/ Boasen, Jared, Sénécal, Sylvain, & Léger, Pierre Majorique (2024): "Responsible artificial intelligence in human resources management: a review of the empirical literature", in: AI and Ethics, 4(4), pp. 1185–1200.

Casovan, Ashley/Shankar, Var/Brooks, Hannah/Faveri, Benjamin/Cairns, Stephanie/Dawson, Phil/Lefaivre, Alyssa/Scully, Kara/ Hererra, Raidhy (2022): "A certification for responsible AI", in: Responsible Artificial Intelligence Institute, pp. 1–14, retrieved from https://assets.ctfassets.net/rz1q59puyoaw/6MYkmmV9wk0ZmrBTOluQG7/fd830fcc413fa54290f5bc9dbbd10f0b/The_Responsible_AI_Certification.pdf.

Chavan, Prasad. R./Chandurkar, Yash/Tidake, Ankita/Lavankar, Gaurav/ Gaikwad, Suhani/Chavan, Rohit (2024): "Enhancing recruitment efficiency: An advanced applicant tracking system (ATS)", in: Industrial Management Advances, 2(1), pp. 6373–6373.

Chen, Zisheng (2023): "Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices", in: Humanities and Social Sciences Communications 10, no. 1, pp. 1–12. https://doi.org/10.1057/s41599-023-02079-x.

Chmielinski, Kasia/Newman, Sarah/Taylor, Matt/Joseph, Josh/Thomas, Kemi/Yurkofsky, Jessica/Qiu, Yue Chelsea (2022): "The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence", in: arXiv preprint arXiv:2201.03954.

Costanza-Chock, Sasha/Raji, Inioluwa Deborah/Buolamwini, Joy (2022): "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem", in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1571–83. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533213.

Dawson, Phil (2022): "AI Audits are Coming to HR". in: Armilla AI. Medium. https://medium.com/@ArmillaAI/ai-audits-are-coming-to-hr-c1386fe6c0e.

De Winter, Joost/Dodou, Dimitra (2010): Five-Point Likert Items: t Test Versus Mann–Whitney–Wilcoxon", in: Practical Assessment, Research, and Evaluation 15(1).

Department of Labor (DOL) (2024): "Artificial Intelligence and Equal Employment Opportunity for Federal Contractors", https://www.dol.gov/agencies/ofccp/ai/ai-eeo-guide.

Digital Regulation Cooperation Forum (DRCF) (2022): "Auditing algorithms: the existing landscape, role of regulators and future outlook", https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022.

Engler, Alex (2021): "Independent Auditors Are Struggling to Hold AI Companies Accountable", in: Fast Company, https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue.

European Commission (2024): "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance)". http://data.europa.eu/eli/reg/2024/1689/oj.

European Parliament, Council of the European Union (2022): "Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 Octo-

ber 2022 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC (Digital Services Act)", in OJ L 277, 27.10.2022, EUR-Lex, pp. 1–102.

Fuchs, Lindsy (2023): "Hired by Machine: Can a New York City Law Enforce Algorithmic Fairness in Hiring Practices?", in: Fordham J. Corp. & Fin. L., 28, pp. 185–266.

Geiger, Stuart R/Tandon, Udayan/Gakhokidze, Anoolia/ Song, Lian/Irani, Lilly (2023): "Rethinking Artificial Intelligence: Algorithmic Bias and Ethical Issues| Making Algorithms Public: Reimagining Auditing From Matters of Fact to Matters of Concern", in: International Journal of Communication, 18, pp. 634–655.

Goodman, Ellen P./Trehu, Julia (2022): "Algorithmic audit Washing and Accountability", in: SSRN Electronic Journal, 2022. https://doi.org/10.2139/ssrn.4227350.

Groves, Lara/Metcalf, Jacob/Kennedy, Alayna/Vecchione, Briana/Strait, Andrew (2024): "Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime", in: arXiv. https://doi.org/10.48550/arXiv.2402.08101.

Harlan, Elisa/Schnuck, Oliver (2021): "Objective or biased: On the questionable use of AI in job applications", Bayerischer Rundfunk. Retrieved from: https://web.br.de/interaktiv/ki-bewerbung/en/

H.R. 6216 — 116th Congress: National Artificial Intelligence Initiative Act of 2020.

Huh Iksoo/Gim Jungsoo (2025): „Exploration of Likert scale in terms of continuous variable with parametric statistical methods", in: BMC Medical Research Methodology, 25(1):218. doi: 10.1186/s12874-025-02668-1. PMID: 41023822; PMCID: PMC12482090.

Hunkenschroer, Anna Lena/Kriebitz, Alexander (2023): "Is AI Recruiting (Un)Ethical? A Human Rights Perspective on the Use of AI for Hiring." In: AI and Ethics 3, no. 1 (2023), pp.199–213. https://doi.org/10.1007/s43681-022-00166-4.

IL HB3773. 103rd General Assembly: Illinois AAI Act of 2024

Jarrahi, Mohammad Hossein/Newlands, Gemma/Lee, Min Kyung/Wolf, Christine T./Kinder, Eliscia/Sutherland, Will (2021): "Algorithmic management in a work context", in: Big data & society, 8(2), 20539517211020332.

Kung, Franki Y.H./Kwok, Navio/Brown, Douglas J. (2018): "Are attention check questions a threat to scale validity?", in: Applied Psychology, 67(2), pp. 264–283.

Lam, Khoa/Lange, Benjamin/Blili-Hamelin, Borhane/Davidovic, Jovana/Brown, Shea/Hasan, Ali (2024) "A Framework for Assurance Audits of Algorithmic Systems.", in: arXiv, https://doi.org/10.48550/arXiv.2401.14908.

Lam, Michelle S./Gordon, Mitchell L./Metaxa, Danaë/Hancock, Jeffrey T./Landay, James A./Bernstein, Michael S. (2022): "End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior", in: Proceedings of the ACM on Human-Computer Interaction. 6, CSCW2, Article 512 (November 2022), 34 pages. https://doi.org/10.1145/3555625

Lassiter, Tina B./Fleischmann, Kenneth R. (2024) ""Something Fast and Cheap" or" A Core Element of Building Trust"?-AI Auditing Professionals' Perspectives on Trust in AI.", in: Proceedings of the ACM on Human-Computer Interaction, *8*(CSCW2), pp.1-22.

Li, Lan/Lassiter, Tina/Oh, Johee/Lee, Min Kyung (2021): "Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring", in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 166–176.

Mökander, Jacob (2023): "Auditing of AI: Legal, Ethical and Technical Approaches", in: Digital Society 2, no. 3, 49. https://doi.org/10.1007/s44206-023-00074-y.

Mökander, Jacob/Floridi Luciano (2021): "Ethics-Based Auditing to Develop Trustworthy AI", in: Minds and Machines 31, no. 2: pp. 323–27. https://doi.org/10.1007/s11023-021-09557-8.

National Telecommunications and Information Administration (NTIA) (2024): AI Accountability Policy Report. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report.

New York City Council—File #: Int 1894–2020. (n.d.). https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9.

Norman, Geoff (2010): "Likert scales, levels of measurement and the "laws" of statistics", in: Advances in Health Sciences Education 15, 625–632, https://doi.org/10.1007/s10459-010-9222-y.

O'Rourke, Diane/ & Lakner, Edward (1989): "Gender bias: Analysis of factors causing male underrepresentation in surveys", in: International Journal of Public Opinion Research, 1(2), pp. 164–176.

Raghavan, Manish/Barocas, Solon/ Kleinberg, Jon/Levy, Karen (2019), in: "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices", in: arXiv:1906.09208 [Cs], https://doi.org/10.1145/3351095.3372828.

Raji, Inioluwa Deborah/Smart, Andrew/White, Rebecca N./Mitchell, Margaret/Gebru, Timnit/Hutchinson, Ben/Smith-Loud, Jamila/Theron, Daniel/Barnes, Parker (2020): "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing", in: arXiv, http://arxiv.org/abs/2001.00973.

Rigotti, Carlotta/Fosch-Villaronga, Eduard (2024): "Fairness, AI & recruitment", in: Computer Law & Security Review, 53, 105966.

Sandvig, Christian/Hamilton, Kevin/Karahalios, Karrie/Langbort, Cedric (2014): "Auditing algorithms: Research methods for detecting discrimination on internet platforms", in: Data and discrimination: converting critical concerns into productive inquiry, 22(2014), pp. 4349–4357.

Scharowski, Nicolas/Benk, Michaela/Kühne, Swen J/Wettstein, Léane/ Brühlmann, Florian (2023): "Certification labels for trustworthy ai: Insights from an empirical mixed-method study", in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 248–260.

Schellmann, Hilke (2024): "The Algorithm", New York, N.Y., Hachette Books.

Selinger, Evan/Leong, Brenda/Cahn; Albert Fox (2021): "AI Audits: Who, When, How… Or Even If?", in: Collaborative Intelligence: How Humans and AI are Transforming our World, MIT Press.

Shen, Hong/DeVos, Alicia/Eslami, Motahhare/Holstein, Kenneth (2021): "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors:, in: Proceedings of the ACM on Human-Computer Interaction. 5, CSCW2, Article 433. 2021. 29 pages. https://doi.org/10.1145/3479577

Smith, William G. (2008): "Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior", Online submission, https://eric.ed.gov/?id=ED501717

US Bureau of Labor Statistics (BLS) (2023): "Occupational Outlook Handbook, Human Resource Managers", https://www.bls.gov/ooh/management/human-resources-managers.htm#:~:text=a%20master's%20degree.-,Education,conflict%20management%20may%20be%20helpful.

Vieira, Pedro Cosme Costa (2016): "T-Test with Likert Scale Variables", available at: SSRN: https://ssrn.com/abstract=2770035 or http://dx.doi.org/10.2139/ssrn.2770035.

Wilson, Christo/Ghosh, Avijit/Jiang, Shan/Mislove, Alan/Baker, Lewis/Szary, Janelle/Trindel, Kelly/Polli, Frida (2021): "Building and auditing fair algorithms: A case study in candidate screening", in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 666–677.

Wright, Lucas/Muenster, Roxana Mike/Vecchione, Briana/Qu, Tianyao/Cai, Pika/Smith, Alan/Matias/J. Nathan/Metcalf, Jacob (2024): "Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability", in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (pp. 1701–1713).

To be Cared for or Deceived?

Operationalizing Ethics in the Case of Elderly Care Robots under the European AI Act

Gaia Contu

Abstract: *Does the use of social robots in elderly care constitute a form of deception toward the individuals concerned? Focusing on the framework of the European AI Act, the paper examines how such concerns can be practically addressed under Article 5 on prohibited AI practices. Drawing on an analysis of the non-neutrality of technology in the fields of artificial intelligence and robotics, and referring to well-known cases such as COMPAS and the Amazon recruitment algorithm, the study highlights the ethical relevance of design choices in social robotics. It argues that ethical considerations are always present in technological development and must be operationalized from the earliest stages of designing robots for elderly care. Through a theoretical investigation grounded in Theory of Affective Coordination, pragmatism and relational ethics, the contribution shows that social robots in elderly care are not deceptive per se, and that potential risks of deception can be meaningfully addressed through concrete design choices and appropriate normative guidelines.*

Keywords: *Social robots; non-neutrality of technology; value-sensitive design; elderly care robots; ethics of technology; AI act; robot ethics*

Introduction

In a world where emerging technologies such as AI and robotics are rapidly expanding and playing a crucial geopolitical and economic role, the European Artificial Intelligence Act (European Commission 2024) represents one of the first attempts to regulate a field that has typically developed in a legislative vacuum. This effort is significant, given that our relationship with technology is

not occasional but an integral part of daily life. Smartphones can almost be seen as extensions of human bodies; many people use large language models like ChatGPT daily, construct identities and affective relationships through social media, wear smartwatches to monitor vital parameters, and artificial intelligence is being used in socially relevant contexts such as healthcare and public administration.

One sensitive domain where this issue warrants particular attention is elderly care. The field is increasingly integrating digital tools, such as telemedicine devices, to ensure high-quality care at home, particularly in marginalized areas and for patients with limited mobility; at the same time, a growing ethical concern relates to the use of robotic assistants (Sharkey/Sharkey 2012, Anderson/Felzmann 2024). In fact, a rapidly developing area is that of "Elderly Care Robots" (ECRs): robotic devices designed to support older adults across a wide range of tasks, from providing physical assistance and help with daily activities, to offering personalized support and even acting as social companions to alleviate loneliness (Hoppe et al. 2020, Yen et al. 2024, Ahmed [Buruk/Hamari] 2024). Among these, the most ethically controversial use is arguably that of social robots. These machines are designed to simulate human-like interaction and characteristics, including language use, appearance, and behavioral cues. This has led several scholars to question the authenticity, reciprocity, and potential for deception in human-robot relationships — particularly when the users belong to vulnerable populations, such as children or the elderly (Sparrow/Sparrow 2006; Sharkey/Sharkey 2012; Turkle 2006; van Wynsberghe 2016).Thus, an ethical, social, and legal analysis is crucial to protect vulnerable users and to provide clear guidelines and normative boundaries for developers, users, and distributors of emerging technologies.

In this paper, I will explore how to integrate ethics and moral values into technology, with a particular focus on robots designed for the care and companionship of older adults. I will examine the phenomenon of human-robot emotional attachment and the associated risk of deception. The topic will also be addressed from a legal standpoint: according to Article 5 of the European AI Act, AI systems that deploy manipulative or deceptive techniques to impair user autonomy must be prohibited. Therefore, it becomes essential to determine whether social robots and Elderly Companionship Robots fall within this regulatory scope.

To this end, the paper will proceed as follows. First, I will explore whether and how ethics can be operationalized in technological design. This includes demonstrating that technology is never neutral and that law, technoscience,

and ethics are deeply entangled. Drawing on the framework of Science and Technology Studies (STS), the Social Construction of Technology (SCOT), and feminist epistemologies, I will illustrate, through various case studies, how technoscientific artifacts inevitably embed values, worldviews, and contextual assumptions. Building on this, I will argue that it has always been possible to shape technological development in alternative directions, countering the myth of technological determinism. In particular, the framework of Value-Sensitive Design suggests that not only is it possible, but indeed necessary, to integrate specific human values into the design of technological systems. Likewise, user-centered design provides concrete methodological guidelines for moral inquiry, emphasizing the importance of incorporating the lived experiences and perspectives of end users.

Having laid this groundwork, I will then focus on a case study concerning social robots — specifically, Elderly Care and Companionship Robots. I will begin by defining what constitutes a social robot and examining several documented cases of their deployment among vulnerable populations. To this end, I will present a brief literature review of empirical studies evaluating the effects of social robot use among elderly individuals, especially in care home settings. Following this, I will engage with the common critique that social robots, by mimicking social cues and emotional expressions, are inherently deceptive. I will challenge this assumption by offering an ethical and epistemological analysis rooted in a relational understanding of mind and emotions. On this basis, I will propose a revised definition of deception, one that is both theoretically grounded and practically applicable. This reconceptualization will serve as a foundation for outlining regulatory and design-oriented guidelines at three levels, from the abstract to the concrete. I will argue for the need for targeted field studies and suggest pathways for rethinking the design and legal framing of social robots in ways that protect users without necessarily hindering innovation or production.

Background

The issue of guaranteeing a high quality of life for the elderly constitutes a main social concern. This is also stated by the third of the UN Sustainable Development Goals (SDGs) (UN, 2015), which aims to promote social inclusion and to ensure healthy lives and well-being at all ages, with particular attention to the most vulnerable, including older adults. According to the WHO, in 2020 the

number of people aged 60 years and older outnumbered children younger than 5 years, and by 2050 they will nearly double from 12% to 22% (WHO 2024). This wing of population is particularly fragile not only for what concerns physical illnesses, but also with regard to mental health: around 14% of adults aged 60 and over live with a mental disorder, and according to the Global Health Estimates (GHE) 2019, "these conditions account for 10.6% of the total disability (in disability adjusted life years, DALYs) among older adults" (WHO, 2023). This constitutes a critical healthcare problem according to the WHO definition of health as a "state of complete physical, mental and social well-being" (WHO 1946).

Loneliness has a particular influence on the psychosocial well-being of the elderly and is often identified as related to chronic illness and poor self-rated health (Jané-Llopis/Gabilondo 2008). This condition might intensify in a socioeconomic context where increasing mobility often leads to a shortage in the availability of family care (Hoppe et al. 2020), and where the growing elderly population comes into conflict with the non-scaling nature of the human care workforce (OECD, 2020).

Moreover, caregiver burden represents a significant yet often overlooked issue. Informal care within families is typically characterized by silent, wearisome labour, disproportionately carried out by women. As feminist scholars have emphasized, this work is often taken for granted, socially and institutionally invisible, and marked by power asymmetries and a lack of formal recognition or support (Tronto 1993, Kittay 1999, Parks 2010).

In response to the needs of care and companionship in old age, increasing attention is being devoted to the role of socially assistive robots, such as homecare robots, personal assistance robots or companionship robots. These robotic devices hold considerable potential: not only can they enhance care, providing caregivers with practical support such as physical assistance, but also respond to social needs and alleviate loneliness in the users (Ahmed [Buruk/Hamari] 2024, Yen et al. 2024). However, in order to be both effective and ethically acceptable, such technologies must be integrated thoughtfully, following a genuinely human-centered approach. This requires careful consideration of moral values and social impacts, as well as of legal boundaries and practical feasibility. It also demands ongoing dialogue and collaboration among all relevant stakeholders: final users, technical developers, engineers, policy makers, families and caregivers, healthcare personnel and other key actors.

But how is it possible to integrate the potential of social robotics with the needs of vulnerable populations in the most *ethical* way possible? How can we

truly develop *ethical* technologies, where the call for human-centeredness is more than just a formal phrase? Is it really possible to embed ethical and human values into technology?

Is it Possible to Operationalize Ethics?

In fact, demanding that technology be "ethical" is a rather vague formulation. What does it actually mean for a technology to be ethical? If we take the definition of ethics as a systematic discourse and philosophical analysis of moral values — that is, of what is right or wrong, good or bad — we are immediately faced with a situation of pluralism (Neri, 2020). Not only are there different ethical approaches and frameworks, but on the moral level as well, values differ from individual to individual, are context-dependent, and are influenced by culture, society, and other factors. So, when we call for ethicality of robots, *which* values are we referring to? And where is the authority or final decision-making power when it comes to evaluation?

To address this question in the context of Elderly Care Robots (ECRs), a process of specification is necessary. We will therefore break down the overarching question into the following sub-questions:

1. Can technology in general embody or integrate moral values?
2. How can such values be implemented or operationalized in technological design and development?
3. Who is entitled to assess the ethicality of a technology?
4. Which specific values are at stake in the case of ECRs – and how can they be assessed and implemented in practice?

In this section, we will address the first three questions, examining the issue of how ethical values can be integrated into technology in general. In the following section, we will turn to the specific case of social robotics.

The Non-Neutrality of Technology

To answer the question of whether technology can embody or integrate moral values (sub-question A), it is important to revisit the long-standing and influential view of technoscience as a value-neutral enterprise. The idea of science

as an impartial, universal and objective endeavour remains popular today. In school, we learn formulas as if they have always been set in stone, and we hear about scientific discoveries and technological inventions as though their only connection to cultural and social contexts is serendipity (Carrada 2005, Bucchi 2002). The image of science conveyed in public discourse, from the news to science communication, is what Latour would have called *ready-made science*: certain, cold, unproblematic, its process of construction and stabilization sealed within a black box. This stands in contrast to the more accurate and dynamic concept of *science in the making*, which is open to controversy, uncertainty, reversals, and mistakes, where truth and efficiency emerge through compromises among relevant actors (Latour 1987).

Alongside this narrative of science, there persists a theoretical assumption that scientific progress follows a linear, inevitable trajectory, largely uninfluenced by cultural, sociopolitical, or geographical contexts. This view is called scientific "determinism" or "inevitability" and comprises the idea that technology develops independently of society and, rather than being shaped by it, determines the course of social development. (Bijker 2009). This idea originates from the 1922 essay by William Ogburn and Dorothy Thomas, *Are Inventions Inevitable? A note on social evolution*, where they argued that revolutionary inventions are simply the inevitable outcome of cultural and technical components. As they put it: "Given the boat and the steam engine, isn't the steamboat inevitable?" (Ogburn & Thomas 1922, Huyskes 2025). This deterministic view continues to shape how emerging technologies are understood and developed. In the case of robotics, for example, there is a widespread view that imagines a linear and necessary trajectory of progress leading toward a very specific prototype: humanoid robots (with masculine body design), equipped with artificial consciousness or artificial general intelligence, and potentially destined to rebel against humans. This image, popular in the West, is shaped by deep-rooted cultural and literary narratives, and it powerfully influences our collective imaginary (Allen [Wallach/Smit] 2006, Van Grunsven 2022). In parallel, is also common the perception of technology, including Artificial Intelligence, as unbiased, objective, and even superior to fallible human judgment. In the words of sociologist Ruha Benjamin, in fact, "many industries and organizations well beyond health care are incorporating automated tools, from education and banking to policing and housing, with the promise that algorithmic decisions are less biased than their human counterpart." (Benjamin 2019). Indeed, a great deal of technology has been introduced precisely to compensate for human error and limitations: from recruitment tools such as the robot

Tengai, originally marketed as an *unbiased* interviewer[1], to Diella, the Albanian virtual minister created with AI in 2025 to fight corruption[2].

Yet, a more critical look reveals that this deterministic, objective, and neutral narrative of scientific progress has been challenged and largely abandoned within the academic communities of social sciences, philosophy of science, and science and technology studies for nearly 60 years.

STS, SCOT and Feminist Epistemologies

From the 1970s onward, the doctrine of technological determinism and the thesis of scientific neutrality and universalism began to show descriptive fragility, and its assumptions have been critically challenged. Technological determinism was argued to be a poor research strategy because it entails a teleological, linear, and one-dimensional view of technological development,

In its place, research approaches have proliferated in which the social dimension of knowledge plays a central role (Bucchi 2002, Bijker [Hughes/Pinch] 1987; Huyskes 2025). The firsts were the Science–Technology–Society (STS) movement and the Sociology of Scientific Knowledge (SSK), which began to investigate scientists' social responsibilities, demonstrating that technologies are shaped by social practices and decisions and supporting the view of a coevolution of technoscience and society (Keulartz et al. 2004, Bijker 2009). These technology studies replaced the old deterministic view with a more constructivist approach, which sees technological artifacts as "the outcome of negotiations, in which many diverse actors are involved." In this approach, "there is rarely one single path of development but rather a number of potentially viable alternatives […], not completely autonomous at all, […] but rather a fairly random result of social interactions" which "require particular role patterns and lay down a specific 'geography of responsibilities.'" (Akrich 1992, Keulartz et al 2004)To strengthen this critique of technological determinism, it was necessary to show that the workings of technology were socially constructed, with an emphasis on the social (Bijker 2009, Keulartz et al 2004). This contribution came from two main frameworks: SCOT (Social Construction of Technology), in which the contributions of STS and SSK have converged, and Feminist Epistemologies.

1 https://www.tengai-unbiased.com, accessed in Feb 2022.

2 https://www.theguardian.com/world/2025/sep/11/albania-diella-ai-minister-public-procurement.

The SCOT movement aimed to unveil the role of "relevant social groups", "interpretive flexibility", "stabilization", and "closure" in technological development. To illustrate these concepts, Bijker uses the bicycle as a case study. He demonstrates that the design of the bicycle — far from being the result of a linear, inevitable, or purely technical process — evolved through complex social negotiations among diverse groups, such as bicycle producers, young athletic ordinary users, women cyclists, and anti-cyclists. Each group attributed different meanings to the technology: men saw it as a symbol of power, while women faced practical barriers like clothing constraints. These conflicting interpretations led to design changes, such as the introduction of pneumatic tires to enhance safety and comfort, and to the dominance of one model over others (Bijker 1995). To this critique, feminist epistemologies added the crucial insight that technological development not only integrates society but also reflects and reinforces its underlying power structures. Feminist scholars have shown that the so-called "universal subject" of science — assumed to produce objective and universally valid knowledge — was, in fact, neither universal nor neutral. Rather, it was a highly specific subject, situated in time, space, class, ethnicity, and gender. Typically, this subject was male, white, middle- or upper-class, and Western. This positionality inevitably shaped the production of scientific and technological knowledge.

For example, the case of the contraceptive pill reveals how technology embodies and reproduces gendered hierarchies and power relations. The development and deployment of the pill were influenced not only by scientific and medical agendas but also by social assumptions about female bodies, reproductive control, and gender roles. As discussed by Keulartz et al., if on the one hand the pill could be considered an emancipatory technology, on the other it continued to assign reproductive responsibility primarily to women, naturalizing this asymmetry through its design (Oudshoorn 2003, Tripaldi 2023, Keulartz et al. 2004). Moreover, the pill exposed women to health risks deemed unacceptable for men, using biological framings to obscure social choices and normatively charged evaluations (Oudshoorn 2003). Feminist epistemologies reveal how such technologies are products of androcentric knowledge systems, defining whose bodies are regulated and how. It is also noteworthy that not only the pill reflected existing norms but also actively reshaped norms, transforming the moral understandings of sexuality, altering practices, expectations, and responsibilities (Keulartz et al. 2004).

The same mechanism of integration of social injustices into technoscience does not only affect women, but also other marginalized groups, such as non-

white ethnicities or lower social classes. For example, in healthcare, pulse oximeters have been shown to be less accurate for Black patients because they were primarily designed and adjusted using lighter skin tones: an implicit bias that can affect life or death. (Al-Halawani et al. 2023). Or again, in urban design, the famous and still debated bridges built by architect Robert Moses in Long Island during the 1930s–1940s were so low that public buses could not pass underneath, effectively excluding poorer, often Black communities from certain areas (Winner, 1980; Huyskes, 2025). The examples of exclusions and practical damages are countless: from clinical trials where women are underrepresented — leading to results calibrated on the male physiology and thus often ineffective or harmful for female patients — to car crash tests based on male body, resulting in a higher likelihood of severe outcomes for women, as well as bulletproof vests and even office temperature settings, all originally designed around male standards (Criado Perez 2019).

The partial and exclusionary framing of technoscientific endeavour, which systematically excludes portions of the population from participating in the construction of knowledge, is therefore not only morally unjust, but has practical consequences and is epistemically limiting. When scientific and technological development is shaped by a limited set of social experiences and values, the resulting outputs are ineffective or even harmful for large segments of the population (Harding 1986; Grasswick 2018).

This dynamic becomes even more evident in the case of Artificial Intelligence and digital technologies, where the societal implications of biased data and exclusionary design perspectives are integrated and amplified (Bartoletti 2020; Huyskes 2025).

Digital Technologies as a Lens to Read Embedded Values

Around the late 2010s, particularly in the United States, a series of troubling cases involving algorithms, software, and digital technologies began to draw public and academic attention. While these technologies were introduced with the stated aim of overcoming human limitations — such as bias in hiring, risk assessment in the criminal justice system, or public housing allocation — they often produced outcomes that were deeply problematic.

In 2018, Joy Buolamwini, an African American researcher at MIT, discovered that the facial recognition software she was working on failed to accurately recognize her face. Further investigation revealed a significant disparity: while the maximum error rate for light-skinned males was just

0.8%, this increased dramatically for other demographic groups. The most misclassified group were dark-skinned females, that faced error rates of up to 34.7% (Buolamwini and Gebru 2018). Disparities like these can carry serious consequences, particularly when facial recognition technologies are used in sensitive domains like law enforcement, where misidentification can lead to wrongful arrests or deportation. Similarly, in 2016, an investigative report by *ProPublica* revealed that U.S. courts were using an algorithm known as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) to estimate the risk of criminal recidivism. Based on a dataset of over 10,000 defendants, the investigation found that the software falsely predicted a high risk of recidivism in 45% of Black defendants — nearly double the rate for white defendants (23%), who were instead more likely to be incorrectly classified as low risk (Angwin et al., 2016). The algorithm was not only ethically unjust but also epistemically flawed and technically ineffective, exhibiting low predictive accuracy. In another example, Amazon ran an AI-based recruitment trial in 2014 to streamline hiring processes. By 2017, it was discontinued after the algorithm was found to systematically disadvantage women (Dastin, 2018).

In these cases, it becomes particularly evident how Artificial Intelligence both integrates and reproduces existing social dynamics. In all three of the previously mentioned cases, the discriminatory outcomes were not simply incidental but rather emerged directly from the type of AI technology employed: supervised machine learning models trained on real-world social data. In supervised machine learning, neural networks are trained on large datasets and learn to recognize patterns by adjusting internal parameters to minimize prediction errors across labelled examples. Once these statistical regularities are identified, the model uses them to make new decisions and perform future tasks. However, in this workflow, the model can encode and amplify existing patterns of bias in the training data, also because it identifies correlations without understanding the underlying causes or broader social contexts (Barocas [Hardt/Narayanan] 2023). Moreover, this process is often opaque, even to the developers themselves, making it difficult to identify or mitigate discriminatory outcomes.

In the case of facial recognition, for instance, training datasets were predominantly composed of lighter-skinned individuals (79.6% in IJB-A and 86.2% in Adience), leading to biased performance (Buolamwini and Gebru, 2018). In the COMPAS and Amazon cases, the training data itself was historically biased, rooted respectively in a long legacy of discrimination against

African American communities in the justice system and gender disparities in the tech labor market.

A similar dynamic can also emerge from the choice and selection of proxy variables. For instance, Obermeyer et al. (2019) found that a widely used algorithm in the U.S. healthcare system predicted patients' health needs based on past health expenditure. Because less money is typically spent on Black patients with the same medical needs as white patients, the software consistently underestimated the needs of Black individuals. As a result, the number of Black patients eligible for extra care was reduced by over half compared to those who would have actually needed it (Obermeyer et al. 2019; Benjamin 2019).

Some biases may emerge from the very structure of the algorithm, independent of the data it processes: for instance, in collaborative filtering systems like those used by Netflix or Amazon, specific algorithmic design choices can lead to biased recommendation patterns. Statistical biases — such as the *cold-start* problem, *popularity bias*, and *homogenization bias* — can directly produce discriminatory outcomes by reinforcing existing stereotypes or systematically favoring content associated with majority groups. This is also evident in the case of search engines, where such biases can result in the preferential ranking of commercial entities or businesses from dominant groups, with tangible economic consequences. (Stinson 2022). Importantly, this mechanism operates even at the foundational level of algorithmic design. The development and architecture of any algorithm — like that of any technology — inevitably reflect human decisions, which are inherently situated and partial (Thaler & Sunstein, 2009). A striking example emerges when AI is applied to the legal domain. Here, attempts to formalize law into machine-readable code require interpretative decisions that become hardcoded into the system, effectively excluding alternative legal interpretations. As Surden (2017) argues, law is inherently interpretive, and the process of translating legal norms into code is never neutral. On the contrary, it necessitates selecting specific legal approaches, assumptions, and values. Even seemingly technical design choices — such as making court documents publicly searchable — reflect value-laden trade-offs, often balancing transparency against privacy. The integration of worldviews, goals, and power relations into technological design is also evident in the field of robotics. When we design a robot, we inevitably make decisions about aesthetics, form, functionality, and even moral behavior. These decisions are never entirely neutral or purely technical. Even when we consider only the aesthetic dimension, the dominant prototype in robotics tends to reflect a male-coded body, while alternative designs are often hyper-sexualized representations of

female forms. This becomes even more evident in the case of sex robots, where the function, features, and appearance are simultaneously technical and ethical, shaped by deeply rooted sociocultural influences (van Grunsven/van Wynsberghe 2019, van Grunsven [Stone/Marin] 2024).

Algorithms and robotic systems clearly show how technology incorporates existing social patterns, inequalities, and normative choices at multiple levels, in various forms, and across a wide range of domains: from the tech industry to the legal system, from healthcare to entertainment. The problem is even greater considering that these socially shaped systems are developed under a narrative of neutrality, which conceals their value-laden nature.

The Implementation of Moral Values into Technological Design

Returning, then, to the question of whether technology can embody or integrate moral values, an affirmative answer is logically entailed. As we have seen in the previous sections, technology not only can integrate moral values, it also cannot not do so. In other words, it is inherently imbued with values by its very nature. If the technoscientific process is intrinsically value-laden, and if technological outcomes are not predetermined — as extensively argued by anti-deterministic and social constructivist approaches — then this opens up a significant space for human agency in shaping technological development. It becomes not only possible, but also unavoidable and necessary to make deliberate decisions about which values should be embedded within technological systems.

We thus arrive at the sub-question B: how can such values be consciously implemented or operationalized in technological design and development? To this end, philosophical and ethical reflection assumes a central and guiding role.

Value-Sensitive Design, Care Ethics and User-Centered Design Approach

One of the earliest approaches to advocate for the possibility of actively integrating specific values into technological design and engineering is Value-Sensitive Design (VSD), defined as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman and Kahn 2003, van Wynsberghe 2012). Friedman and Kahn criticize approaches that treat ethics as a marginal or after-the-fact concern in technological develop-

ment as insufficient. Instead, they advocate for the systematic integration of fundamental human values — such as human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, identity, and environmental sustainability — from the early stages of the design process. To this end, their VSD methodology proposes an iterative and integrative process that integrates conceptual, empirical and technical investigations. The *conceptual* component uses philosophical analysis to clarify how values are supported or diminished by technological designs, who is affected, and how trade-offs among competing values should be navigated in the design. The *empirical* investigation draws on social science to understand how people experience and interpret those values. The *technical* part assesses existing or potential technological solutions in terms of their capacity to uphold or undermine those values (Friedman and Kahn, 2003).

A particularly relevant variant of Value-Sensitive Design (VSD) can be found in the field of robotics, especially in the context of robots designed for the care of vulnerable populations: the Care-Centered Value-Sensitive Design (CCVSD) (Van Wynsberghe, 2012). Drawing on the philosophical framework of care ethics, van Wynsberghe applies the values identified by care ethicist Joan Tronto — attentiveness, responsibility, competence, and reciprocity (Tronto, 1993) — to the domain of elderly care robotics. This approach explores how these values can be operationalized within different robotic designs. In particular, the focus is on making each issue more specific and less abstract, taking into consideration various factors such as context (place), the actors involved, the care practices (e.g., lifting, bathing, feeding, delivery of goods, social interaction), the type of robot (assistive, enabling, replacement), and the manifestation of moral elements. A paradigmatic example of CCVSD is the practice of lifting elderly patient with low mobility, traditionally treated as a purely technical or engineering task. Yet, as van Wynsberghe illustrates, lifting is ethically charged: it occurs at a moment of maximal patient vulnerability and can result in feelings of objectification. This practice comprises value-laden and sensitive actions like eye contact that are integral to establishing and maintaining a bond of trust and preserving the patient's dignity (Van Wynsberghe, 2012). It is therefore possible to consider this ethically relevant practice right from the initial stages of design and functionality choices. When lifting is entirely delegated to autonomous robots (e.g., RIBA), these relational cues are often lost. In contrast, one can choose to favor the creation and use of enabling technologies, such as exoskeletons, which allow the caregiver to pre-

serve the ethical sensitivity of the practice (ibd.). However, if moral evaluation is context-specific, dependent on the individual characteristics of a patient and shaped by cultural variation, pluralistic values, and situational nuance, then a central question emerges: who decides which values are embedded in the technology? We thus arrive at the sub-question C, namely who is entitled to assess the ethicality of a technology.

A deeply rooted view in Western philosophy holds that the final say in ethical decisions belongs to the citizens directly affected (Dewey 1927, Arendt 1958, Rawls 1993). Only those who live the problems have the experiential standpoint needed to judge them, and only their participation can make science and technology genuinely democratic.

This is where methodological frameworks such as the User-Centered Design (UCD) become crucial. UCD encompasses approaches such as *participatory design* and *co-design*, which aim to include stakeholders, and especially end-users, throughout the design process (Sanders & Stappers, 2008). In this view, the active engagement of users is not merely a desirable feature but a necessary condition for ensuring that the technology remains truly human-centered (Gould/Lewis, 1985). By involving users through iterative cycles of co-construction, these approaches help to align technical functionalities with the lived values, concerns, and actual needs of end users. Many studies in healthcare show that integrating patient feedback is key to improving services (De Rosis [Ferrè/Pennucci] 2022).

In this way, the dimension of public trust and consumer adoption can be directly connected to that of moral reasoning. On one hand, the social acceptance of a technology is a pivotal for its success and long-term adoption. On the other, listening to end-users' voices serves to operationalize a range of ethical frameworks: from consequentialist ethics, where moral evaluation is based on promoting the greatest well-being for the greatest number (Neri 2020), to the relational and non-hierarchical approach of care ethics (Gilligan 1977, Botti 2015), to the mutual and dynamic processes addressed in pragmatist ethics (Keulartz et al., 2004), and even to non-paternalistic, self-defined forms of deontological reasoning. Importantly, this approach can also inform legal and regulatory frameworks. By grounding regulation in bottom-up evaluations and user participation, it can offer concrete guidance for policies that reflect real-world contexts and ethical concerns (WHO, 2007).

Operationalizing Ethics in the Case of Elderly Care Robots

After examining how technology in general can embody or integrate moral values, with a focus on various cases of AI and robotics, we have outlined several approaches for implementing such values in technological design and development and identified who should have the final say in the ethical assessment. With this foundation, we can now return to the specific case of robots for elderly care. There remains, in fact, the final and central sub-question D: which specific values are at stake in the case of Elder Care Robots (ECRs), and how can they be assessed and implemented in practice? To this end, the discussion will be divided into three parts. First, we will define what social robots and elderly companionship robots are, providing a general overview along with a review of their documented effects. Second, we will address the question: what does it mean for ECRs to be "ethical"? Finally, we will explore how such ethical considerations can be implemented in legal and regulatory frameworks.

Social Robotics and Robots to Combat Loneliness

Robots are currently among the most widely discussed and anticipated technological innovations, both in Western and Eastern contexts. In public discourse and the collective imagination, they are often seen as symbols of progress and future potential (Alnajjar et al. 2021). Originally designed to carry out physical tasks, robots are now increasingly being developed to engage in social interaction. One of the most sensitive and pressing social needs today is companionship, particularly for vulnerable groups such as the elderly. In response, growing attention is being directed toward the development of "companion robots", leading to the emergence of a dedicated and rapidly evolving field: social robotics. Social robots are "a new class of machines designed to function as 'social partners' for humans" (Damiano/Dumouchel, 2020). They are used in a wide variety of contexts: wellbeing, education, socialization, healthcare, disability assistance, elderly care, motivation and behavioral influence, rehabilitation, entertainment, navigation and guidance, shopping, and general assistance (Ahmed [Buruk/Hamari], 2024). They have proven particularly effective in contexts where human companionship is unavailable or limited (e.g. during pandemics) or in contexts where individuals are socially isolated because of physical or mental conditions (ibd.). The central strategy of social robotics is to create "socially interactive robots" (Fong [Nourbakhsh/Dautenhahn] 2003), meaning robots capable of generating a

believable social presence. This is defined as the robot's ability to give users the "sense of being with another" (Biocca et al., 2003) or the "feeling of being in the company of another" (Heerink et al., 2008). Achieving this involves transferring and adapting features of human face-to-face social interaction to human-robot interaction. One of the most crucial features is the ability to communicate through emotion (Damiano/Dumouchel, 2020). To this end, social robots are often equipped with advanced capabilities that enable them to foster deep emotional connection and facilitate social engagement, while also assisting with practical tasks (Ahmed [Buruk/Hamari], 2024). These characteristics do not necessarily need to be human-like. According to Fong et al. (2003), social robots can be categorized based on their appearance and features into four main types: anthropomorphic (human-like), zoomorphic (animal-like), functional (machine-like), and caricatured (object-like) (Fong et al. 2003, Ahmed [Buruk/Hamari] 2024). What makes a robot socially effective is not its resemblance to a human, but rather its capacity to trigger certain innate human responses. As Sherry Turkle explains, robots can activate our "Darwinian buttons", for example, by making eye contact, which causes people to respond as if they were engaging in a real social relationship (Turkle 2007). Designers often include features like large eyes and rounded heads to resemble infants. According to Konrad Lorenz, founder of ethology, such "baby-like" traits instinctively elicit tenderness and care from humans (Lorenz, 1970. Kaplan 2001). Some well-known social robots include Nao, Pepper, Buddy, Jibo, and Amazon's Astro: Nao is a small humanoid robot used mainly in research, education, and therapy, capable of speech, movement, and emotion recognition; Pepper, also from SoftBank, is designed to detect emotions and interact socially, and is widely adopted in research and pilot programs. Buddy is a friendly home companion with an expressive screen face, helping with reminders, entertainment, and social connection; Jibo and Astro, both aimed at private home use, combine features like voice interaction, mobility, and social presence — though with mixed commercial success. These robots can foster social presence, empathy, and contextual awareness. They often respond to voice, recognize faces, interpret emotions, and support users in both practical and emotional ways (Ahmed [Buruk/Hamari] 2024).

The field of social robotics for the elderly seeks to address the challenges posed by the aging population, shortages in care services, and the increasing burden on caregivers, as ensuring a high quality of life for older adults has become an increasingly important priority (WHO 2024). In this context, robots have been developed to assist with a wide range of tasks: physical support

(e.g., lifting, mobility aid), housework, guided physical exercise, cognitive assistance (e.g., reminders), and telemedicine (Hoppe et al 2020, Ahmed et al 2024). However, one of the most widespread and often overlooked issues in elderly care is social isolation, which is strongly linked to poor mental health outcomes such as depression, anxiety, and cognitive decline (Jané-Llopis/Gabilondo 2008, Yen et al. 2024). This is where Elderly Companionship Robots (ECRs) come into play. These social robots aim to alleviate loneliness by providing interactive, emotionally supportive presence. A particularly researched subgroup within elderly care is individuals living with dementia, for whom a wide variety of robotic devices have been developed. Examples include Paro, a baby seal-like robot covered in soft fur that reacts to touch and sound; Hyodol, a doll-like companion robot widely used in South Korea, even with cognitively healthy older adults; and animal-like robots such as the robotic dog Aibo and the Joy-for-All cat. Studies have shown positive outcomes, including reduced loneliness, improved communication, and emotional engagement (Tamura et al. 2004, Sharkey/Sharkey 2012, Robinson et al. 2013, Lee et al. 2023, Yen et al 2024). Similarly, studies on Aibo report comparable emotional benefits to those seen with real animals (Collins [Millings/Prescott] 2013). The same robots are also used in therapy with cognitively healthy older adults, where they serve as tools for stimulation, interaction, and emotional support. These include humanoid robots such as Nao and its therapeutic version Zora, Pepper, Aibo, and many others. Research suggests that these robots can support wellbeing, stimulate conversation, and help reduce feelings of isolation and anxiety, at least in the short term (Skarkey/Sharkey 2012, Anderson/Felzmann, 2024). For instance, a randomised controlled trial in a New Zealand care facility found that interactions with Paro significantly reduced loneliness compared to usual activities or even a resident dog. Residents engaged in more physical interactions and discussion, and Paro also appeared to act as a social catalyst within the group (Robinson et al. 2013).

However, evidence is mixed. While several narrative and qualitative reviews highlight clear psychosocial benefits, especially in terms of engagement, emotional expression, and reduced use of medication, some meta-analyses have found no statistically significant effects on outcomes such as quality of life or depression (Pu et al. 2019, Lu et al. 2021, Anderson/Felzmann, 2024). This may also be due to the inherent difficulty in detecting such effects through standardised measures, as highlighted by several narrative and qualitative analyses that instead report perceived benefits (Pu et al., 2019). Moreover, the biggest challenge regards long-term outcomes: for example, the study by Lee

et al. (2023) on Hyodol showed reduced depressive symptoms at three months, but no lasting effect at six, and some Japanese reports describe a rapid decline in users' interest over time (Sharkey/Sharkey, 2010).

How Can Companion Robots be Ethical?

Despite potential benefits, social robots have raised significant concerns among scholars, especially for what concerns the formation of affective bonds in human-robot relationship. What happens when our parents or grandparents become emotionally attached to robots that, at least *prima facie*, cannot genuinely reciprocate their feelings? Do the abovementioned features of social robotics (such as human-like appearance and behavior, the ability to evoke tenderness, personalized memory, and kind, believable speech) not only enhance user engagement but also involve a form of deception? Is it ethical to design companion robots with the specific aim of eliciting affective attachment, especially in vulnerable users like older adults?

The Accusation of Deception

When discussing ethics in the context of social robots, particularly companion robots, the specific moral values at stake are typically those related to human dignity, affective reciprocity, and the risk of emotional deception or manipulation. These moral values lie at the heart of a complex and ongoing ethical debate, with many scholars arguing that the use of social robots in vulnerable contexts, such as elderly care, is problematic. Among the first and most influential critiques is that of Sparrow and Sparrow, who argue:

> "Any beneficial effects of robot pets or companions are a consequence of deceiving the elderly person into thinking that the robot is something with which they could have a relationship. […] We believe that it is not only misguided, but actually unethical, to attempt to substitute robot simulacra for genuine social interaction." (Sparrow/Sparrow 2006)

Their position reflects a broader deontological perspective, concerned with authenticity, truthfulness, and respect for human dignity. Similar arguments have been developed by many others. For instance, Sherry Turkle, based on extensive empirical and theoretical work, including case studies of older adults interacting with social robots, has raised critical concerns. She asks:

> "The fact that our parents, grandparents and our children might say 'I love you' to a robot who will say 'I love you' in return, does not feel completely comfortable; it raises questions about the kind of authenticity we require of our technology." (Turkle 2006)

Van Wynsberghe also takes a critical stance toward this technology, focusing on issues of relational authenticity and vulnerability in care contexts:

> "Through an analysis of the value of reciprocity from care ethics, […] it becomes clear that HRI cannot achieve this bidirectional value of reciprocity; a robot must deceive users into believing it is capable of reciprocating to humans or is deserving of reciprocation from humans. […] social robots designed for reciprocity use reciprocity as an instrumental value to enhance acceptability of the robot." (van Wynsberghe 2022)

From these perspectives, emotional authenticity is considered central. The idea is that genuine emotions are understood as first-person, internal experiences; since social robots do not possess such inner emotional states and cannot genuinely feel or reciprocate emotions, their simulations of social or emotional behavior are inherently deceptive. Therefore, designing robots to evoke emotional attachment is, from this standpoint, ethically wrong. But is this truly the only, or the best, way to frame the issue? According to some scholars, the answer may be more nuanced.

Another Perspective: Emotions as Coordination and the Relational View of the Self

In recent years, new perspectives in the philosophy of mind and emotions, and more generally in understandings of the human mind and self, have begun to emerge as alternatives to the classical, standard views. According to Dumouchel and Damiano (2017), for example, who have developed a new theory of emotions applied to Human-Robot Interaction called the "Theory of Affective Coordination", the common argument of deception is based on a fundamental misunderstanding. They argue that, while modern philosophy of mind formally rejects Cartesian dualism and acknowledges the mind as a cognitive system like any other, it nevertheless tends to preserve the view of the human mind as a paradigmatic epistemic agent—almost ontologically distinct. The deception argument relies precisely on this residual dualism: it assumes that external expressions are only genuine if they mirror internal, private states.

Thus, when robots simulate emotions, they are accused of misleading users into believing that those emotions are authentic (Dumouchel/Damiano 2017). This view, however, is regarded as philosophically untenable. Even Descartes had to postulate an external agent — the evil genius — to demonstrate solipsism, since without an external point of reference, the very notion of being wrong loses coherence: one cannot recognize an error without some form of comparison or feedback. This suggests that truth and meaning are not purely internal, but necessarily arise within relational and intersubjective contexts (ibid.). Affectivity, too, is embedded in intersubjective interaction: agents adjust their behavior in response to others, whose reactions shape our strategies, expectations, and relationships. Dumouchel and Damiano thus propose that the mind, as well as emotions and affective processes, exists within a relational space. Emotions, in Dumouchel's view, are not private mental states or internal attributes of the individual, but relational properties emerging from, and functional to, a mechanism of strategic interindividual coordination (Dumouchel 1995, Dumouchel/Damiano 2017).

Interestingly, this reconceptualization of emotions aligns with emerging views of the self, mind, and affectivity that challenge traditional notions of the mind as internal, private, rational, and representational. Rooted in American pragmatism, such perspectives go back to philosophers like William James, who famously inverted the intuitive model of emotion. According to James, emotions are not internal states that precede bodily expression, but rather the bodily expression constitutes and influences the emotional experience itself (Murphy, 1997). Numerous psychological experiments have attempted to demonstrate how bodily or environmental cues influence emotional states (Laird/Lacasse 2014): eye contact can increase feelings of affection (ibd.); the classic Strack, Martin, and Stepper (1988) study showed that holding a pen in one's mouth to simulate smiling could influence emotional evaluation; the "suspension bridge" experiment on misattributed arousal showed that participants were more likely to feel attracted under conditions of physiological stress (Dutton/Aron 1989). Social influences also play a role: for example, simply being told that your heart is beating differently may change your emotional perception (Valins 1966, Taylor 1975). This is consistent with research in behavioral sciences that challenge the idea of the mind as a rational, introspective unit. Asch's conformity experiment (1955) and moral psychology studies suggest that moral judgments are often intuitive rather than deliberative. According to Bem (1972), introspection is often a post-hoc rationalization of behavior, not an access to internal causes. From pragmatist foundations,

radical embodied and enactive approaches to cognition have emerged. Radical embodied cognitive science argues that cognition, perception, and emotion are not internal processes but are grounded in bodily and environmental interaction (Chemero 2013). Similarly, enactive accounts claim that cognition is not individual but arises from dynamic interaction between agents (De Jaegher & Di Paolo 2007). This shift gives new importance to the body and environment in shaping mental life. In psychology, sensorimotor psychotherapy emphasizes how working on the body rather than rational cognition can be effective in trauma treatment (Ogden & Minton, 2000). In robotics, an embodied approach has shown promise: morphology can reduce computational demands and enhance task efficiency (Hoffmann & Pfeifer, 2012); neuromorphic and active perception research suggests that vision and understanding arise through movement and sensorimotor interaction rather than internal mental representation (D'Angelo et al., 2025). At a deeper level, these shifts invite us to reconceptualize the very notion of the self. As seen above, our scientific models are deeply shaped by cultural, social, and moral values. The individualist, rationalist image of the self reflects Western socio-economic and cultural paradigms, particularly those of capitalism and liberalism. By contrast, African approaches to the selfhood emphasize its shared and emergent nature through communal relationships (Jecker/Atuire, 2025). Feminist and care ethics similarly criticize the notion of the atomized individual and promote ideas such as relational autonomy and interdependence (Barclay, 2000).

If we then connect these views to proposals such as Bruno Latour's Actor-Network Theory (ANT), which holds that both human agents and technological artefacts contribute symmetrically to hybrid networks of relations, we arrive at a radically relational and distributed perspective (Latour 2005). Within this framework, robots may be seen as active participants in affective and social interactions, not because they possess personal or inner consciousness, but because subjectivity itself is co-constructed within a relational network. If there is no longer a Cartesian dichotomy, and affectivity is instead distributed across a relational network of various actants, then the affective dimension is no longer confined to human beings. This means that the assumptions underlying the deception argument lose their force: if, in order to avoid deception, it was once required that external emotional expressions match internal emotional states, and that requirement no longer holds, then human–robot interaction is no longer inherently deceptive. As a consequence, social robots can be understood as social partners, and the human–robot relationship can be

framed as a new form of interaction within the human affective ecosystem, rather than as the more controversial replacement for human relationships.

New Ways to Determine Deception

Social robots, therefore, are not deceptive *per se*. But this does not mean they are necessarily non-deceptive either: they can be. Hence, caution is needed, especially because we are dealing with vulnerable users, whose protection is of primary importance. Therefore, we must carefully assess under what specific conditions deception might occur. Looking more closely at the Theory of Affective Coordination, another consequence emerges: what is morally relevant is not the "truth" of another's emotion, understood as correspondence with an internal, private mental state, but the mutual commitment to act within a reciprocal emotional coordination. This standpoint aligns with the pragmatist account of truth, according to which truth is not about an unattainable dimension of objectivity, but about its function in action. (Murphy 1987) As Charles Sanders Peirce put it: "Reality […] consists in the particular sensible effects that things participating in it produce"; or in William James' words: "Ideas […] become true only insofar as they help us to establish a satisfactory relationship with all the other parts of our experience." (ibd.) Bringing these perspectives together, we can redefine deception: an agent deceives us not when its actions fail to correspond to some internal "true" emotional state, but when it betrays our expectations of mutual behavior. Consequently, determining when social robots may be deceptive cannot rely solely on abstract theorizing; it requires bottom-up, context-specific, culturally sensitive, and user-centered research. To this end, future empirical work is needed to explore older adults' emotional expectations toward social robots. While some studies have addressed this question in general terms (Hoppe et al. 2020), more targeted research focusing specifically on affective expectations is still lacking.

This reflects a first, foundational layer of the concept of deception. But several scholars have proposed more nuanced distinctions to better guide both ethical reflection and user protection, particularly in the case of vulnerable populations (Danaher 2020, Umbrello/Natale 2024). Most notably, two categories often emerge: the first is what we have discussed so far – a form of structural or design-based deception, resulting from the robot's social cues or the user's psychological tendencies (such as agency attribution or personification). The second, more serious, form is what Danaher (2020) calls "hidden state deception" and Umbrello and Natale (2024) refer to as "strong deception". This stronger form of deception refers to situations in which a robot deliber-

ately conceals or misdirects attention away from certain capacities or functions, e.g. covertly recording users through a hidden camera (Danaher, 2020), or when users are led to misunderstand the artificial nature of the system (Umbrello/Natale 2024). In these cases, "the deception serves some purpose other than truth, one that is typically to the advantage of the deceiver and the disadvantage of the deceived" (Danaher, 2020). According to Umbrello and Natale, strong deception involves three features: (1) lack of transparency, (2) an intent to mislead, and (3) the potential to undermine user autonomy or control. While the first type of deception (structural or affective) can be meaningfully examined within a philosophical framework, as discussed above, the second type (intentional and harmful deception) may be more appropriately addressed through legal and regulatory instruments.

Ethics by Law in Social Robotics: Preliminary Reflections on Article 5 of the AI Act

Digital technologies evolve at a pace that often exceeds the capacity of legal and regulatory systems to respond effectively. Regulatory frameworks tend to operate on significantly slower structural timelines, leaving many emerging technologies unregulated or only partially addressed. The European Union is one of the few institutional bodies that has formalized ethical principles and translated them into regulatory frameworks for AI technologies, albeit with significant challenges such as the abstract nature of many guidelines and their limited binding power.

The most comprehensive effort in this direction is the AI Act, which was proposed in 2021 and officially entered into force on August 1st 2024 (European Commission, 2024). Some legal scholars working on social robotics have pointed out that the issue of potentially deceptive social robots may fall under Article 5 of the AI Act, which concerns prohibited AI practices (Bertolini, 2024). The article reads as follows:

1. The following AI practices shall be prohibited:
    a. the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to

make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm; Related: Recital 29

b. the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm; Related: Recital 29. (European Commission, 2024)

The most relevant part of this provision for the present analysis is the explicit prohibition of deceptive technologies, under which we can include including robotic systems, that employ "manipulative or deceptive techniques, with the objective or the effect of materially distorting the behavior of a person". This provision may represent a striking legal translation of the ethical concerns previously discussed, especially those addressed in the philosophical literature on deception, emotion simulation, and relational ethics, since a misplaced emotional attachment could lead to unintended or harmful behavioral changes. As such, philosophical and ethical conclusions remain highly relevant to legal interpretation and implementation. If we accept the ethical arguments explored earlier — particularly those informed by relational theories and the Theory of Affective Coordination — then social robots created to generate feelings of affection in the user should not be considered inherently deceptive, and therefore not inherently subject to the bans outlined in Article 5.Companion robots for the elderly and other vulnerable population, therefore, can be considered generally permissible and not automatically excluded by the regulation.

However, the fact that not all social robots fall under the scope of Article 5 does not imply that none of them do. Following the ethical analysis previously outlined, the lighter and structural form of "design-based deception" must be distinguished from more problematic forms of deception that involve manipulation or exploitation. Given the vulnerability and delicate moral status of the end-users, it is crucial to safeguard their human dignity, ensuring that both no unintended harmful effects occur and no explicit intention to deceive is present. Within the regulatory framework of the AI Act, this highlights the need for a more nuanced evaluation and the development of more specific

guidelines, aimed at more narrowly and precisely identifying which types of companion robots are ethically and legally permissible, and which may raise deeper concerns.

To this end, much work remains to be done. However, even from the embryonic analysis provided here, it is possible to outline some very preliminary guidelines. On the one hand, this concerns the methodological approach. As shown by the value-sensitive design perspective, the integration of moral values can—and should—occur at the very beginning of the design process. The form, appearance, and functionality of a robot can take many different shapes and not necessarily follow the depictions found in dominant cultural narratives. Furthermore, as highlighted by the need for user-centered design approaches, the final say in this technological construction must lie with the end-users themselves; in the case of ECRs, with elderly people. As Sharkey & Sharkey emphasize, "it is important to ensure that robots introduced into elder care do actually benefit the elderly themselves, and are not just designed to reduce the care burden on the rest of society." (Sharkey/Sharkey 2012). It is therefore essential that any design and development processes be informed by, and co-constructed with, elderly users. On the other hand, it is necessary to employ theoretical tools from ethics and philosophy alongside those of the law, using legal constraints to operationalize ethical considerations in ways that maximize the protection of vulnerable populations. For instance, legal constraints can be applied to manufacturers regarding product features, or to distributors and retailers concerning marketing practices—thus minimizing risks related to profit-driven exploitation. This could include, for example, applying specific regulations used for medical devices. By combining the top-down theoretical insights of ethics and philosophy with the bottom-up, user-centered and value-sensitive design approaches, it becomes possible to begin outlining context-sensitive guidelines tailored to different types of deception and specific use cases.

Preliminary Guidelines

Building on the previous analysis, I propose some preliminary regulatory and design guidelines, which need to be further developed and specified in future work. These recommendations follow the above-mentioned distinction between banal and strong deception, to which I suggest adding a third area of concern: commercial exploitation, a particular form of strong deception. Indeed, the main risk posed by social robots is that they may be designed to

exploit users' psychological vulnerabilities for commercial gain, much like social media platforms have proven to do — enabled by privatization, deregulation, monopolies, and disproportionate power distribution. For example, they could manipulate users' attention and emotions to increase engagement or use personalized interactions to influence behavior in ways that primarily serve profit-driven goals rather than users' own interests. This includes tactics such as reinforcing addictive behaviors, subtle persuasion without clear consent, and controversial personalization up to undisclosed advertising.

In light of this, and following the threefold definition of deception, the following guidelines can be outlined, addressed to different stakeholders:

- Avoiding Banal Deception
  - Ensure emotional safety: conduct field studies using user-centered and value-sensitive design focused on users' affective expectations and inform regulation on these findings. (researchers, companies, national and supranational institutions…)
  - Promote understanding of human-robot interaction as a novel relational form and not as a replacement for human-human relations, favoring non-human-like designs in elderly care and discouraging the employment of hyper-realistic anthropomorphic robots. (companies, manufacturers, developers…)
  - In this perspective, design and usage guidelines should ensure that robots are not used exclusively nor perceived as replacements. (lawyers, manufacturers, final users…)
- Avoiding Strong Deception
  - Ensure transparency and privacy protection: all features must be clearly declared, with clear communication to guarantee informed consent as much as possible. (lawyers, companies, retailers and distributors, final users…)
  - Prohibit intentional manipulation (e.g., hidden cameras, covert recording, advertising purposes, and other deceptive practices). (lawyers, institutions, manifacturers…)
- Avoiding Commercial Exploitation
  - Explore reclassification of certain social robots under stricter regulatory categories, such as medical devices, to provide stronger control and protection for vulnerable populations. (lawyers, researchers, institutions…)

- Promote transparency about commercial intents and practices behind social robots. (lawyers, companies, retailers and distributors…)
- Encourage business models through legal boundaries that prioritize user well-being over profit maximization. (lawyers, researchers, institutions…)

Conclusion

Despite the complexity and abstract nature of regulations such as the AI Act, it is nonetheless possible and necessary to specify clear frameworks to guide the ethical governance of technologies. This ensures the protection of end users without resorting to blanket bans on potentially beneficial innovations, such as elderly care and companionship robots. Indeed, legal instruments are crucial to operationalize ethics and embed values within technology, acknowledging that technology is never neutral but always reflects social biases, values, and worldviews. At the same time, technology is neither inevitable nor predetermined. On the contrary, it can be shaped towards multiple possible futures and diverse forms by integrating specific values and purposes into its design and deployment.

In the specific case of Elderly Care Robots (ECRs), particular attention must be paid to the risks of deception and lack of reciprocity, with a focus on human dignity and well-being. Several paths can be pursued to achieve this goal: on the one hand, developing ethical evaluations grounded in rigorous philosophical and theoretical work that incorporate emerging relational perspectives; and on the other, undertaking bottom-up, field-based analyses centered on the interests, care, and lived experiences of end users. The overarching purpose is to translate ethical and empirical insights into robust regulatory measures that effectively guide innovation, while acknowledging the complexity of this task and that substantial work remains to be done.

References

Ahmed, E., Buruk, O. O., & Hamari, J. (2024). Human–robot companionship: Current trends and future agenda. *International Journal of Social Robotics*, *16*(8), 1809–1860.

Akrich, M. (1992). The de-scription of technical objects. In: W.E. Bijker and J. Law, eds. Shaping technology/building society. Cambridge, MA: MIT Press, pp.205–224.

Al-Halawani, R., Charlton, P. H., Qassem, M., & Kyriacou, P. A. (2023). A review of the effect of skin pigmentation on pulse oximeter accuracy. *Physiological measurement*, *44*(5), 05TR01.

Allen, C., Wallach, W., Smit, I. (2006). Why machine ethics? IEEE Intelligent Systems, 21(4), pp.12–17.

Alnajjar, F., Bartneck, C., Baxter, P., Belpaeme, T., Cappuccio, M., Di Dio, C., Eyssel, F., Handke, J., Mubin, O., Obaid, M. and Reich-Stiebert, N. (2021). *Robots in education: An introduction to high-tech social agents, intelligent tutors, and curricular tools*. Routledge.

Anderson, M. L., & Felzmann, H. (2024). 3. Ethical complexities within the appearance and usage of social robots: A scoping review. Irish Journal of Applied Social Studies, 24(1), 3.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 May.

Arendt, H. (2022). The human condition. University of Chicago press.

Asch, S. E. (1955). Opinions and social pressure. Scientific American, 193(5), 31–35.

Barclay, L. (2000). Autonomy and the social self. Relational autonomy, 52–71.

Barocas, S., Hardt, M., Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bartoletti, I. (2020). An artificial revolution: on power, politics and AI. Black Spot Books.

Benjamin, R. (2019). Assessing risk, automating racism. Science, 366(6464), pp.421–422.

Bem, D. J. (1972). Self-perception theory. In: L. Berkowitz, ed. Advances in experimental social psychology, Vol. 6. New York: Academic Press.

Bertolini, A. (2024). The subtle line between personalization and user manipulation in a European regulatory perspective. A proposal for a technology-assessment methodology for Artificial Intelligence Systems. In *2024*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4777–4784). IEEE.

Bijker, W.E. (1995). Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change. Cambridge, MA: MIT Press.

Bijker, W.E. (2009). Social construction of technology. In: J.K.B. Olsen, S.A. Pedersen and V.F. Hendricks, eds. A companion to the philosophy of technology. Oxford: Wiley-Blackwell, pp.88–94.

Bijker, W.E., Hughes, T.P., Pinch, T., eds. (1987). The social construction of technological systems. Cambridge, MA: MIT Press.

Botti, C. (2015). Prospettive femministe nel dibattito bioetico contemporaneo. In *Donne, diritto, diritti. Prospettive del giusfemminismo* (pp. 97–115). Giappichelli.

Bucchi, M. (2002). Scienza e società: introduzione alla sociologia della scienza. Bologna: Il Mulino.

Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR, pp.77–91.

Carrada, G. (2005). Comunicare la scienza: kit di sopravvivenza per ricercatori. Vol. 12. Milano: Alpha Test.

Chemero, A. (2013). Radical embodied cognitive science. Review of General Psychology, 17(2), 145–150.

Coalition for Future Mobility (n.d.) *Benefits of self-driving vehicles*. Available at: https://coalitionforfuturemobility.com/benefits-of-self-driving-vehicles/ (Accessed: 23 June 2025).

Collins, E.C., Millings, A., Prescott, T.J. (2013). Attachment to assistive technology: A new conceptualisation. In: Assistive technology: From research to practice. IOS Press, pp.823–828.

Criado-Perez, C., 2019. Invisible women. New York: Abrams Press.

D'Angelo, G. V., Clerico, V., Bartolozzi, C., Hoffmann, M., Furlong, P. M., Hadjiivanov, A. (2025). Wandering around: A bioinspired approach to visual attention through object motion sensitivity. Neuromorphic Computing and Engineering.

Damiano, L., Dumouchel, P.G. (2020). Emotions in relation. Epistemological and ethical scaffolding for mixed human-robot social ecologies. HUMANA.MENTE Journal of Philosophical Studies, 13(37), pp.181–206.

Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. Ethics and Information Technology, 22(2), pp.117–128.

Dastin, J. (2018). 'Amazon scraps secret AI recruiting tool that showed bias against women', *Reuters*, 10 October. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (Accessed: 23 June 2025).

De Jaegher, H., Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. Phenomenology and the cognitive sciences, 6, 485–507.

De Rosis, S., Ferrè, F., Pennucci, F. (2022). Including patient-reported measures in performance evaluation systems: patient contribution in assessing and improving the healthcare systems. *The International Journal of Health Planning and Management*, *37*, 144–165.

Dewey, J. (1927). The Public and Its Problems, Henry Holt.

Dumouchel, P. (1995). *Émotions: essai sur le corps et le social*. Synthélabo.

Dumouchel, P., Damiano, L. (2017). Living with robots. Cambridge, MA: Harvard University Press.

Dutton, D. G., Aron, A. (1989). Romantic attraction and generalized liking for others who are sources of conflict-based arousal. Canadian Journal of Behavioural Science, 21(3), pp.246–255.

European Commission (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689).

Fong, T., Nourbakhsh, I., Dautenhahn, K. (2003). A survey of socially interactive robots. Robot Auton Syst 42:143–166. https://doi.org/ 10.1016/S0921-8890(02)00372-X.

Friedman, B., Kahn, P.H. Jr. (2003). Human values, ethics, and design. In: J.A. Jacko and A. Sears, eds. The human–computer interaction handbook. Mahwah, NJ: Lawrence Erlbaum Associates, pp.1177–1201.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard educational review*, *47*(4), 481–517.

Gould, J. D., Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM*, *28*(3), 300–311.

Grasswick, H. (2018). *Feminist Social Epistemology*. In: Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition). Available at: https://plato.stanford.edu/archives/fall2018/entries/feminist-social-epistemology/ (Accessed: 24 June 2025).

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108(4), pp.814–834.

Harding, S. G. (1986). The science question in feminism. Ithaca: Cornell University Press.

Hoffmann, M., Pfeifer, R. (2012). The implications of embodiment for behavior and cognition: animal and robotic case studies. arXiv preprint. arXiv:1202.0440.

Hoppe, J. A., Johansson-Pajala, R. M., Gustafsson, C., Melkas, H., Tuisku, O., Pekkarinen, S., Hennala, L., Thommes, K. (2020). Assistive robots in care: Expectations and perceptions of older people. In: Aging between participation and simulation: Ethical dimensions of socially assistive technologies in elderly care, pp.139–156.

Huyskes, D. (2025). Constructing automated societies: Socio-cultural determinants and impacts of automated decision-making in public services, Università degli Studi di Milano.

Institute for Health Metrics and Evaluation (IHME) (2019). Global Health Data Exchange (GHDx). [online] Available at: https://vizhub.healthdata.org/gbd-results/.

Jecker, N. S., Atuire, C. A. (2025). Authors meet critics: What is a person? Untapped insights from Africa. Journal of medical ethics, 51(4), 249–250.

Jané-Llopis, E., Gabilondo, A., eds. (2008). Mental health in older people: Consensus paper. Luxembourg: European Communities.

Kaplan, F. (2001), January. Artificial attachment: Will a robot ever pass Ainsworth's strange situation test. In *Proceedings of humanoids* (pp. 125–132).

Keulartz, J., Schermer, M., Korthals, M., Swierstra, T. (2004). Ethics in technological culture: A programmatic proposal for a pragmatist approach. Science, Technology & Human Values, 29(1), pp.3–29.

Kittay, E. (1999). Love's labor: Essays on women, equality, and dependency. New York: Routledge.

Laird, J. D., Lacasse, K. (2014). Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion. Emotion Review, 6(1), pp.27–34.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge, MA: Harvard University Press.

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University press.

Lee, O. E., Nam, I., Chon, Y., Park, A., Choi, N. (2023). Socially assistive humanoid robots: Effects on depression and health-related quality of life among low-income, socially isolated older adults in South Korea. *Journal of Applied Gerontology*, 42(3), 367–375.

Lorenz, K (1970). Studies in animal and human be haviour. Harvard University Press.

Lu, L. C., Lan, S. H., Hsieh, Y. P., Lin, L. Y., Lan, S. J., Chen, J. C. (2021). Effectiveness of companion robot care for dementia: a systematic review and meta-analysis. *Innovation in aging*, *5*(2), igab013.

Murphy, J. P. (1997). Il pragmatismo. Bologna: Il Mulino.

Neri, D. (2020). *Filosofia morale. Manuale introduttivo*. goWare & Guerini editore.

Obermeyer, Z. et al., 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), pp.447–453.

OECD (2020), *Who Cares? Attracting and Retaining Care Workers for the Elderly*, OECD Health Policy Studies, OECD Publishing, Paris, https://doi.org/10.1787/92c0ef68-en.

Ogden, P., Minton, K. (2000). Sensorimotor psychotherapy: One method for processing traumatic memory. Traumatology, 6(3), 149–173.

Ogburn, W.F., Thomas, D. (1922). Are inventions inevitable? A note on social evolution. Political Science Quarterly, 37(1), pp.83–98.

Oudshoorn, N. (2003). *The male pill: A biography of a technology in the making*. Duke University Press.

Parks, J. A. (2010). Lifting the burden of women's care work: Should robots replace the "human touch"? Hypatia, 25(1), pp.100–120.

Pu, L., Moyle, W., Jones, C., Todorovic, M. (2019). The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. *The gerontologist*, *59*(1), e37-e51.

Rawls. J. (1993), Political Liberalism, Columbia University Press.

Robinson, H., Macdonald, B., Kerse, N., Broadbent, E. (2013). [Various contributions]. Journal of the American Medical Directors Association, 14. DOI: 10.1016/j.jamda.2013.02.007.

Sanders, E. B. N., Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, *4*(1), 5–18.

Sharkey, A., Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. Ethics and Information Technology, 14, pp.27–40.

Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering. AI and Ethics, 2(4), pp.763–770.

Sparrow, R., Sparrow, L. (2006). In the hands of machines? The future of aged care. Minds and Machines, 16, pp.141–161

Strack, F., Martin, L. L., Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. Journal of Personality and Social Psychology, 54(5), pp.768–777.

Surden, H. (2017). Values embedded in legal artificial intelligence. University of Colorado Law Legal Studies Research Paper, No. 17–17. SSRN.

Tamura, T., Yonemitsu, S., Itoh, A., Oikawa, D., Kawakami, A., Higashi, Y., Nakajima, K. (2004). Is an entertainment robot useful in the care of elderly people with severe dementia? *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *59*(1), M83-M85.

Taylor, S. E. (1975). Relationship between physiological feedback and the attribution of emotion. Journal of Personality and Social Psychology, 31(2), pp.306–314.

Tengai (n.d.) Our story. Available at: https://tengai.io/our-story/ (Accessed: 23 June 2025).

Thaler, R. H., Sunstein, C.R. (2009). Nudge. London: Penguin.

Tripaldi, L. (2023). Gender tech: Come la tecnologia controlla il corpo delle donne. Roma: Gius. Laterza & Figli.

Tronto, J. (1993). Moral boundaries: A political argument for an ethic of care. New York: Routledge.

Turkle, S., Taggart, W., Kidd, C. D., Dasté, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. Connection Science, 18(4), pp.347–361.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction studies*, *8*(3), 501–517.

Umbrello, S., Natale, S., 2024. Reframing deception for human-centered AI. *International Journal of Social Robotics*, *16*(11), 2223–2241.

United Nations (2015). Transforming our world: The 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September &2015, A/RES/70/1, pp.1–13.

Valins, S. (1966). Cognitive effects of false heart-rate feedback. Journal of Personality and Social Psychology, 4(4), pp.400–408.

van Grunsven, J., van Wynsberghe, A. (2019). A semblance of aliveness: How the peculiar embodiment of sex robots will matter. *Techné: Research in Philosophy and Technology*.

van Grunsven, J. (2022). Anticipating sex robots: A critique of the sociotechnical vanguard vision of sex robots as 'good companions'. In: I. van de Poel and S. Royakkers, eds. Being and value in technology. Cham: Springer, pp.63–91.

van Grunsven, J., Stone, T., Marin, L. (2024). Fostering responsible anticipation in engineering ethics education: how a multi-disciplinary enrichment of the responsible innovation framework can help. *European journal of engineering education*, *49*(2), 283–298.

van Wynsberghe, A. (2012). Designing robots for care: Care centered value-sensitive design. Science and Engineering Ethics, 19(2), pp.407–433.

van Wynsberghe, A. (2022). Social robots and the risks to reciprocity. AI & Society, 37(2), pp.479–485.

Winner, L., (1980). 'Do Artifacts Have Politics?', *Daedalus*, 109(1), pp. 121–136.

World Health Organization (WHO) (1946). Constitution of the World Health Organization. [online] Available at: https://www.who.int/about/governance/constitution

World Health Organization (WHO) (2007). People centred health: A framework for policy. [online] Available at: http://www.wpro.who.int/NR/rdonlyres/

World Health Organization (WHO) (2023). Mental health of older adults. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/mental-health-of-older-adults

World Health Organization (WHO) (2024). Ageing and health. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

Yen, H. Y., Huang, C. W., Chiu, H. L., Jin, G. (2024). The effect of social robots on depression and loneliness for older residents in long-term care facilities: A meta-analysis of randomized controlled trials. *Journal of the American Medical Directors Association*, *25*(6).

Practioner's Corner

Navigating Challenges and Opportunities of Standards Developing Organizations in Times of Rapid Technological Advancement

A DIN e.V. Perspective

Dr. Maria Mensch, Adrian Seeliger, Johannes Wellhöfer

Abstract: *The standardization of AI ethics is an essential yet complex endeavor, necessitating a balance between technological progress, regulatory frameworks, and societal values. As AI systems increasingly shape economic, political, and social realities, ensuring their ethical use requires robust and legitimate standardization mechanisms. DIN e.V. as a national Standards Developing Organization (SDO) acknowledges the challenges of defining a universal ethical framework, given the vast differences in values, legal systems, and governance structures across the globe. In this publication we will address the involvement of for-profit companies in the standardization process, the distribution and visibility of stakeholders and their interests, and whether ethics can be formalized and how the certification of AI could proceed. The certification of ethics in AI with regard to defined specifications on principles of fairness, transparency, accountability and privacy, shows trustworthy compliance through a trusted structural process. The specifications used as a basis for certification need to have a solid understanding of their implications. This requires the participation of all stakeholders. Critics point out that traditional standardization bodies often struggle to achieve a balance between open democratic participation and efficiency, which leads to the criticism that certification mechanisms are either too bureaucratic or not inclusive enough (Werle and Iversen, 2006). Another point of criticism is the perceived increasing influence of private companies and industry-led bodies on the standardization process behind closed doors, leading to concerns about industry dominance and regulatory capture (Corporate Europe Observatory, 2025). Addressing these points, we show which mechanisms we have within the SDOs, which institutions (e.g. Consumer Council and Commission for Occupational Health and Safety and Standardization (KAN)) exist and how they*

contribute to a formalization of ethical principles in order to enable a fair and inclusive standardization process. Using case studies, we show that standardization is a means that can serve as a mechanism to align technological advancements with societal values. In conclusion, the certification of AI ethics and the governance of AI standardization at first glance appears fraught with challenges, from industry influence to geopolitical disparities. However, history demonstrates that standardization can serve as an important moderating system for harmonization, ensuring responsible innovation while safeguarding fundamental rights. SDOs work as a mediator between many different stakeholder groups, advocating for a balanced approach that fosters global cooperation while upholding ethical principles in AI development and deployment.

Keywords: *Standardization; Standards Developing Organization; DIN; Artificial Intelligence*

Introduction

Artificial intelligence (AI) influences various facets of human lives ranging from economic to social domains like chatbots or automated creditworthiness assessments (Genovesi et al., 2024). Given the range of AI through its integration into everyday processes, the demand for regulation to ensure ethical use is on the rise as shown by the European Artificial Intelligence Act. Robust ethical standards within evolving AI technology should encompass key principles such as fairness, transparency, accountability, and privacy (Hermann, 2022; Hunt and Vitell, 1986). Official international standards guide the development and deployment of AI systems and must therefore ensure they themselves are aligned with societal values and do not inadvertently compromise human rights.

Standards Developing Organizations (SDOs), such as DIN e.V. on a national level, or ISO and IEC as international institutions, play a critical role in this endeavor. These organizations are tasked with orchestrating the development of standards that harmonize technological advancements with ethical imperatives. SDOs face the challenge of establishing frameworks that are not only technologically sound, but also socially responsible, contributing actively to global efforts in setting legitimate standardization mechanisms (Hoek, 2015; Van Eck and Waltman, 2010). Through these frameworks, SDOs aim to shape the future landscape of AI development, advocating for standards that also reflect diverse ethical considerations and promote responsible innovation.

Ethics in AI and Business

When talking about ethics, common understanding is the rational and systematic examination of the norm of what is right and wrong, and moral actions whose results do not cause harm to people. Another aspect is the ability to distinguish between good and bad actions/guidelines and standards that prescribe the decision to act in different circumstances (Graham, 2013; Kazim, 2017; Weiss, 2014). It should be noted that morality in philosophical ethics compares inner states that are directed towards external constraints, the so-called virtue. Meanwhile, AI ethics can be oriented towards the legislature to cover the key issues of governance, accountability and transparency. Here, it is important that there are rules and guidelines that are covered and implemented by the executive branch of the government, a so-called coercive core (Kazim and Koshiyama, 2021).

The most relevant ethical theories for business ethics and AI ethics are virtue, consequentialism, and deontology ethics (Kapteint and Wempe, 2002). Virtue ethics has its roots in pre-Enlightenment Aristotelianism and states that a person's character traits influence their behavior in certain situations. Character traits or virtues such as honesty, self-control, integrity, courage, generosity and fairness are desirable. The person should develop into a perfect self, the action itself or its consequences are not the focus of virtue ethics (Kapteint and Wempe, 2002; Kazim and Koshiyama, 2021).

In contrast, consequentialism theory has the consequence of an action at its center. The maximization of pleasure or minimization of displeasure is sought as a consequence through the formulation of principles. Here, the outcome literally justifies the means, and the character traits of a person can be disregarded. This ethical approach is often found in government policy and decision-making, for example the maximization of gross domestic products as an argument in economic policy (Kazim and Koshiyama, 2021; Okello, 2021). The third deontology ethics theory, which is also known as duty ethics, sees the duty and obligation to commit an action as the main essence of its theory. This distinguishes it from the consequences and outcomes of consequentialism ethics, where the weakness lies in the inflexibility and the use of strict rules (Kapteint and Wempe, 2002; Okello, 2021). When ethics is applied practically, it is often referred to as morality, as it is more concrete in its application and more helpful to society. Morals and ethics are related, but not the same, which often leads to misunderstandings and discussions in literature. Morality is based on realistic standards of behavior and norms of action and is therefore

easier for businesses to grasp (Place, 2019; Yeh et al., 2020; Vadera & Pathki, 2021;).

Business ethics emphasizes organizational responsibility to society over organizational profit (Kolb, 2008). This responsibility to society has been discussed in literature since the 1930s and 40s, with the question of what specific social responsibility companies have (Caroll, 1999). With the current environment and the increased interest in ethics, businesses have no choice regarding their moral responsibility and how they behave, as misconduct can lead to profit and image loss (Brimmer 2007; Okello, 2021; Kamila and Jasrotia, 2023). Environmental, social and governance issues (ESG factors) have received a lot of attention in recent years and can influence a business positively or negatively. Without ethical practices, the business opens itself up to a high risk of being boycotted by consumers, fines, and penalties or litigation claims, whereas the positive effects of ethical behavior in an organization include increased profitability and strengthened internal and external relationships with stakeholders (Nelson, Campfield and Weeks, 2008; Jin, Drozdenko and DeLoughy, 2013; Okello, 2021). The lack of ethical behavior and guidelines can lead to very diverse business risks such as data leaks from technology companies, the dissemination of false information on platforms or the fast fashion industry, with its poor working conditions in production countries and harmful environmental impact (Binet et al., 2019; Jing and Murugesan, 2019; Marsden, Meyer and Brown, 2020; Niinimäki et al., 2020).

There are two ethically relevant aspects in AI development, on the one hand AI itself, how it works and "acts", and on the other hand the development of AI, under what working conditions the developers work, whether the programming work has been outsourced to countries with lower wages and possibly inhumane working conditions for example. Both aspects offer approaches for standardization in order to take ethical aspects into account in both the development and application of AI, and to give industry and users guidelines and technical basis for their behavior and processes. In the discussion of AI ethics, however, the focus is primarily on the processes and development of AI, consumer safety in its application and the ethical concepts that AI developers follow and that are used in programming.

Challenges in Standardizing AI Ethics

AI ethics is concerned with the psychological, social and political implications of AI. This includes e.g. protection against undue manipulation, the right to know when you are not interacting with a human and questions of mental autonomy in the psychological sphere. The issues of justice and fairness fall under the social aspect of above mentioned ESG factors and influence on economic and democratic processes falls under the political area (Floridi, 2014; Kriebitz and Lütge, 2020; Müller, 2020).

The process of standardizing AI ethics is complex, primarily due to the diverse cultural and legal landscapes across the globe reflecting a profoundly dynamic and novel technology with diverse societal interaction. Different regions uphold varying ethical values, legal systems and regulatory approaches making it challenging to formulate a universal framework that encompasses these differences (Hoek, 2015; ÓhÉigeartaigh et al. 2020).

Secondarily, conflicting objectives and the unclear state of technology regarding metrics for AI monitoring present further barriers to standardization. Also, the diversity in use cases necessitates tailored ethical considerations, making a one-size-fits-all approach impractical at this stage (Genovesi et al., 2024; Radanliev et al., 2024).

The literature highlights the need for dynamic and adaptive standards that can navigate these intricacies, allowing for the global dissemination of ethical AI practices without sacrificing cultural specificity (Ho, Wang and Vitell, 2012; Vitolla, Raimo and Rubino, 2021).

Furthermore, some SDOs, such as DIN e. V. have internal guidelines and documents that prohibit the standardization of ethics as such. Since ethics is a term that can be interpreted in different ways, ethical aspects and guidelines should be defined and anchored in national and international standards. Here the task of the SDOs is not to define ethical principles but to process the legislative framework and support regulation through technical definitions, requirements or guidance documents. This puts the SDOs and thus the respective topical stakeholders in a position to help shape AI regulation within the legislator's intended boundaries. These standards act as principles in the development of AI products and can integrate ethical considerations. To ensure trust of other market actors in one's own adherence to standards and specifications, actors can certify their AI products. This means having a third party audit the implementation of standards and specifications and document this audit to allow for higher confidence in these AI products. (Radanliev et al., 2024).

Another significant challenge lies in incorporating adequate stakeholder participation. A large amount and high variety of participants boosts inclusiveness, while it reduces speed and flexibility creating a conflict of interest. The main objective of the international standardization system and its involved national SDOs is to dissolve this conflict of interest and to create a diverse and high participation.

An example of the demand for mechanisms of participation and moderation is evident in the numerous definitions of artificial intelligence systems. As AI is subject to potential regulation or oversight, having a clear and distinct definition is crucial to understanding its scope. However, various definitions of AI exist. For instance, the EU AI Act and the broadly accepted OECD definition differ, and the foundational document ISO/IEC 22989:2022 offers yet another interpretation of 'AI system' (Holistic AI, 2024). This diversity in definitions likely stems from the involvement of many stakeholders, resulting in consensus based on the minimal common denominator for each individual working group.

While democratic inclusiveness is vital in capturing a wide array of perspectives, there is often criticism that standardization mechanisms either become too bureaucratic or not sufficiently inclusive (Corporate Europe Observatory, 2025; Raj, Jasrotia, Rai and Ansari, 2022). Additionally, the influence of private companies and industry-led bodies can lead to concerns regarding regulatory capture and industry dominance in setting ethical guidelines. This dominance can skew ethical standards towards commercial interests, potentially perceived as undermining broader societal values (Schuklenk, 2020).

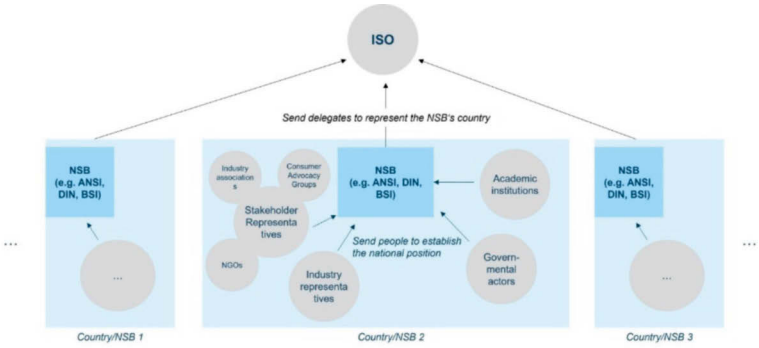
While standardization historically focused only on specific technologies, standardization today is becoming increasingly interconnected and complex. Almost all surrounding aspects of a given technology, e.g. logistics, process management, safety and environmental impact, are now also standardized. Additionally, international standards are becoming more politicized because of their wide influence. While governments have historically stayed out of the standardization process, a change is noticeable here. As such, the influence of the Western actors is often pitted against the growing influence of China within the world of international standardization. A metric frequently used to monitor country involvement in international standardization is the number of proposals for new working items or the chair of working groups by country or their representative national standardization body (NSB) providing the administrative work of a given committee. The various incentives for govern-

ments to participate in standardization, whether through tax breaks or by paying experts for their contributions, illustrate the harsher tone (Barrett, 2024).

Despite the increased political interest, SDOs, such as ISO and IEC, have mechanisms in place to avoid over-politicization and to ensure their functionality and legitimacy. These mechanisms are based on the two core principles of consensus and voluntarism (Barrett, 2024). These two principles state that only a maximum of a quarter of negative votes and a positive two-thirds majority are required for votes. These guidelines presuppose the willingness of the various parties to discuss and compromise in order to reach an agreement and virtually prevent a single party from doing it alone. The agreed standards are not binding but are to be applied voluntarily. As a result, application through market actors ultimately decides whether the standard is fit for purpose, unless governments intervene and make the standard binding by law. A special category are Harmonized European Standards, whose development is commissioned by the European Commission and complying with them demonstrates compliance with requirements of the relevant EU legislation. However, the industry is still free to choose a different solution to meet the requirements (Barret, 2024, Your Europe). The ethical challenges in the field of AI require proactive strategies from governing bodies, including mechanisms that promote equal stakeholder representation, mitigate the risk of industry dominance, while fostering constructive dialog between all parties involved (Rothschild, 1999; Quazi & O'Brien, 2000; Brenkert, 2002). To accomplish this task, SDOs rely on incorporating stakeholder with different backgrounds' perspectives for a holistic approach and inclusive standardization processes (Agudelo, Jóhannsdóttir and Davídsdóttir, 2019). As the means of different parties vary greatly, one of the simplest first steps is to facilitate access to meetings, e.g. by setting up web meetings or hybrid solutions.

Another problem is the influence of interest groups from industry. Silicon Valley giants have historically used their considerable resources to shape academic research and ethical guidelines to align with their commercial goals (Ochigame, 2019; Schuklenk, 2020) This practice risks unduly influencing regulatory institutions that are supposed to oversee AI development, thereby distorting AI standards in favor of commercial interests and potentially at the expense of the common good.

Figure 1: Structure of opinion formation in ISO as an exemplary Standards Development Organization (SDO). The National Standardization Bodies (NSB) represented by NSB 1, NSB 2 and NSB 3 get input from different sources, such as stakeholder representatives, industry representatives, academic institutions and governmental actors, forming a national opinion. This opinion is then represented by a delegate sent by the NSB to the SDO, here ISO.



To prevent this, transparency in the standardization process is essential. Ensuring that dialogues and negotiations between stakeholders are open and documented will help to maintain accountability and prevent undue influence by any single party. The distribution and visibility of the stakeholders involved, such as in the foreword of DIN SPECs developed by DIN e. V., can make a contribution here. Conventional standardization mechanisms are often criticized for being either too bureaucratic or not inclusive enough, which limits the participation of critical voices such as smaller technical innovators or consumer groups. Current strategies to improve stakeholder engagement include public consultations, open forums and the use of digital platforms to broaden participation. By increasing stakeholder visibility and accountability, SDOs are better able to develop standards that align technological advances with societal values.

Formalizing Ethics and Certification Mechanisms in AI

In the previous sections, various ethical principles and theories were discussed. Nowadays, ethics no longer only refers to interpersonal interactions, but also includes a technical component, on the one hand the exchange be-

tween humans and technology and, on the other hand, what influence new technologies have on life. As AI-driven processes become more widespread and integrated into our everyday lives, the formalization of ethics in AI systems is a crucial process that underpins the integrity and social acceptance of artificial intelligence. AI technologies permeate different sectors. Establishing robust certification mechanisms enables compliance with ethical principles such as fairness, transparency, accountability and privacy. These principles are crucial to protect (individual) liberties and enable regulatory mechanisms.

In order to maintain an ethical AI system, the expectations of the system's outcome as well as the expectations of how the system has arrived at these outcomes must be considered with ethical principles (DIN and DKE, 2022). It is also important that the algorithms in question are accountable throughout the application's lifecycle, i.e. they operate transparently, fairly and in line with societal expectations (Radanliev et al., 2024).

The implementation of these principles should be considered from the outset. Here developers in particular are called upon, to implement the various demands and requirements in consultation with different interest groups (ethicists, civilians, tech companies, lawyers). If you take into consideration the whole life cycle, this includes not only testing AI, but also continuous monitoring in the form of security protocols, error monitoring and reliability. There are various tests for this, such as permutation importance, drop-column importance, model-specific methods, and post-hoc interpretation methods each with their own strength and pitfalls (Radanliev et al., 2024). Another aspect is certification and regular audits, which can strengthen the trust of the civilian population in the applications. In the area of cybersecurity, there is the so-called Red-teaming, in which the system is attacked and thus the resistance is tested; this process would also be conceivable in the AI area (Radanliev et al., 2024).

Another aspect is data protection, which is a fundamental aspect of the development of AI systems, if sensitive data are part of the training or deployment of a system. and could include issues such as prejudice based on socioeconomic class, gender, skin color and other factors. Companies should and often are obliged to provide resources to address these. The inclusion of various stakeholders in development, as mentioned above, can help with issues such as prejudice by representing broader societal values and offering a range of perspectives. Feedback options from users and other stakeholders after the implementation of the system help to identify problems and adapt to changing circumstances (Radanliev et al., 2024). Transparency reports containing infor-

mation on new AI implementations, problems and solutions are another way to increase trust in AI applications.

While the implementation and realization of ethical principles is done by developers and companies, the formalization process stems from a different level of philosophical and societal discourse. From the perspective of technical implementation in democratic contexts these are foremost perspectives moderated by election, law making, laws and law enforcement as well as further discussed stakeholder requirements. Here, criteria that AI systems must meet in order to be considered ethical are established, comparable to the development of marketing ethics, where the need for reliable data is emphasized to improve ethical standards and practices (Hunt and Vitell, 1986). Similarly, AI ethics should be based on empirical rigor and supported by comprehensive methodologies detailing ethical requirements and compliance. This empirical input is needed to lower the challenge to translate ethical principles into quantifiable standards that can be universally applied. The concept of "ethics by design" discussed in the AI ethics literature advocates incorporating ethical considerations at the earliest stages of AI development (Schuklenk, 2020). This proactive approach emphasizes the importance of embedding ethical considerations into the technological design process to ensure that AI systems inherently uphold societal values. Care and consideration should be exercised to maximize the benefits to society while limiting potential harm. Establishing standards and certification procedures based on these standards, to ensure that AI is developed and applied in an ethical manner, is a crucial element in putting Ethics by Design into practice. Industry-relevant guidelines covering various aspects of AI, e.g. data management, transparency of algorithms, explainability and fairness, can be created through collaboration with international SDOs such as ISO. The development of structures, guidelines and principles should be based on a collaborative approach involving multiple stakeholders, such as governments, industry, academia and civil society. The inclusion of different perspectives is essential to develop ethical standards that are both relevant and representative, as noted in previous discussions on marketing responsibility (Kamila and Jasrotia, 2023). Another point is the promotion of global cooperation, which due to the global nature of AI technologies, standards and legal frameworks should transcend national borders (Radanliev et al., 2024).

The standards developed can be verified through innovative certification models, which helps to streamline the process, reduce bureaucratic hurdles and improve inclusivity at the same time. For example, privacy-friendly tech-

nologies such as federated learning are proposed as models that embed ethical standards directly into the development process and guarantee the confidentiality of data. Through certification, independent third parties can objectively verify whether an organization's AI system installations are ethical and in line with best practices. By formalizing ethical values represented in state law and establishing the basis for effective certification mechanisms, SDOs can promote AI systems that are ethically compliant and socially accepted. Achieving this goal not only strengthens the credibility of AI systems but also advances the mission of aligning technological progress with ethical demands on a global scale.

In parallel with industry advancements, the European Union has adopted a widely discussed regulation for AI: the EU AI Act. Based inter alia on different publications by the High-Level Expert Group on AI set up by the European Commission, it features a human- and risk-centered approach. The most critical AI applications will be categorized in the high-risk class, depending on the areas of use, leading to a vast number of obligations (Regulation (EU) 2024/1689, Busch et al., 2024). Harmonized European standards, a tool well known in the EU product safety legislation, are currently developed to implement and further formalize the EU AI legislation.

Compliance with these standards enables providers of AI systems to issue a written EU declaration of conformity stating that the system meets all relevant requirements thus reducing the involvement of third parties in most cases. Organizations will still be able to comply with the EU's obligations without using harmonized standards using a formal third-party assessment by a notified body.

Successful Co-Creation of AI Standards: Insights from Practice

To ground our discussion in current efforts, exploring examples and case studies where standardization processes have succeeded can provide insights into the practicalities of aligning AI technological advancements with societal values.

The following two examples originate from the healthcare and the financial sector. As both are highly regulated areas already due to their substantial impact on our life, they were early in developing and adopting AI governance instruments such as standards.

One significant example comes from the healthcare sector, where AI-driven diagnostic tools have revolutionized patient care. By adhering to strict ethical guidelines and certification processes, these tools have managed to enhance diagnostic accuracy while respecting patient privacy and ensuring data security (Hermann, 2022). The implementation of AI ethics standards has involved multi-stakeholder collaboration, including medical professionals, patient advocacy groups, and technology developers. This diverse engagement has helped balance innovation with ethical principles, promoting trust in AI applications.

Another area of interest is the financial industry, where AI algorithms are used for credit scoring and fraud detection. Ethical standardization has been proven as crucial in preventing discriminatory practices and ensuring accountability in decision-making processes (Ho, Wang, and Vitell, 2012; Vitolla, Raimo and Rubino., 2021). The standardization of AI within the financial sector brings several benefits. By establishing publicly available frameworks, it enhances transparency and allows customers and supervisors to better understand the mechanisms in use. For financial institutions, standardized guidelines reduce the burden of developing solutions from scratch, enabling them to rely on shared resources and established best practices. Furthermore, standards created through structured processes often integrate ethical principles, promoting fairness and alignment with societal values. However, given instances of unethical behavior in the finance industry, continuous oversight is necessary to ensure these standards are effectively implemented and adhered to. Anticipating the need for adequate AI standardization, the German Standardization Roadmap on Artificial Intelligence (DIN and DKE, 2022) was launched as an open stakeholder-involvement effort to identify needs for standardization with active participation of over 500 experts. The area of the financial industry was a main topic of the second edition of the document, leading to the development of three specifications projects. The DIN SPEC 91512 "Fairness of AI-based applications in the financial sector", DIN SPEC 91527 "Goals, Methods and Metrics for Automated/Semi-Automated Runtime Monitoring of AI Systems for Non-Adversarial Performance" and DIN SPEC 92001–3 "Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 3: Explainability" can serve as the basis and starting point for consensus-based international standardization. Industry, researchers, regulatory and governmental bodies and users have collaborated in these projects, laying important groundwork. Especially the standardization initiative DIN SPEC 92001–3 serves as a practical example of how formalization of ethical

aspects can be supported by standardization processes (DIN, 2023; DIN, 2024; DIN 2025).

A key topic, horizontally affecting all AI applications, is transparency. Being at the intersection of technical and ethics and technology it plays a key role in most frameworks or regulations trying to manage trust and accountability in autonomous systems. Transparency in AI refers to the extent to which the operations, decision-making processes, and impacts of AI systems can be understood and traced. Various organizations, both public and private, have attempted to define transparency, leading to a highly fragmented research landscape (Haresamudram, Larsson, and Heintz, 2023). A technical specification effort involving a diverse range of contributors (DIN SPEC 91528, currently being developed) seeks to lay the groundwork for standardizing the terms and definitions and overview of relations and dependencies of technical transparency dimensions that are widely recognized. In addition, the involvement of various stakeholders can help to align the currently divergent concepts of regulatory frameworks, research, legal norms and societal expectations, thereby strengthening trust in AI systems and mitigating risks while promoting regulatory compliance and societal acceptance. Thus, the creation of technical standards aims to ease discussion of these issues through staying at the descriptive level. The prescriptive level of normative social rules has to be dealt with on each individual application's level

Mechanisms for Ensuring Inclusive and Fair Standards

The pursuit of ethical AI standardization demands mechanisms that ensure inclusivity and fairness in stakeholder participation. Without such measures, the process risks sidelining important voices and values, leading to standards that may not fully address the needs of diverse communities.

One critical mechanism to realize the above goals is the implementation of public consultation processes, allowing a wide spectrum of stakeholders, including consumer groups, academia, and industry representatives, to contribute to standard development. Such consultations can help democratize the standardization process, ensuring that various perspectives are considered and valued (Schuklenk, 2020; Raj, Jasrotia, Rai, and Ansari, 2022). Within the EN/ISO set of rules for creating standards documents, organizations are invited to participate, and participating organizations are able to comment and vote.

In addition to consultations, the formation of dedicated councils and commissions can support inclusive participation. Institutions, such as, for example, the Consumer Council and Commission for Occupational Health and Safety and Standardization (KAN) at DIN e.V., play vital roles in advocating for professional end users and private individuals' interests, and ensuring that ethical standards reflect societal values (Kamila and Jasrotia, 2023). These bodies can facilitate ongoing dialogue and monitor the standardization process to safeguard against dominance by any one group.

Furthermore, cross-sector and cross-stakeholder collaborations can enhance the inclusivity of standardization efforts. By fostering partnerships between various sectors – e.g. IT, manufacturing, quality assurance – and various stakeholders – public, private, and non-profit – SDOs can encourage a holistic approach that integrates contributions from all relevant parties. These collaborations are instrumental in creating standards that not only address technological advancements, but also ensure with ethical principles on a global scale, as shown for example in the above-mentioned (DIN, 2024 *DIN SPEC 91512*), where industry and researchers collaborated with the banking supervision.

Through these mechanisms, SDOs can champion fair and inclusive standardization processes, promoting ethical AI development that upholds societal values and protects fundamental rights.

Conclusion

In this article, we have looked at various ethical theories and discussed the different requirements and challenges of ethics in AI.The standardization of AI ethics is complex due to the many areas of society that are now influenced by AI, as well as the global nature of AI and the different interpretations of ethics and their application. In addition, SDOs are confronted with accusations of industry influence and opaque processes, the lack of balance between efficiency and inclusivity and are simultaneously exposed to a changing political dimension of standardization. However, both history and recent developments have shown, standardization capacities for all stakeholders are of fundamental importance. They act as essential mechanisms for synchronization, ensuring coordinated development of standards alongside technological innovation and are a tool for protecting basic human rights, by observing conventions and rights.

To enable the standardization of ethical AI, a multi-layered approach that involves all relevant stakeholders, employs different methodologies and promotes global collaboration is necessary. SDOs play a crucial role in this process as they provide the framework for the necessary processes. By promoting transparent stakeholder engagement, lowering organizational barriers and advocating for inclusive frameworks, SDOs can drive ethical AI practices. In the example of DIN e.V., as a national SDO, examples on the development of standards that can help formalize ethical aspects and particularize legislation were shown. With internal, independent institutions such as the Consumer Council and the Commission for Occupational Health and Safety and Standardization (KAN), DIN e.V. shows a way to involve and protect stakeholders, such as end users, who are primarily affected by AI applications.

In summary, there are many challenges in the standardization of ethical AI and some changes can be expected in the future. Common technical rules will be set, if this will happen by a coordinated approach through the processes of official international standardization (e.g. ISO or IEC), by de facto means of big actors within the AI ecosystem, or otherwise has to be seen. However, standardization and the SDOs have proven in the past their ability to unite different perspectives, build consensus and ensure responsible progress. It can therefore be assumed that, in the case of AI ethics, standardization is also a valuable instrument that continues these principles and leads to a result that forms the basis for common technical implementation and certification and thus provides industry and users with a tool to facilitate understanding of the processes and trust in AI applications.

References

Agudelo, L. C., Jóhannsdóttir, L., and Davídsdóttir, B. (2019). A literature review of the history and evolution of corporate social responsibility. International Journal of Corporate Social Responsibility, 4(1), 2–23.

Athwal, N., Wells, V. K., Carrigan, M., and Henninger, C. E. (2019). Sustainable luxury marketing: A synthesis and research agenda. International Journal of Management Reviews, 21(4), 405e426. https://doi.org/10.1111/ijmr.12195

Barrett, T. (2024). "Standards Development Organisations in an era of strategic competition," United States Studies Centre at the University of Sydney.

Binet, F., Coste-Manière, I., Decombes, C., Grasselli, Y., Ouedermi, D., and Ramchandani, M. (2019). Fast fashion and sustainable consumption. Fast

fashion, fashion brands and sustainable consumption, 19e35. https://doi.org/10.1007/978-981-13-1268-7_2

Brenkert, G. G. (2002). Ethical Challenges of Social Marketing. Journal of Public Policy & Marketing, 21(1), 14–25.

Brimmer, S. (2007). The Role of Ethics in 21st Century Organizations.

Busch, F., Kather, J.N., Johner, C. et al. (2024). Navigating the European Union Artificial Intelligence Act for Healthcare. npj Digit. Med. **7**, 210. https://doi.org/10.1038/s41746-024-01213-6

Carroll, A. B. (1999). Corporate social responsibility. Business & Society, 38(3), 268–295.

Corporate Europe Observatory. (2025). Bias Baked In: How Big Tech Sets Its Own AI Standards.

Daza MT and Ilozumba UJ (2022) A survey of AI ethics in business literature: Maps and trends between 2000 and 2021. Front. Psychol. 13:1042661. doi: 10.3389/fpsyg.2022.1042661

DIN, DKE (2022): German Standardization Roadmap on Artificial Intelligence (2nd edition); https://www.din.de/go/roadmap-ai

DIN e.V. (2024). DIN SPEC 91512 Fairness of AI-based applications in the financial sector. https://dx.doi.org/10.31030/3573086

DIN e.V. (2025) DIN SPEC 91527 Goals, Methods and Metrics for Automated/Semi-Automated Runtime Monitoring of AI Systems for Non-Adversarial Performance. https://dx.doi.org/10.31030/3649777

DIN e.V. (2023). DIN SPEC 92001–3. Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 3: Explainability. https://dx.doi.org/10.31030/3446955

Floridi, L. (2014). The 4th Revolution (Oxford University Press).

Genovesi, S., Mönig, J.M., Schmitz, A., Poretschkin, M., Akila, M., Kahdan, M., Kleiner, R., Krieger, L. and Zimmermann, A. (2024) Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans. AI Ethics 4, 537–553. https://doi.org/10.1007/s43681-023-00291-8

Graham, J., (2013). The Role of Corporate Culture in Business Ethics. s.l., s.n.

Goworek, H. (2011). Social and environmental sustainability in the clothing industry: A case study of a fair trade retailer. Social Responsibility Journal, 7(1), 74e86. https://doi.org/10.1108/17471111111114558

Haresamudram, K., Larsson, and S., Heintz, F. (2023). Three Levels of AI Transparency. Computer, vol. 56, no. 2, pp. 93–100, doi: 10.1109/MC.2022.3213181

Hermann, E. (2022). Leveraging artificial intelligence in marketing for social good: An ethical perspective. Journal of Business Ethics, 179(1), 43–61. https://doi.org/10.1007/s10551-021-04843-y

Hoek, J. (2015). Informed choice and the nanny state: Learning from the tobacco industry. Public Health, 129(8), 1038e1045. https://doi.org/10.1016/j.puhe.2015.03.009

Holistic AI (2024). Lost in Transl(A)t(I)on: Differing Definitions of AI [Updated]. Available at: https://www.holisticai.com/blog/ai-definition-comparison. Accessed on 13th March 2025.

Ho, F. N., Wang, H. M. D., and Vitell, S. J. (2012). A global analysis of corporate social performance: The effects of cultural and geographic environments. Journal of Business Ethics, 107(4), 423–433. https://doi.org/10.1007/s10551-011-1047-y

Hunt, S. D., and Vitell, S. (1986). A general theory of marketing ethics. Journal of Macromarketing, 6(1), 5–16. https://doi.org/10.1177/027614678600600103

Jing, T. W., and Murugesan, R. K. (2019). A theoretical framework to build trust and prevent fake news in social media using blockchain. In Recent trends in data science and soft computing: Proceedings of the 3rd international conference of reliable information and communication technology (IRICT 2018) (pp. 955e962). Springer International Publishing. https://doi.org/10.1007/978-3-319-99007-1_88.

Jin, K. G., Drozdenko, R. G. and DeLoughy, S., (2013). The Role of Corporate Value Clusters in Ethics, Social Responsibility, and Performance: A Study of Financial Professionals and Implications for the Financial Meltdown. Journal of Business Ethics, 112(1).

Kamila, M. K. and Jasrotia, S. S., (2023). Ethical issues in the development of artificial intelligence: recognizing the risks, International Journal of Ethics and Systems, Emerald Group Publishing Limited, vol. 41(1), pages 45–63, July.

Kamila, M. K. and Jasrotia, S. S. (2023) Ethics and marketing responsibility: A bibliometric analysis and literature review, Asia Pacific Management Review, Volume 28, Issue 4, Pages 567–583, ISSN 1029–3132, https://doi.org/10.1016/j.apmrv.2023.04.002

Kaptein, M. and Wempe, J., (2002). Three General Theories of Ethics and the Integrative Role of Integrity Theory. SSRN Electronic Journal.

Kazim, E. (2017). Kant on Conscience (Brill).

Kazim, E. and Koshiyama, A. S. (2021) A high-level overview of AI ethics, Patterns, Volume 2, Issue 9, 100314, ISSN 2666–3899, https://doi.org/10.1016/j.patter.2021.100314.

Kolb, R. W., (2008). Encyclopedia of Business Ethics and Society.. s.l.:SAGE Publications, Inc.

Kriebitz, A., and Lütge, C. (2020). Artificial intelligence and human rights: a business ethical assessment. Bus. Hum. Rights J. 5, 84–104.

Marsden, C., Meyer, T., and Brown, I. (2020). Platform values and democratic elections: How can the law regulate digital disinformation? Computer Law & Security Report, 36, Article 105373. https://doi.org/10.1016/j.clsr.2019.105373

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In The Stanford Encyclopedia of Philosophy, E.N. Zalta, ed. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/.

Nelson, W. A., Campfield, J. M. and Weeks, W. B., (2008). The Organizational Costs of Ethical Conflicts. Journal of healthcare management / American College of Healthcare Executives, 53(1), pp. 41–52.

Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T., and Gwilt, A. (2020). The environmental price of fast fashion. Nature Reviews Earth & Environment, 1(4), 189e200. https://doi.org/10.1038/s43017-020-0039-9

Ó hÉigeartaigh, S. S., Whittlestone, J., Liu, Y. et al. Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. Philos. Technol. 33, 571–593 (2020). https://doi.org/10.1007/s13347-020-00402-x

Ochigame, R. (2019). The invention of 'Ethical AI': How big tech manipulates academia to avoid regulations. The Intercept. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

Okello, O. A. (2021). Impact of Culture on Business Ethics: a Literature Review, EU Business School.

Place, K. R. (2019). Moral dilemmas, trials, and gray areas: Exploring on-the-job moral development of public relations professionals. Public Relations Review, 45(1), 24e34. https://doi.org/10.1016/j.pubrev.2018.12.005

Quazi, A. M., and O'brien, D. (2000). An empirical test of a cross-national model of corporate social responsibility. Journal of Business Ethics, 25(1), 33e51. https://doi.org/10.1023/A:1006305111122

Radanliev, P., Santos, O., Brandon-Jones, A. and Joinson, A. (2024) Ethics and responsible AI deployment. Front. Artif. Intell. 7:1377011. doi: 10.3389/frai.2024.1377011

Raj, S., Jasrotia, P., Rai, P., and Ansari, M. (2022). Democracy, Technology and Marketing Responsibility. Journal of Responsibility Studies, 29(2), 123–140.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)." Official Journal of the European Union, L 2024/1689, 12 July 2024, ppRothschild, M. L. (1999). Carrots, sticks, and promises: A conceptual framework for the management of public health and social issue behaviors. Journal of Marketing, 63(4), 24e37. https://doi.org/10.1177/002224299906300404

Schuklenk, U. (2020). On the ethics of AI ethics. Bioethics, 34(2), 146–147.

Vadera, A. K., and Pathki, C. S. (2021). Competition and cheating: Investigating the role of moral awareness, moral identity, and moral elevation. Journal of Organizational Behavior, 42(8), 1060e1081. https://doi.org/10.1002/job.2545

Valenzuela, L. M., Merigó, J. M., Johnston, W. J., Nicolas, C., and Jaramillo, J. F. (2017). Thirty years of the Journal of Business & Industrial Marketing: A bibliometric analysis. Journal of Business & Industrial Marketing. https://doi.org/10.1108/JBIM-04-2016-0079

van Eck, N., and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Technological Forecasting and Social Change, 77(5), 632–634. https://doi.org/10.1016/j.techfore.2010.03.001

Vitolla, F., Raimo, N., and Rubino, M. (2021). The impact of corporate governance on the disclosure of digital information in the context of fair, accountable, and transparent AI. Sustainability, 13(1), 302.

Weiss, J. W., (2014). Business Ethics : A Stakeholder and Issues Management Approach. 6 ed. s.l.:Berrett-Koehler.

Werle, R., and Iversen, E. J. (2006). Promoting Legitimacy in Technical Standardization. Science, Technology & Innovation Studies.

Yeh, C. C., Lin, F.,Wang, T. S., &Wu, C. M. (2020). Does corporate social responsibility affect cost of capital in China? Asia Pacific Management Review, 25(1), 1e12. https://doi.org/10.1016/j.apmrv.2019.04.001

Your Europe (2023) Europäische Normen. Available at: https://europa.eu/youreurope/business/product-requirements/standards/standards-in-europe/index_de.htm