Responsible AI Governance in the Public Sector: Explaining Contextual Dynamics through a Realist Synthesis Review

Gagua, Ana; van der Voort, H.G.; Goyal, N.; Verbraeck, A.

DOI
10.1609/aies.v8i1.36606

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 2025 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Responsible AI Governance in the Public Sector: Explaining Contextual Dynamics through a Realist Synthesis Review

Ana Gagua, Haiko van der Voort, Nihit Goyal, Alexander Verbraeck
Multi-Actor System Department, Delft University of Technology
a.gagua@tudelft.nl, h.g.vandervoort@tudelft.nl, nihit.goyal@tudelft.nl, a.verbraeck@tudelft.nl

Abstract

Responsible AI (RAI) governance is increasingly understood not as a static checklist of principles, but as a dynamic process embedded in institutional, organisational, and sociotechnical contexts. While several ethical frameworks exist, translating high-level principles into situated organisational practices remains challenging. Empirical studies examining how public sector organisations operationalise RAI remain fragmented, limiting cumulative insights. To address this gap, we conduct a realist synthesis review of 21 empirical studies. Our analysis shows that similar interventions in different contexts activate distinct mechanisms and produce divergent outcomes with varying degrees of alignment to RAI principles. From these variations, we identify three cross-cutting dynamics explaining outcomes: organisational embeddedness, power-expertise tensions, and trust-transparency relationships. Together, we term it the situated dynamics of RAI governance. This approach moves beyond asking whether interventions "work" to explain why similar interventions succeed in some contexts and fail in others.

Introduction

Responsible AI (RAI) is commonly described as a set of ethical principles, such as fairness, accountability, and transparency, intended to guide the development and use of AI systems (Jobin, Ienca, and Vayena 2019; Hagendorff 2020; Morley et al. 2020). Rooted in socio-technical perspectives, RAI rejects the idea of treating AI as a neutral technological artefact and instead emphasises how technologies are shaped by social, political, and institutional conditions (Verbeek 2006; Winner 1980). Responsible AI governance builds on this view: it is increasingly understood not as a static checklist of principles but as a dynamic governance process embedded in organisations and shaped by governance structures, technical realities, political contexts, and institutional capacity (Dignum 2019; Floridi and Cowls 2022).

In recent years, governments, companies, and international bodies have issued over 100 ethical frameworks for AI (Jobin, Ienca, and Vayena 2019). However, translating these high-level principles into actionable organisational practices remains a significant challenge (Hagendorff 2020; Morley et al. 2020; Taeihagh 2021). The issues include a lack of concrete methods or metrics (Mittelstadt 2019; Peters et al. 2020), spread of responsibility across teams, blurring accountability (Schiff et al. 2021), and organisational dynamics that discourage ethical reflection or whistleblowing (Hagendorff 2020; Morley et al. 2023). In response, the RAI research focus has shifted from defining abstract ideals to investigating how responsibility is operationalised (Morley et al. 2020, 2023).

The practice of RAI is even more pressing in the public sector. Governments and institutions must uphold key public values such as fairness, justice, democratic participation, and accountability (Bozeman 2007; Symes 1999). Their value-driven mandates require approaches that align AI development and use with public service obligations (Fatima, Desouza, and Dawson 2020). At the same time, they face unique organisational constraints, legal obligations, and political pressure.

Recent studies focus on different aspects of Responsible AI practices in public sector organisations. Some investigate national or municipal strategies, such as algorithm registers or oversight instruments to enhance transparency (Kuziemski and Misuraca 2020). Others explore internal practices, such as fairness in predictive analytics (Veale, Van Kleek, and Binns 2018; Dankloff et al. 2025) or governance routines during system design (Henriksen and Blond 2023). A growing number of studies investigate sector-specific implementation challenges, such as welfare automation (Rinta-Kahila et al. 2021), predictive policing (Donatz-Fest 2025) and infrastructure planning in low-capacity settings (Nisar et al. 2022).

Yet, most research relies on single or small-n case studies, which offer rich detail but limited generalisability, leaving findings scattered. While there are emerging systematic reviews of RAI governance (Batool, Zowghi, and Bano 2023; Lu et al. 2024) and of AI use in governments (Zuiderwijk, Chen, and Salem 2021), none focus on empirical public sector implementation or the factors behind different outcomes. This review addresses the gap by examining how public sector organisations put Responsible AI into practice.

Research Design

To do so, we adopt a realist synthesis approach (Pawson et al. 2005), which explains how interventions embedded in specific contexts activate mechanisms that lead to divergent outcomes (ICMO). This method is well-suited to understanding complex interventions (Rycroft-Malone et al. 2012; Sheldon 2005) to move beyond whether interventions work to understanding why their effects vary.

Search Strategy and Study Selection

Following RAMESES guidelines for realist and meta-narrative evidence syntheses (Wong et al. 2013), we conducted a systematic literature search in March 2024 using Dimensions database. We selected Dimensions because it indexes peer-reviewed and policy-relevant publications across disciplines, capturing diverse RAI governance literature. We aimed to identify empirical studies on Responsible AI operationalisation within public sector organisations.

Search terms (Table 1) were structured around three core themes: AI, RAI concepts, and public sector settings. RAI search terms were based on established AI ethics and governance terminology (Dignum 2019; Hagendorff 2020; Schiff et al. 2020). Initial search produced 1,301 documents.

Search Terms
("artificial intelligence" OR "AI" OR "machine learning" OR "Large language model*" OR "Natural language processing" OR “algorithmic decision-making" OR "algorithmic governance") AND (ethic* OR responsib* OR trustworth* OR accountab* OR privacy OR fair* OR safe* OR transparen* OR robust* OR just* OR explainab* OR “human oversight” OR “human autonomy” OR diversity OR discriminat* OR “bias”) AND ("AI governance" OR "Artificial Intelligence Governance" OR "public admin*" OR "public sector" OR "public service*" OR “public agenc*” OR “public organisation”)

Table 1. Search Keywords used in Dimensions Database

We manually screened abstracts in Rayyan using the following inclusion criteria: Focus specifically on AI (not general digital technologies), address RAI concerns (e.g., fairness, transparency), take place in public sector settings, be empirical in nature (e.g., interviews, case studies), and examine AI implementation at the organisational level. This process identified 89 studies for full-text review. After reapplying the inclusion criteria, 21 studies were retained for final analysis (Figure 1).

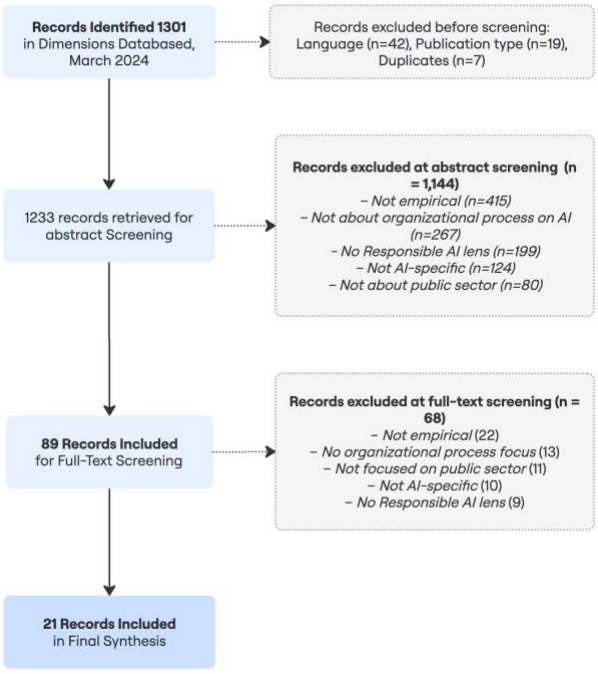

Figure 1. Study identification and selection process

Data Analysis

For the 21 selected studies, we first adapted realist synthesis definitions of intervention, context, mechanism, and outcome (ICMO) to the focus of this review (Table 2). Using these definitions, we extracted relevant data in the original language of the source materials, then applied inductive thematic analysis within each ICMO element (Braun and Clarke 2006) in Atlas.ti. This process involved familiarisation, coding, and iterative refinement, with interpretive judgement particularly important for identifying mechanisms, which were rarely explicit in source materials (Greenhalgh et al. 2011; Wong et al. 2013). The resulting themes were organised into typologies for each ICMO element, also presented in Table 2. These typologies then served as the basis for constructing ICMO chains, allowing us to link elements across studies and identify recurring cross-case patterns (Figure 2).

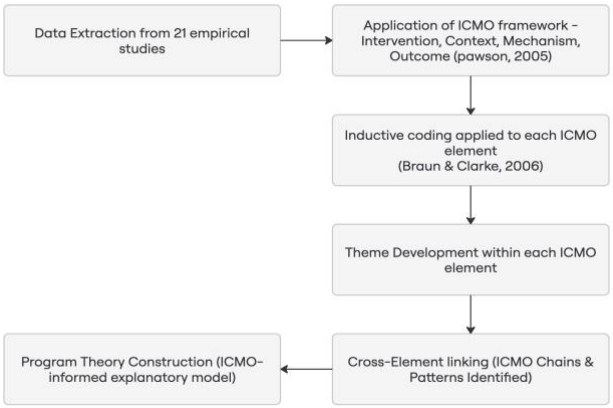

Fig 2. Analytical process for thematic classification and synthesis of ICMO elements across 21 empirical studies

Intervention types were grouped according to their nature in RAI governance: 1) Policy and Legal, (2) Organisational, (3) Participatory, (4) Technological.

Contexts were grouped as internal or external factors. Internal factors included organisational culture and norms, role definitions, workforce capacity, and technological infrastructure. External factors captured influences beyond organisational control, including regulatory frameworks, political priorities, and socio-cultural conditions, which created opportunities and constraints for RAI governance.

Mechanisms were grouped by the responses they activated: Proactive Compliance (anticipatory alignment), Symbolic Compliance (performative actions), Reactive Compliance (crisis-driven responses), Ad Hoc Implementation (inconsistent or improvised application of rules), Reliance Mechanisms (delegation of decision-making to AI systems), Oversight and Contestation (enabling challenge), and Learning and Adaptation (iterative improvement).

Outcomes were coded inductively from empirical patterns while aligning with established RAI principles: Accountability (formal oversight and responsibility), Transparency (visibility of system operations), Fairness & Bias Mitigation (preventing discrimination), Privacy & Data Protection (data governance compliance), and Public Trust (perceived legitimacy). Each outcome was examined for both intended and unintended effects. Working definitions were developed to address variation in how these principles are interpreted across studies. These categories provide the structure for the cross-case synthesis presented in the findings.

Findings

The 21 papers span diverse countries but are notably concentrated in Western contexts, particularly Scandinavia (9 studies), suggesting that both research and documented applications of RAI are predominantly situated in Western, high-income countries. Most studies focus empirically on AI-based decision-support systems within domains such as welfare, immigration, education, and public safety. Across the sample, no single use case emerged as dominant, highlighting the exploratory and experimental stage of public-sector AI deployment. Publication outlets cover policy, public administration, information systems, and socio-technical fields, highlighting the interdisciplinary nature of research on public sector AI governance. An overview of selected papers is presented in Table 3.

Across these studies, we identified 75 distinct interventions: Policy/Legal (12), Internal Governance Interventions (24), Participatory (16), and Technological (22). The most frequently cited contexts were workforce capacity & literacy (33) and structures & roles (23), while common activated mechanisms included contestation & oversight (19) and learning & adaptation (14).

ICMO element	Definition	Types
Intervention	Organisational actions undertaken to support RAI governance.	Policy & Legal, Internal Governance, Participatory, Technological
Context	Internal or external conditions as part of the setting in which an intervention was implemented	Organisational Culture & Norms, Structures & Roles, Workforce Capacity & Literacy, Technological Capacity; Regulatory & Policy Environment, Political and Economic priorities, Public Trust & Socio-cultural Norms
Mechanism	Underlying processes triggered by interventions explain how they lead to outcomes	Proactive Compliance, Symbolic Compliance, Reactive Compliance, Ad-hoc Mechanism, Discretionary Mechanism, Contestation & Oversight, Reliance Mechanism
Outcome	Intended or unintended effects related RAI	Accountability, Transparency, Fairness & Bias Mitigation, Privacy and Data Protection, Public Trust and Legitimacy

Table 2. Adapted ICMO Definitions (from Pawson et al., 2005) and Inductive Typologies Derived from Empirical Findings in RAI Governance

No.	Study	Journal/Conference	Case Country
1	Kuziemski and Misuraca (2020)	Telecommunications Policy	Canada, Poland, Finland
2	Rinta-Kahila et al. (2021)	European Journal of Information Systems	Australia
3	Saxena and Guha (2023)	Journal of Responsible Computing	USA
4	Nisar et al. (2022)	ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization	India
5	Criado, Valero, and Villodre (2020)	Information Polity	Spain
6	Alshahrani, Dennehy, and Mäntymäki (2022)	Government Information Quarterly	Saudi Arabia
7	Dankloff et al. (2025)	AI & Society	Netherlands
8	Henriksen and Blond (2023)	Social Studies of Science	Denmark
9	Veale, Van Kleek, and Binns (2018)	CHI - Conference on Human Factors in Computing Systems	5 OECD country
10	Leikas et al. (2022)	Frontiers in Artificial Intelligence	Finland
11	Chenou and Rodríguez Valenzuela (2021)	The Law, State and Telecommunications Review	Colombia
12	Kinder et al. (2023)	Government Information Quarterly	Finland
13	Jørgensen and Nissen (2022)	Big Data & Society	Denmark
14	Koskimies and Kinder (2022)	Public Management Review	Finland
15	Mahomed et al. (2023)	pre-print	UK
16	Tangi, van Noordt, and Rodriguez Müller (2023)	Conference on Digital Government Research	Various EU countries
17	Hinton (2023)	International Journal of Technotics	Estonia
18	Figueras, Verhagen, and Cerratto Pargman (2022)	Scandinavian Journal of Information Systems	Sweden
19	Berman, de Fine Licht, and Carlsson (2024)	Technology in Society	Sweden
20	Tsourma et al. (2023)	IHIET 2023 - Human Interaction & Emerging Technologies	Italy, Greece, Norway
21	Fest et. al (2023)	Public Management Review	Netherlands

Table 3. An overview of Selected Studies

Notably, intended outcomes focused on transparency (26) and accountability (14), while unintended outcomes clustered around accountability failures (28) and fairness issues (17). The following sections examine each subtype of intervention, tracing how contexts, mechanisms, and outcomes interact. Findings are organised by intervention type to compare how similar interventions produced different effects across settings.

Policy and Legal Interventions

Policy and legal interventions appeared 12 times across 10 studies, including impact assessments, audit procedures, and AI-specific mandates. While they achieved transparency and accountability (3 cases each), they also frequently resulted in compliance failures (5 cases) due to the symbolic implementation or bureaucratic obstacles. Two contextual factors shaped these outcomes: regulatory clarity and enforcement (5 cases) and organisational culture and structure (4 cases). These conditions activated different proactive, symbolic, ad hoc, or reactive compliance mechanisms, leading to divergent outcomes (Table 4).

Audit procedures were mentioned four times across three studies. In contexts with stronger regulatory frameworks and capacity (Studies 5, 7), audits activated proactive compliance that supported accountability and fairness, though with added burdens or externalised responsibility. By contrast, in fragmented organisational settings (Study 14), audits slipped into symbolic compliance, improving external transparency but failing to secure accountability.

Impact assessments (Studies 13, 15, 21) showed a similar drift from proactive to symbolic compliance. While in the design phase, gains in transparency and privacy protection were recorded, they eroded during implementation as organisational cultures prioritised documentation over action.

AI-specific mandates further presented regulatory dependence: Canada's Automated Decision-Making Directive (Study 1) triggered transparency through proactive compliance, yet weak enforcement and limited organisational embedding led to biased testing. Australia's absent framework (Study 2) activated reactive responses only after Robodebt's public backlash, resulting in unfairness and financial hardship for affected citizens.

Intervention	Context	Mechanism	Outcome
Audit[5,7,14]	Regulatory environment; Organisational Structure & Capacity	Proactive, Symbolic compliane	✓ Accountability, Transparency, Fairness & Bias Mitigation; ✕ Accountability
Impact Assess-ments[13,15,21]	Organisational Culture & Norms	Proactive, Symbolic compliance	✓ Transparency, Privacy & Data Protection
AI Legal Safeguards[1,2,16]	Regulatory & policy environment	Proactive,Reactive Compliance	✓Transparency; ✕ Accountability

Table 4. An overview of the policy and legal interventions and their most associated contexts, mechanisms, and outcomes

Internal Governance Interventions

Internal Governance interventions appeared 25 times across 16 studies, including capacity building and role design. While achieving transparency (12 cases) and accountability (9 cases), they also notably created accountability gaps (16 cases) through eroded professional discretion and misaligned roles. Two contextual factors shaped outcomes: Workforce Capacity and Literacy (11 cases) shaped whether staff could interpret and challenge new responsibilities, while Organisational Structures and Role Distribution (9 cases) influenced whether governance roles were clearly defined and practically supported. As a result, these conditions activated learning and adaptation or discretionary mediation mechanisms (Table 5).

Capacity building and learning programs appeared in 10 studies. In low-literacy, high-turnover contexts, they activated learning mechanisms that improved transparency by helping staff interpret AI outputs (Studies 4, 7, 12). However, training sometimes increased AI reliance and weakened oversight, leading to accountability failures (Studies 3, 14, 15). Where organisational roles were better defined, programs activated discretionary mediation, strengthening professional discretion and improving accountability (Study 12). Yet they also revealed context-specific trade-offs: persistent scepticism (Study 5), decision-making imbalances (Study 12), disconnects between system design and day-to-day application (Study 21), and ethics training often failed to translate into concrete oversight practices (Study 19). In these cases, training raised awareness but lacked contextual adaptation and support.

Organisational structures and role design appeared in 13 studies, including ethics boards and newly created AI units. In defined structures, they activated oversight and discretionary mediation mechanisms, improving transparency (study 3,16), accountability (study 6,9,10,14,17,18,21), and fairness (study 9, 17). For instance, Finland's AuroraAI ethics board enabled cross-disciplinary deliberation and supported anonymised data use (Study 10), while expert-led quarterly validations improved employment service fairness (Study 16). However, even effective structures faced constraints like late-stage setup limiting influence (Study 10). In these cases, accountability was not fully realised due to a lack of institutional power (Study 3), leadership misalignment (Study 6), unclear enforcement (Studies 9, 16), and role confusion or procedural fragmentation (Studies 17, 18, 21).

Participatory Interventions

Participatory interventions appeared 16 times in 12 studies, including citizen engagement (co-design workshops, citizen panels), internal stakeholder processes (feedback loops, workshops), and ecosystem initiatives (open data publishing). They most frequently achieved transparency (10 cases), and public trust & legitimacy (9 cases). Yet also produced accountability failures and fairness limitations (4 cases each). Outcomes were shaped by two key contexts: Public Trust and Socio-Cultural Settings (8 cases) influenced participatory process uptake and inclusiveness, while organisational structures and workforce capacity (7 cases) shaped whether the input was formally embedded or

Intervention	Context	Mechanism	Outcome
Capacity Building & Learning Programs [3, 4, 5, 7, 12, 13, 14, 15, 17, 19, 21]	Structures & Roles, Workforce capacity & Literacy	Learning & Adaptation; Discretionary Mediation; Reliance	✓ Transparency, Accountability; ✕ Accountability
Structures & Role Design [3, 6, 9, 10, 14, 16, 17, 18, 21]	"	Contestation & Oversight; Discretionary Mediation	✓Accountability, Fairness & Bias Mitigation, Transparency; ✕ Accountability

Table 5. An Overview of Internal Governance Interventions and their associated contexts, mechanisms, and outcomes

remained advisory. These conditions activated learning and adaptation or contestation and oversight mechanisms (Table 6).

Citizen-facing participatory interventions appeared in 8 studies. In high-trust contexts with strong participatory expectations, they activated contestation and oversight mechanisms, achieving transparency and legitimacy (Studies 1,10,12). For example, Finland's co-design workshops (Study 10) integrated diverse stakeholders into AI service development, with participants seeing their input shape features. However, outcomes were often partially realised due to coordination burdens and limited integration capacity. In contrast, in developer-driven contexts, the same interventions activated a reactive compliance mechanism, focusing on technical fixes while excluding broader fairness concerns (Study 17). Similarly, weak institutional structures undermined initial gains: involving children in design created developer reflection (Study 15) but lacked follow-through mechanisms to embed input into final decisions.

Beyond citizen input, *Internal Stakeholder-Oriented Interventions* appeared in 5 studies (6 times). In well-regulated contexts, they activated learning mechanisms that improved fairness by integrating local expertise into prediction models (Study 9), reorienting priorities from cost-efficiency toward child protection ethics (Study 13) and facilitating stakeholder alignment on fairness principles and data-sharing protocols (Study 14). However, they revealed context-specific trade-offs. These included decision paralysis from competing priorities (Study 14), exclusion of vulnerable voices when technically confident stakeholders dominated agenda-setting (Study 12), feedback manipulation due to weak validation (Study 16), and unrealistic expectations as political shifts forced frontline staff beyond institutional capacity (Study 13). In low literacy contexts, developer dominance similarly led to misaligned fairness considerations when technical teams set priorities without challenge from domain experts (Study 7).

At a broader level, *ecosystem-oriented* participatory interventions appeared in 2 studies. In high-trust, open collaboration contexts, they improved public trust and transparency by enabling citizen scrutiny (Studies 12, 16). However, opening data infrastructure to private companies created accountability ambiguity when responsibility for AI-driven decisions remained unclear (Study 12).

Technological Interventions

Technological interventions covered: AI model design, data practices, and explainability tools, appearing 22 times across 13 studies. While achieving public trust (5 cases) and privacy protection (4 cases), unintended effects were more common than intended ones (17 vs. 11), particularly fairness failures (9) and transparency limitations (6). These interventions consistently improved operational efficiency (10 cases) but at the cost of due process and professional judgment, manifesting in rigid assessments. Workforce capacity (7 cases) and technological infrastructure (5 cases) frequently shaped outcomes, often activating contestation mechanisms where users challenged system use (Table 7).

Explainability features and transparency tools appeared in 7 cases, including interpretability methods and tailored communication interfaces. In contexts with front-line discretion and high public trust, they activated contestation mechanisms that increased public trust and legitimacy (Studies 9,10,12,13). For example, Finnish municipal staff adapted explanations for different audiences (Study 12), while citizen-facing tools helped users understand how data-informed recommendations (Study 10). However, these tools consistently fell short of their transparency goals across multiple cases: Explanations constrained enabling understanding (Study 13), weak organisational support for training left non-technical staff unable to translate outputs and act on outputs (Study 12), and formal compliance measures provided visibility without substantive understanding (Study 10, 19).

Intervention	Context	Mechanism	Outcome
Citizen Oriented[1,4,10,12,15,16,17,18]	Public Trust & Socio-Cultural; Structures & Roles; Workforce Capacity & Literacy	Contestation & Oversight, Ad & Hoc, Reactive	✓ Transparency; Fairness & Bias Mitigation; Public Trust & Legitimacy; ✕ Accountability
Internal Stakeholder oriented [7,9,12,13,14]	Structures & Roles; Workforce Capacity & Literacy; Regulatory & Policy environment, Political and Economic Priorities	Contestation & Oversight, Learning & Adaptation	✓ Transparency; Public Trust & Legitimacy; Fairness & Bias Mitigation; ✕ Fairness & Bias Mitigation, Transparency
Ecosystem Oriented [12,16]	Public Trust & Socio-Cultural	Contestation & Oversight, Learning & Adaptation	✓ Transparency; Public Trust & Legitimacy; ✕ Accountability

Table 6. Overview of Participatory Interventions along with their most associated contexts, mechanisms, and outcomes

Intervention	Context	Mechanism	Outcome
AI Model Design & Data Practices[2,3,5,8,9,12,13,16,17,20, 21]	Technology Capacity & Infrastructure; Structure; &Roles; Workforce Capacity & Literacy	Proactive, Discretionary, Symbolic	✓ Privacy & Data Protection ✕ Fairness & Bias Mitigation, Privacy & Data Protection
Explainability features & transparency Tools[9,10,12,13,16,19]	Public Trust & Socio-Cultural Setting; Organisational Culture	Contestation & Oversight	✓ Public Trust & Legitimacy; ✕ Transparency

Table 7. An Overview of Technological Interventions and their most associated contexts, mechanisms, and outcomes.

AI model design and data practices appeared in 14 cases, including input simplification, exclusion of sensitive variables, and model repurposing across domains. In capacity-limited settings, they activated mixed mechanisms with predominantly poor outcomes: only 1 of 5 cases achieved intended results, anonymization enhancing privacy (Study 21), while 4 resulted in fairness failures through inappropriate design choices: transferring a financial AI model to healthcare without domain adaptation (Study 8), and using simplistic income averaging that systematically overestimated casual workers' earnings (Study 1). Similar patterns emerged in contexts with rigid structures or workforce constraints, again producing fairness failures (4 cases). Across both contexts, interventions consistently achieved efficiency gains but at the cost of fairness, professional judgment, and procedural consistency.

Discussion

Our findings show that interventions intended to advance Responsible AI principles rarely worked in straightforward ways. For example, audits and impact assessments improved transparency only when supported by regulatory enforcement and organisational embedding; otherwise, they drifted into symbolic exercises. Training initiatives, meant to enhance professional judgment, sometimes deepened automation bias when not adapted to local capacity. Organisational structures like ethics boards created spaces for deliberation yet often lacked authority to translate recommendations into practice. Participatory processes provided voice, but too often without uptake, reinforcing existing role ambiguities and power asymmetries. Explainability tools, which in high-trust contexts improved visibility, frequently delivered little more than symbolic transparency. Taken together, these patterns highlight a central insight: interventions cannot be judged in isolation, as their effects depend on the mechanisms they activate under specific internal and external contexts.

What became visible across cases is that certain cross-cutting dynamics consistently shaped whether interventions were sustained or drifted into symbolic forms. We refer to these as the Situated Dynamics of Responsible AI Governance, which consists of three recurrent and interconnected dynamics: organisational embeddedness, power-expertise, and trust-transparency.

Organisational Embeddedness Dynamics

Across the cases, interventions achieved their goal when embedded into internal contexts, with external conditions such as regulatory clarity acting as enablers. For instance, audits and impact assessments, for instance, increased transparency when reinforced by clear regulatory oversight and supported by organisational structures for follow-up (Studies 5, 7). Yet in other contexts, the same interventions drift into symbolic compliance when they were reduced to documentation exercises without enforcement or role clarity (Studies 13, 14, 15, 21). Ethics boards and newly created governance roles showed a similar split: Finland's AuroraAI board enabled cross-disciplinary deliberation and supported anonymised data use (Study 10), but in settings where such bodies lacked institutional power, leadership alignment, or enforcement mechanisms, leaving accountability gaps (Studies 3, 6, 9, 16). Similarly, technological interventions such as explainability tools depended on this embedding: when frontline staff had the literacy and organisational support to adapt outputs for different audiences, they fostered trust and transparency (Studies 9, 10, 12). Without training or structural follow-through, however, the same tools produced only symbolic transparency (Studies 10, 12, 19).

We call this recurring pattern Organisational Embeddedness Dynamics. It captures how the effectiveness of Responsible AI interventions depends less on their nature than on whether they are internalised into structures, roles, and everyday routines of organisations. When embedded, interventions became part of decision-making processes, activating mechanisms such as proactive compliance or learning, and leading to accountability and transparency. When left as one-off fixes, trainings without follow-up, audits reduced to documentation, or ethics boards without authority, they drifted into symbolic or ad hoc responses, producing accountability gaps and oversight failures. External conditions, such as regulatory clarity or public trust, shaped whether embedding was sustained. Still, they

were not sufficient on their own: without organisational routines that embedded new practices, interventions remained fragile and easily eroded.

The gap between Responsible AI principles and practice is well recognised in the literature (Mittelstadt 2019; Morley et al. 2023), often explained through the difficulty of translating broad ethical commitments into situated organisational routines. Our findings resonate with this observation but point to a deeper root: an Organisational Embeddedness Dynamic that shapes whether interventions consolidate or erode in practice. This dynamic reflects what institutional theory terms incomplete institutionalisation, practices introduced but not yet stabilised as norms (Tolbert and Zucker 1996; Scott 2014) and in some cases approaches decoupling, where changes are maintained primarily for symbolic legitimacy (Meyer and Rowan 1977). Rather than being separate or sequential, our analysis shows how these processes intersect: decoupling pressures are especially strong in early stages of institutionalisation, when practices remain fragile and norms are absent (Bromley and Powell 2012). This is precisely where Responsible AI governance currently sits, which explains why Responsible AI interventions are so vulnerable. Novelty, technical opacity, and the absence of established standards mean that embedding inevitably takes time (Hajer 2003; Klievink, van Wegberg, and van Eeten 2017), yet regulatory timelines and legitimacy demands (AI HLEG 2019; OECD 2022) push organisations into strategic survival mode, adopting interventions as one-off fixes rather than integrating them into workflows. In this sense, organisational embeddedness is not only about internal structures and roles but about the capacity to resist drifting into symbolic compliance under persistent external pressure. Therefore, RAI governance depends on whether organisations can reduce decoupling pressures and strengthen institutionalisation by embedding interventions into everyday organisational practices.

Power-Expertise Dynamics

Another dynamic that consistently shaped the outcomes of Responsible AI interventions concerned the distribution of power and expertise. Across cases, authority over interventions frequently rested with technical teams, narrowing the scope of influence for other actors. The design and implementation of audits and impact assessments, for example, were defined by developers, sidelining organisational and domain expertise (Studies 4, 15, 21). In model design and data practices, authority concentrated in technical specialists or blurred by role ambiguity led to expertise dominance, including the transfer of systems across domains without contextual adaptation, which excluded practitioners with local knowledge (Studies 3,8,9,15). Participatory interventions reflected similar asymmetries: in low-literacy settings, technically confident stakeholders dominated deliberations, while vulnerable voices were marginalised and fairness priorities primarily set by developers (Studies 7, 12). Thus, interventions intended to broaden accountability instead reinforced epistemic hierarchies, with technical authority outweighing other perspectives.

We refer to this recurring pattern as Power-Expertise Dynamics. It captures how Responsible AI interventions are shaped by how authority and expertise are distributed among technical teams, domain practitioners, and affected stakeholders. When technical expertise dominated, defining the scope and interpretation of audits, leading model transfers without contextual adaptation, or steering participatory processes, interventions undermined accountability and fairness. Where expertise was shared across domains, roles were clearly defined, and follow-up mechanisms ensured that input translated into practice, interventions led to accountability and fairness.

Public administration scholarship has long examined how discretion shifts across levels of bureaucracy. In Lipsky's (1980) account, street-level bureaucrats exercise judgment in direct encounters with citizens, while later work highlighted the rise of system-level bureaucracy, where discretion is encoded upstream in IT systems and central procedures (Bovens and Zouridis 2002). More recent studies describe screen-level bureaucrats, whose interactions with citizens are mediated by system outputs, leaving them with limited scope to contest or reinterpret decisions (Landsbergen 2004). Our cases suggest that AI governance pushes this trajectory further, into what might be called code-level bureaucracy: developers and data scientists embed rules and value-laden choices directly in code and model parameters. Discretion did not simply relocate from frontline staff to administrators. Still, it was consolidated in technical design, narrowing the role of domain expertise and concentrating authority in technical specialists, who not only implement but also define the parameters that shape accountability and fairness outcomes, an intensified form of the information asymmetries highlighted by agency theory (Jensen and Meckling 1976).

AI amplifies these dynamics in three ways. First, expertise is unstable: models degrade, and standards shift, making knowledge of fairness or privacy quickly obsolete (Veale, Van Kleek, and Binns 2018). Second, opacity persists even for developers, who nevertheless arbitrate decisions (Pasquale 2015; Lipton 2018; Passi and Barocas 2019). Third, in contexts of role ambiguity and uneven literacy, technical staff assumed de facto responsibility, not by mandate but by default. Together, these conditions concentrate authority in technical teams, displacing domain expertise and embedding normative choices in system design. The result is that interventions meant to strengthen

Responsible AI struggle to deliver accountability and fairness when expertise remains unevenly distributed.

Trust-Transparency Dynamics

Across cases, we also found that the trajectory of RAI interventions pivoted on how trust and transparency were configured. Across intervention types, we observed a recurring pattern: interventions that were expected to build public trust through greater transparency often produced very different results depending on context. Transparency took the form of both technical explainability measures (dashboards, feature visualisations) and information-oriented efforts (training and literacy sessions intended to help staff or citizens interpret AI outputs). Yet across these variants, when introduced without sufficient interpretive capacity, transparency measures tended to trigger misplaced reliance and automation bias rather than fostering contestation and critical oversight (Studies 2, 3, 7, 9, 13, 19). At the same time, baseline trust relations shaped whether transparency landed as meaningful. In supportive environments, transparency measures could be reinforced through participatory norms and feedback loops, enabling contestation and strengthening legitimacy (Studies 1, 10). In less supportive settings, however, formal or technical transparency was more likely to collapse into symbolic compliance, producing reliance rather than scrutiny (Study 19).

We refer to this recurring pattern as Trust-Transparency Dynamics, a dynamic that helps explain why interventions often diverge in practice. It captures how interventions oriented toward openness and visibility, whether through technical explainability tools or broader information-oriented measures, could either foster or erode Responsible AI, depending on whether transparency was rendered meaningful. In some cases, trust was framed as the intended outcome of transparency, yet visibility without understanding tended to produce reliance or symbolic compliance instead. In other cases, trust acted as a contextual condition: supportive trust relations and interpretive capacity amplified the effectiveness of transparency measures, while their absence redirected them into misplaced reliance. This dynamic underscores that transparency cannot be treated as a straightforward lever for legitimacy, but as a relational process contingent on the social and organisational scaffolding that makes information actionable.

Seen in this light, Trust-Transparency Dynamics resonate with but also complicate existing assumptions in Responsible AI frameworks. These frameworks consistently position transparency as a pathway to trust, operating under the assumption that visibility generates confidence and legitimacy (AI HLEG 2019; OECD 2022). This view - transparency as input, trust as output - is reinforced by experimental studies showing positive correlations between explanation provision and user trust (Ribeiro, Singh, and Guestrin 2016; Dodge et al. 2019; Park and Yoon 2024). Yet these studies typically examine individual user responses to isolated interventions in controlled settings, overlooking the institutional and contextual conditions our synthesis reveals as critical. Where visibility existed without enabling scrutiny (Ananny and Crawford 2018), our cases showed that transparency not only failed to deliver its intended outcomes but could also reshape user relations to AI in ways that discouraged critical engagement. Explainability tools often created technical transparency, making system operations visible, but without interpretive scaffolding, they tended to encourage reliance and narrow the space for contestation. In this sense, our cases highlight the gap between technical transparency and democratic transparency: the former delivers visibility, the latter enables meaningful oversight and institutional accountability. Realising the latter requires governance arrangements that anchor transparency in deliberation and democratic control over AI (Wong et al. 2025).

This conditional relationship between trust and transparency reflects but also extends existing critiques. AI trust research increasingly acknowledges the contextual and multi-dimensional character of trust (Knowles, Richards, and Kroeger 2022) yet has not examined how these dimensions interact with transparency in institutional settings. Organisational studies emphasise that trust depends not on disclosure alone but on the conditions under which information is interpreted (Schnackenberg and Tomlinson 2014) but these accounts were developed for traditional organisational contexts. Public-sector AI adds new challenges: technical opacity, automated decision-making at scale, and the difficulty of explaining algorithmic logic to diverse stakeholder groups. Our findings suggest that the trust-transparency dynamic in AI governance cannot be understood as a simple input-output relation but as an interdependent process shaped by both institutional context and the form of transparency enacted. For practitioners, the implication is clear: transparency is not the destination but the start of a process. Unless it is paired with interpretive capacity and conditions for contestation, it risks producing only symbolic visibility or misplaced reliance, rather than the legitimacy Responsible AI aspires to secure.

Together, these dynamics form the Situated Dynamics of RAI Governance, interacting forces that shape how responsible AI interventions are enacted, sustained, or derailed in the public sector. While each dynamic captures a distinct dimension, embeddedness, expertise distribution, and the trust-transparency relationship, they are deeply interdependent: shifts in one can reinforce, destabilise, or reconfigure the others. Current evidence suggests that AI-

specific characteristics such as technical opacity, model fluidity, and the scarcity of relevant expertise amplify all three. Yet, the precise causal pathways linking these dynamics remain underexplored.

Limitations

As with all realist syntheses, our findings depend on the quality and completeness of the source studies. The literature in the public sector RAI is fragmented, with inconsistent terminology, reporting standards, and causal framing. Because many studies did not clearly define mechanisms or outcomes, we often had to make interpretive inferences during coding. Ambiguity was common; for instance, it was sometimes unclear whether a passage described a mechanism or a context, or whether outcomes reflected unintended effects or weak implementation. While this introduced subjectivity, such inference is an accepted and necessary feature of realist synthesis. As Pawson et al. (2005), Wong et al. (2013), and Greenhalgh et al. (2011) emphasise, realist reviews rely on theory-informed judgement to surface plausible explanations of how interventions work across varying contexts. To mitigate bias, we used memoing, collaborative review, and repeated checks of the source material. These interpretive judgments are particularly essential in value-laden domains such as Responsible AI governance, where institutional norms, legal ambiguities, and sociopolitical dynamics shape causal pathways. Although realist synthesis originated in health and social care (Jagosh et al. 2012; Pawson et al. 2005), and it is increasingly used in education, policy, and public health, the RAI governance literature remains less coherent than in those fields, where interventions tend to be more standardised. Additionally, we acknowledge that not every observed pattern can be attributed to a single context shaping a mechanism. As in other forms of causal inference, confounding factors may be present, unmeasured variables that influence both the intervention and the outcome, making it difficult to establish direct causality. Our aim is therefore not to claim definitive causal relationships, but to propose plausible, theory-informed explanations for how particular outcomes may arise in different contexts. Finally, we note that our focus on English-language, peer-reviewed, and policy-relevant empirical work may have excluded valuable insights from non-academic or non-English sources, particularly from underrepresented regions.

Conclusion

Based on a realist synthesis of 21 empirical studies on RAI governance in public sector settings, our analysis shows that widely promoted interventions - such as audits, training programs, and participatory processes - are not inherently effective. Their outcomes are shaped by organisational norms and structures, capacity and expertise, and the broader sociotechnical contexts. Moreover, interventions intended to advance RAI principles often generated trade-offs that undermined those same principles: transparency tools obscured more than they clarified, training reinforced reliance on AI rather than critical oversight, and participatory initiatives that offered voice without meaningful uptake. These examples highlight that what matters is not the intervention in isolation, but the mechanisms it activates under particular conditions.

To explain these patterned effects, our synthesis develops three interdependent dynamics of Situated RAI Governance: organisational embeddedness dynamics, power-expertise, and trust-transparency dynamics. These dynamics were made visible by applying realist synthesis, rarely used in AI governance, and by developing an inductive ICMO taxonomy that structured comparisons across cases. Together, they show that responsible AI governance is not secured through design alone but requires continuous reinforcement within organisational routines and institutional structures. Even well-aligned interventions can erode over time, sliding into symbolic compliance when capacity, follow-through, or role clarity weaken. Rather than offering a prescriptive checklist, this paper explains why interventions succeed or fail across different conditions, contributing to a more explanatory, practice-informed understanding of Responsible AI governance.

At the same time, we note important gaps for future research. Despite the large and growing literature on AI governance, only 21 empirical studies directly examine what works in practice, most situated in high-income Western contexts. Very little is known about how RAI governance operates in low-resource or non-Western settings. While this paper contributes to the emerging focus on how to operationalise RAI governance, what constitutes RAI itself is also evolving. The reviewed studies centred on transparency, accountability, fairness, privacy, and trust, while environmental considerations were notably absent. Future research should examine their intersections empirically, tracing how they unfold over time and across governance contexts. Rather than asking which intervention "works," the critical question is under what conditions interventions are maintained as substantive practices rather than reduced to symbolic gestures. This shift helps move the field from isolated implementation fixes toward a systemic, practice-informed understanding of RAI governance as an evolving socio-technical practice and, through that, more responsible AI governance in the public sector.

References

Alshahrani, A., D. Dennehy, and M. Mäntymäki. 2022. An attention-based view of AI assimilation in public sector organizations: The case of Saudi Arabia. Government Information Quarterly 39(4): 101617. doi.org/10.1016/j.giq.2021.101617.

Ananny, M.; and Crawford, K. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. New Media & Society 20(3): 973–989. doi.org/10.1177/1461444816676645.

Batool, A.; Zowghi, D.; and Bano, M. 2023. Responsible AI Governance: A Systematic Literature Review. arXiv preprint. arXiv:2401.10896 [cs.CY]. Ithaca, NY: Cornell University Library. doi.org/10.48550/arXiv.2401.10896.

Berman, A., K. de Fine Licht, and V. Carlsson. 2024. Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system. Technology in Society 76: 102471. doi.org/10.1016/j.techsoc.2024.102471.

Bozeman, B. 2007. Public Values and Public Interest: Counterbalancing Economic Individualism. Washington, DC:Georgetown University Press. doi.org/10.1353/book13027.

Bovens, M.; and Zouridis, S. 2002. From Street-Level to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control. Public Administration Review 62(2): 174–184. doi.org/10.1111/0033-3352.00168.

Braun, V.; and Clarke, V. 2006. Using Thematic Analysis in Psychology. Qualitative Research in Psychology 3(2): 77–101. doi.org/10.1191/1478088706qp063oa.

Bromley, P., and Powell, W. W. 2012. From Smoke and Mirrors to Walking the Talk: Decoupling in the Contemporary World. The Academy of Management Annals 6(1): 483–530. doi.org/10.1080/19416520.2012.684462.

Chenou, J.-M., and L. E. Rodríguez Valenzuela. 2021. Habeas Data, Habemus Algorithms: Algorithmic intervention in public interest decision-making in Colombia. Law, State and Telecommunications Review 13(2): 56–77. doi.org/10.26512/lstr.v13i2.34113.

Criado, J. I.; Valero, J.; and Villodre, J. 2020. Algorithmic Transparency and Bureaucratic Discretion: The Case of SALER Early Warning System. Information Polity 25(4):449–470. doi.org/10.3233/IP-200260.

Dankloff, M., V. Skoric, G. Sileno, T. Ghebreab, J. van Ossebruggen, and F. Beauxis-Aussalet. 2025. Analysing and organising human communications for AI fairness assessment. AI & Society 40: 2347–2367. doi.org/10.1007/s00146-024-01974-4.

Dignum, V. 2019. Responsible Artificial Intelligence. Cham, Switzerland: Springer International Publishing. doi.org/10.1007/978-3-030-30371-6.

Dodge, J.; Liao, Q. V.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19), 275–285. New York, NY: Association for Computing Machinery. doi.org/10.1145/3301275.3302310.

Donatz-Fest, I. C. 2025. Values? Camera? Action! An Ethnography of an AI Camera System Used by the Netherlands Police. Policing & Society 35(1): 50–67.doi.org/10.1080/10439463.2024.2370939.

Fatima, S.; Desouza, K. C.; and Dawson, G. S. 2020. National Strategic Artificial Intelligence Plans: A Multi-Dimensional Analysis. Economic Analysis and Policy 67:178–194. doi.org/10.1016/j.eap.2020.07.008.

Fest, I., M. Schäfer, J. van Dijck, and A. Meijer. 2023. Understanding data professionals in the police: A qualitative study of system-level bureaucrats. Public Management Review 25(9): 1664–1684. doi.org/10.1080/14719037.2023.2222734.

Figueras, C., H. Verhagen, and T. Cerratto Pargman. 2022. Exploring tensions in responsible AI in practice: An interview study on AI practices in and for Swedish public organizations. Scandinavian Journal of Information Systems 34(2): Article 6. aisel.aisnet.org/sjis/vol34/iss2/6.

Floridi, L.; and Cowls, J. 2022. A Unified Framework of Five Principles for AI in Society. In Machine Learning and the City, 535–545. Hoboken, NJ: Wiley. doi.org/10.1002/9781119815075.ch45.

Greenhalgh, T.; Wong, G.; Westhorp, G.; Pawson, R.; and Green, R. 2011. Protocol – Realist and Meta-Narrative Evidence Synthesis: Evolving Standards (RAMESES). BMC Medical Research Methodology 11: 115. doi.org/10.1186/1471-2288-11-115.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30(1): 99–120. doi.org/10.1007/s11023-020-09517-8.

Hajer, M. 2003. Policy Without Polity? Policy Analysis and the Institutional Void. Policy Sciences 36(2): 175–195. doi.org/10.1023/A:1024834510939.

Henriksen, A.; and Blond, L. 2023. Executive-Centered AI? Designing Predictive Systems for the Public Sector. Social Studies of Science 53(4): 738–760. doi.org/10.1177/03063127231163756.

High-Level Expert Group on Artificial Intelligence (AI HLEG). 2019. Ethics Guidelines for Trustworthy AI. Brussels, Belgium: European Commission.

Hinton, C. 2023. The state of ethical AI in practice: A multiple case study of Estonian public service organizations. International Journal of Technoethics 14(1): 1–15. doi.org/10.4018/IJT.322017.

Jagosh, J.; Macaulay, A. C.; Pluye, P.; Salsberg, J.; Bush, P. L.; Henderson, J.; Sirett, E.; Wong, G.; Cargo, M.; Herbert, C. P.; Seifer, S. D.; Green, L. W.; and Greenhalgh, T. 2012. Uncovering the Benefits of Participatory Research: Implications of a Realist Review for Health Research and Practice. The Milbank Quarterly 90(2): 311–346.doi.org/10.1111/j.1468-0009.2012.00665.x.

Jensen, M. C.; and Meckling, W. H. 1976. Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure. Journal of Financial Economics 3(4): 305–360. doi.org/10.1016/0304-405X(76)90026-X.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence 1(9): 389–399. doi.org/10.1038/s42256-019-0088-2.

Jørgensen, A. M., and M. A. Nissen. 2022. Making sense of decision support systems: Rationales, translations and potentials for critical reflections on the reality of child protection. Big Data & Society 9(2). doi.org/10.1177/20539517221125163.

Kinder, T., J. Stenvall, E. Koskimies, H. Webb, and S. Janenova. 2023. Local public services and the ethical deployment of artificial intelligence. Government Information Quarterly 40(4): 101865. doi.org/10.1016/j.giq.2023.101865.

Klievink, B.; van Wegberg, R.; and van Eeten, M. 2017. EenGezamenlijke Rekening? Bestuurskunde 26(1): 56–64.doi.org/10.5553/Bk/092733872017026001010.

Knowles, B.; Richards, J. T.; and Kroeger, F. 2022. The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency. arXiv preprint. arXiv:2208.00681 [cs.CY]. doi.org/10.48550/arXiv.2208.00681.

Koskimies, E., and T. Kinder. 2024. Mutuality in AI-enabled new public service solutions. Public Management Review 26(1): 219–244. doi.org/10.1080/14719037.2022.2078501.

Kuziemski, M.; and Misuraca, G. 2020. AI Governance in the Public Sector: Three Tales from the Frontiers of Automated Decision-Making in Democratic Settings. Telecommunications Policy 44(6): https://doi.org/10.1016/j.telpol.2020.101976.

Landsbergen, D. 2004. Screen level bureaucracy: Databases as public records. Government Information Quarterly. 21(1): 24–50. https://doi.org/10.1016/j.giq.2003.12.009.

Leikas, J., A. Johri, M. Latvanen, N. Wessberg, and A. Hahto. 2022. Governing ethical AI transformation: A case study of AuroraAI. Frontiers in Artificial Intelligence 5: 836557. doi.org/10.3389/frai.2022.836557

Lipton, Z. C. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery. Queue 16(3): 31–57. doi.org/10.1145/3236386.3241340.

Lu, Q.; Zhu, L.; Xu, X.; Whittle, J.; Zowghi, D.; and Jacquet, A. 2024. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. ACM Computing Surveys 56(2): 1–35.doi.org/10.1145/3626234.

Mahomed, S., M. Briggs, J. Wong, and M. Aitken. 2023. Navigating Children's Rights and AI in the UK: A roadmap through uncertain territory. Preprint. doi.org/10.21203/rs.3.rs-3377300/v1.

Meyer, J. W.; and Rowan, B. 1977. Institutionalized Organizations: Formal Structure as Myth and Ceremony. American Journal of Sociology 83(2): 340–363. doi.org/10.1086/226550.

Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. Nature Machine Intelligence 1(11): 501–507. doi.org/10.1038/s42256-019-0114-4.

Morley, J.; Floridi, L.; Kinsey, L.; and Elhalal, A. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods, and Research to Translate Principles into Practices. Science and Engineering Ethics 26(4): 2141–2168. doi.org/10.1007/s11948-019-00165-5.

Morley, J.; Kinsey, L.; Elhalal, A.; Garcia, F.; Ziosi, M.; and Floridi, L. 2023. Operationalising AI Ethics: Barriers, Enablers, and Next Steps. AI & Society 38(2): 411–423. doi.org/10.1007/s00146-021-01308-8.

Nisar, H.; Gupta, D.; Kumar, P.; Murapaka, S. R.; Rajesh, A. V.; and Upadhyaya, A. 2022. Algorithmic Rural Road Planning in India: Constrained Capacities and Choices in Public Sector. In Proceedings of the 2022 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 1–11. New York, NY: Association for Computing Machinery. doi.org/10.1145/3551624.3555299.

Organisation for Economic Co-operation and Development (OECD). 2022. OECD Recommendation on Artificial Intelligence. Paris, France: OECD Publishing.

Orr, W.; and Davis, J. L. 2020. Attributions of Ethical Responsibility by Artificial Intelligence Practitioners. Information, Communication & Society 23(5): 719–735. doi.org/10.1080/1369118X.2020.1713842.

Park, K.; and Yoon, H. Y. 2024. Beyond the Code: The Impact of AI Algorithm Transparency Signaling on User Trust and Relational Satisfaction. Public Relations Review 50(1):102507. doi.org/10.1016/j.pubrev.2024.102507.

Pasquale, F. 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

Passi, S.; and Barocas, S. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19), 39–48. New York, NY: Association for Computing Machinery. doi.org/10.1145/3287560.3287567.

Pawson, R.; Greenhalgh, T.; Harvey, G.; and Walshe, K.2005. Realist Review – A New Method of Systematic Review Designed for Complex Policy Interventions. Journal of Health Services Research & Policy 10(1_suppl): 21–34. doi.org/10.1258/1355819054308530.

Peters, D.; Vold, K.; Robinson, D.; and Calvo, R. A. 2020. Responsible AI-Two Frameworks for Ethical Design Practice. IEEE Transactions on Technology and Society 1(1): 34–47. doi.org/10.1109/TTS.2020.2974991.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv preprint. arXiv:1602.04938 [cs.LG]. doi.org/10.48550/arXiv.1602.04938.

Rinta-Kahila, T., I. Someh, N. Gillespie, M. Indulska, and S. Gregor. 2021. Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. European Journal of Information Systems 31(3): 313–338. doi.org/10.1080/0960085X.2021.1960905.

Rycroft-Malone, J.; McCormack, B.; Hutchinson, A. M.; DeCorby, K.; Bucknall, T. K.; Kent, B.; Schultz, A.; Snelgrove-Clarke, E.; Stetler, C. B.; Titler, M.; Wallin, L.; and Wilson, V. 2012. Realist Synthesis: Illustrating the Method for Implementation Research. Implementation Science 7(1): 33. doi.org/10.1186/1748-5908-7-33.

Saxena, D., and S. Guha. 2023. Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. arXiv:2308.05224. doi.org/10.48550/arXiv.2308.05224.

Schiff, D.; Biddle, J.; Borenstein, J.; and Laas, K. 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 153–158. New York, NY: Association for Computing Machinery. doi.org/10.1145/3375627.3375804.

Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; and Lennon, M. 2021. Explaining the Principles to Practices Gap in AI. IEEE Technology and Society Magazine 40(1): 81–94. doi.org/10.1109/MTS.2021.3056286.

Schnackenberg, A. K.; and Tomlinson, E. C. 2014. Organizational Transparency: A New Perspective on Managing Trust in Organization–Stakeholder Relationships. Journal of Management 42(7): 1784–1810. doi.org/10.1177/0149206314525202.

Scott, W. R. 2014. Institutions and Organizations: Ideas, Interests, and Identities. 4th ed. Thousand Oaks, CA: Sage Publications.

Sheldon, T. A. 2005. Making Evidence Synthesis More Useful for Management and Policy-Making. Journal of Health Services Research & Policy 10(1_suppl): 1–5. doi.org/10.1258/1355819054308521.

Symes, A. 1999. Review of Creating Public Value: Strategic Management in Government by Mark Moore. International Public Management Journal 2(1): 158–167.doi.org/10.1016/S1096-7494(00)87438-3.

Taeihagh, A. 2021. Governance of artificial intelligence. Policy and Society. 40(2): 137–157. https://doi.org/10.1080/14494035.2021.1928377

Tangi, L., C. van Noordt, and A. P. Rodriguez Müller. 2023. The challenges of AI implementation in the public sector: An in-depth case studies analysis. In Proceedings of the 24th Annual International Conference on Digital Government Research (DGO '23). ACM, New York, NY, 414–422. doi.org/10.1145/3598469.3598516.

Tolbert, P. S.; and Zucker, L. G. 1996. Institutionalization of Institutional Theory. In Handbook of Organization Studies, edited by S. Clegg, C. Hardy, and W. Nord. London, UK: Sage Publications.

Tsourma, M., N. Carmeno, J. Codagnone, S. Mancini, J. Krognos, A. Drosou, and D. Tzovaras. 2023. User experience of a web-based platform that enables ethical assessment of artificial intelligence in the public sector. In Human Interaction & Emerging Technologies (IHIET 2023): Artificial Intelligence & Future Applications. doi.org/10.54941/ahfe1004047.

Veale, M., M. Van Kleek, and R. Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, Paper 440, 1–14. doi.org/10.1145/3173574.3174014

Verbeek, P.-P. 2006. Materializing Morality: Design Ethics and Technological Mediation. Science, Technology, & Human Values 31(3): 361–380. https://doi.org/10.1177/0162243905285847

Winner, L. 1980. Do Artifacts Have Politics? Daedalus 109(1): 121–136. http://www.jstor.org/stable/20024652.

Wong, G.; Greenhalgh, T.; Westhorp, G.; Buckingham, J.; and Pawson, R. 2013. RAMESES Publication Standards: Realist Syntheses. BMC Medicine 11(1): 21. doi.org/10.1186/1741-7015-11-21.

Wong, J.; Morgan, D.; Straub, V. J.; Hashem, Y.; and Bright, J. 2025. Key Challenges for the Participatory Governance of AI in Public Administration. In Handbook on Governance and Data Science, 179–197. Cheltenham, UK: Edward Elgar Publishing. doi.org/10.4337/9781035301348.00017.

Zuiderwijk, A.; Chen, Y.-C.; and Salem, F. 2021. Implications of the Use of Artificial Intelligence in Public Governance: A Systematic Literature Review and a Research Agenda. Government Information Quarterly 38(1): 101577. doi.org/10.1016/j.giq.2021.101577.