Resolving value conflicts in public AI governance: A procedural justice framework

Citation for the original published paper (version of record):
de Fine Licht, K. (2025). Resolving value conflicts in public AI governance: A procedural justice framework. Government Information Quarterly, 42(2). http://dx.doi.org/10.1016/j.giq.2025.102033

Government Information Quarterly 42 (2025) 102033

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Resolving value conflicts in public AI governance: A procedural justice framework

Karl de Fine Licht *

Chalmers University of Technology, Technology Management and Economics, SE-412 96 Gothenburg, Sweden

ARTICLE INFO

Keywords:
Artificial intelligence
Trustworthy AI
Value conflicts
Public decision-making

ABSTRACT

This paper addresses the challenge of resolving value conflicts in the public governance of artificial intelligence (AI). While existing AI ethics and regulatory frameworks emphasize a range of normative criteria—such as accuracy, transparency, fairness, and accountability—many of these values are in tension and, in some cases, incommensurable. I propose a procedural justice framework that distinguishes between conflicts among derivative trustworthiness criteria and those involving fundamental democratic values. For the former, I apply analytical tools such as the Dominance Principle, Supervaluationism, and Maximality to eliminate clearly inferior alternatives. For the latter, I argue that justifiable decision-making requires procedurally fair deliberation grounded in widely endorsed principles such as publicity, inclusion, relevance, and appeal. I demonstrate the applicability of this framework through an indepth analysis of an AI-based decision support system used by the Swedish Public Employment Service (PES), showing how institutional decision-makers can navigate complex trade-offs between efficiency, explainability, and legality. The framework provides public institutions with a structured method for addressing normative conflicts in AI implementation, moving beyond technical optimization toward democratically legitimate governance.

1. Introduction

In recent years, there has been a significant debate surrounding the use of artificial intelligence (AI) in public decision-making. Examples include its application in areas like social welfare, medicine, criminal justice, and employment services (see e.g., Eubanks, 2018, Kaur et al., 2022, Berman et al., 2024). It is increasingly agreed that AI in public decision-making should be trustworthy (Reinhardt, 2023; Zanotti et al., 2023). This trustworthiness encompasses aspects such as accuracy, robustness, transparency, accountability, and fairness (Kaur et al., 2022; Reinhardt, 2023; Zanotti et al., 2023). These principles are not only integral to the ethical guidelines adopted by the private sector but also to legislation, like the European Union's AI act. The preamble was the suggestion from EU's High-Level Expert Group on AI, which published the 'Ethics Guidelines for Trustworthy Artificial Intelligence' (HLEG, 2019). This document has since become a foundational reference in studies exploring the practical implementation of trustworthiness in AI. Similarly, the EU AI Act, which came into force in August 2024, represents a legally binding framework for AI governance (European Parliament and Council, 2024).

Despite broad agreement on the need for trustworthy AI, debates persist over how to prioritize *competing values*. Conflicts frequently arise between accuracy and transparency, particularly in high-performing AI models like neural networks, which often trade interpretability for predictive accuracy (see e.g., Berman et al. 2024). This lack of transparency makes it difficult for public officials and affected individuals to understand or challenge decisions. While future solutions—such as explainable AI (XAI) or hybrid decision-making models—may help bridge this gap, they remain underdeveloped or impractical in public-sector decision-making due to their complexity, limited technical expertise, and institutional constraints. A similar challenge emerges in AI-driven legal decision-making in cases of e.g., racial bias (Angwin et al., 2022). The COMPAS algorithm, used in sentencing, illustrates a fundamental fairness conflict: ensuring equal treatment would require identical false positive and false negative rates across racial groups, while prioritizing predictive accuracy could justify different rates if they reflect real disparities in historical data. This trade-off is not just technical but deeply normative, as resolving it inevitably privileges one ethical principle over another (see e.g., Loi et al., 2023). Despite extensive discussion, there is still no clear framework for recognizing,

* Corresponding author at: Chalmers University of Technology, Address: Vera Sandbergs Allé 8, 411 33 Gothenburg, Sweden.
E-mail address: karl.definelicht@chalmers.se.

https://doi.org/10.1016/j.giq.2025.102033
Received 19 March 2024; Received in revised form 3 May 2025; Accepted 5 May 2025
Available online 13 May 2025

analyzing, and balancing these conflicts in trustworthy AI (see e.g., Petersen, 2021; Reinhardt, 2023; Ryberg & Petersen, 2022).

This paper aims to utilize philosophical theorizing to understand conflicts in AI usage, discerning when such conflicts arise and how to address them. I propose a framework for analyzing what constitutes truly trustworthy AI. Trustworthiness in AI extends beyond embedding the right values; it also involves recognizing value conflicts and responding appropriately. I argue that there are fundamental values or principles essential for trustworthy AI in public decision-making, but their interpretation and content may vary based on the *specific context* of application. Additionally, when unsolvable clashes between values or principles occur, they should be resolved through pure procedural justice procedures. Thus, rather than assuming that AI systems must meet all trustworthiness conditions simultaneously, this paper argues that AI governance is a balancing act in which decision-makers must systematically weigh competing values. This structured prioritization ensures that AI implementations remain justifiable within democratic institutions rather than being assessed through a rigid pass/fail evaluation. To demonstrate the practical applicability of the framework, I will utilize an empirical case where an AI system is used in the Swedish Public Employment Service to determine whether applicants should get access to some of their services such as educational programs etcetera.

The structure of the paper is as follows: I begin by constructing a framework for identifying and addressing genuine conflicts concerning AI in public decision-making. I then apply this framework to the case of the Swedish Public Employment Agency (PES) to demonstrate its practical use and to provide concrete examples. Following this, I discuss the implications of the findings, leading to a conclusion and outlook on future research and application possibilities.

2. Theory

In this section, I develop a framework for analyzing and resolving value conflicts in trustworthy AI governance. I begin by identifying types of value conflicts that arise in AI systems, distinguishing between conflicts among non-fundamental (derivative) values and those involving fundamental values. I then present approaches for resolving each type of conflict: analytical methods for non-fundamental conflicts and procedural justice mechanisms for fundamental ones. The framework integrates these elements into a practical decision-making structure that public institutions can implement across various contexts and decision levels.

2.1. Identifying and resolving nonfundamental value conflicts

There is a broad spectrum of values and principles associated with "trustworthy AI." These encompass performance, calibration, interpretability, explainability, intelligibility, fairness, legality, and accountability, among others (Kaur et al., 2022, Reinhardt, 2023, Zanotti et al., 2023, Berman et al. 2024).

Performance is assessed through the accuracy of the AI system in making judgments or decisions at all levels, the enhancement of human decision-makers accuracy due to the AI system, and the communication of the system's performance to stakeholders. Calibration addresses the confidence estimates to stakeholders and the accuracy of these estimates, ensuring they are well-calibrated with the system's actual performance. The combined principles of interpretability and explainability concern themselves with whether the AI system's decision-making logic can, in principle, be understood by stakeholders and whether the explanations provided are faithful to the actual decision-making process. Lastly, intelligibility and availability focus on making the decision-making logic accessible and comprehensible to various stakeholders, ensuring that explanations are not just theoretically available but practically understandable as well.

Equal and fair treatment should according to many people in the debate, at minimum, include what is referred to in the debate on fair and equitable AI as "predictive fairness." Predictive fairness refers to the ethical ideal that algorithmic systems making probability predictions about individuals should distribute errors and accuracy equally across different demographic groups (Loi et al., 2023). Legality, accountability, appeal, and human oversight pertain to the system built around the AI or can be seen as part of the socio-technical AI system. Trustworthy AI must abide by the law, ensure that someone is accountable for decisions made as a result of the AI system, provide a mechanism for appealing decisions, and include a human agent "in the loop" to oversee these processes.

That these conditions clash has been demonstrated in case studies (Berman et al., 2024), proven mathematically in the context of predictive fairness (Loi et al., 2023), and is theoretically defensible based on the literature on incommensurability, which suggests that incommensurability is pervasive (Raz, 1986, Anderson, 1995; Daniels, 2007). For example, it is widely recognized that accuracy and transparency can conflict in AI systems (Kaur et al., 2022). Increasing transparency may require simplifying or disclosing aspects of an algorithm, which can reduce its predictive accuracy. Conversely, optimizing for accuracy often involves complex, opaque models that hinder transparency and interpretability. Another example is the tension between transparency and security, where enhancing AI transparency by disclosing system operations directly conflicts with security goals aimed at preventing adversarial attacks.

With this being said, value conflicts in trustworthy AI governance arise in two primary forms. First, conflicts occur within AI trust conditions, and second, conflicts arise between AI trust conditions and values external to them or more fundamental than them. The framework outlined in this paper addresses both types of conflicts. In determining clashes between values or principles, it is crucial to first precisely define concepts and understand the core values underpinning these principles (cf. Ryberg & Petersen, 2022). By examining (i) precise definitions of trustworthy AI and (ii) fundamental values, we may uncover previously unnoticed conflicts or discover that apparent conflicts are not morally significant. Thus, the first two steps in our framework for handling conflicts are, first, to examine whether a conflict exists by clearly defining our core terms, and second, to determine whether the conflict is genuine by identifying the more fundamental values underlying it. Concerning (ii), if underlying values justify the principles of trustworthy AI, resolving conflicts may not compromise these values.

Fundamental or external norms and values play a crucial role in how we act when the conditions of trustworthy AI are in conflict. To develop a framework that public agencies can follow—one that is neither too rigid nor too lenient—it is useful to consider what Rawls famously refers to as a "realistic utopia" (Rawls, 1999). This approach involves setting aspirational yet feasible goals by identifying both the ideals a society should strive for and the practical constraints within which these goals must be pursued. This implies that the fundamental values underpinning the framework should be those that most people share or would endorse upon reflection, as well as those that decision-makers in the state are obligated to uphold. The core principles of the democratic state—such as constitutional commitments and deeply ingrained bureaucratic norms—serve as a foundation. Examples include prioritizing the worst-off, ensuring transparency, promoting efficiency, upholding legality, and maintaining accountability.

Thus, within the context of trustworthy AI, some values can be considered foundational, while others may be more contingent or context-dependent. Identifying core values and ensuring their protection is crucial for developing an AI governance framework that is both principled and practical. Furthermore, some conditions of trustworthy AI reflect core or fundamental values. For example, according to most theories of rule legitimacy, governance must be transparent—at least in the sense that high-stakes decisions must be justified to the public. Additionally, when such decisions have the potential to harm individuals or groups, there must be mechanisms for appeal. This suggests that certain conditions of trustworthy AI, such as these norms or values,

cannot be reduced to more fundamental principles.

When derivative or non-fundamental values clash, we can use analytical methods or principles to resolve these conflicts while preserving the underlying fundamental values they serve. Analytical principles can help eliminate clearly suboptimal options, narrowing the field of consideration to those alternatives that best respect our underlying commitments. The principles we are going to discuss here are the Dominance Principle, Supervaluationism, and the Maximality Principle well known from philosophical discussions on value incommensurability and decision theory (Andersson, 2017; Chang, 2002; Fine, 1975; Herlitz, 2019; Sen, 1997).

To reduce complexity and improve decision efficiency, we first apply the Dominance Principle, eliminating AI systems clearly inferior to others—specifically, those performing worse on at least one dimension of trustworthiness without compensating improvements elsewhere (Savage 1951, Sen, 1970, Broome, 1991). This step simplifies decision-making by focusing only on non-dominated alternatives, ensuring that we do not expend effort comparing options that are objectively suboptimal. For instance, if one AI system consistently shows higher accuracy without compromising fairness compared to another, the inferior system can be excluded.

Next, building upon our earlier emphasis on precise definitions, we apply supervaluationism when nondeterminacy arises from vague evaluative criteria such as "fairness" or "transparency" (Andersson, 2017; Fine, 1975). This formal approach systematically implements the definitional clarity we identified as crucial in step (i), ensuring that decision-making remains coherent by considering all admissible precisifications (sharpenings) of vague concepts. This approach prevents arbitrary or inconsistent evaluations while maintaining flexibility in ethical reasoning. An AI system can be rationally eliminated if it performs worse across all plausible interpretations.

Lastly, Sen's conception of rationality based on "maximality" rather than optimization helps eliminate remaining clearly suboptimal AI systems in nondeterminate contexts. Since full comparability among AI systems is often impossible due to value incommensurability, maximality provides a rational decision rule that avoids the need for a complete ranking. According to Sen, a rational choice needs only be no worse than any alternative, rather than demonstrably optimal. This allows for structured decision-making even when optimization is infeasible. Thus, even without complete rankings, AI systems determinately inferior to at least one alternative can still be eliminated. For example, if AI system C consistently demonstrates poorer explainability compared to AI system B, it can be rationally excluded, even if neither AI system A nor B can be definitively ranked. Applying maximality at this stage ensures that only defensible, non-dominated choices remain, preserving ethical pluralism while maintaining decision feasibility. See Table 1 for an overview.

Table 1
Decision-making principles for narrowing AI system alternatives in cases of value conflict and nondeterminacy.

Principle	Description	Example Application
Dominance Principle	Eliminate AI systems clearly inferior to others, specifically those worse on one dimension without improvement elsewhere.	An AI system consistently shows lower accuracy without compensating fairness improvements compared to another.
Supervaluationism	Clarifies vague evaluative criteria by eliminating options worse across all plausible precisifications.	Eliminating AI algorithms consistently less transparent across all plausible definitions of transparency.
Sen's Maximality Criterion	Eliminates clearly suboptimal AI systems that are determinately inferior to at least one alternative, without requiring a full ranking.	Excluding an AI system that consistently demonstrates poorer explainability compared to another system, even if full rankings are unclear.

2.2. Resolving fundamental value conflicts

When derivative values clash, we can use non-derivative values or principles to adjudicate the conflict. However, when *non-derivative values clash*, we face a case of value incommensurability. According to a widely accepted definition of value incommensurability, two items, x and y, are incommensurable if it is not determinately true that x is better than y, y is better than x, or that x and y are equally good (e.g., Herlitz, 2024). This condition arises when the inability of a normative framework to fully determine a ranking of values prevents an *at least as good as* relation from holding determinately between all pairs of items (Broome, 2022; Chang, 2022). This challenge is not unique to AI governance—incommensurability is a pervasive phenomenon across ethical domains (Raz 1986; Anderson 1993; Daniels 2008), and independent criteria often fail to determine a single best alternative (Broome 2004; Chang, 2002; Rabinowicz, 2008).

Value incommensurability is thus a core challenge in trustworthy AI, where principles such as transparency, fairness, accountability, robustness, privacy, and human oversight lack a clear ranking or definitive method for resolving trade-offs. This complexity increases when considering broader external and fundamental values beyond AI-specific criteria. While elimination methods help maintain rigor in decision-making by removing irrational alternatives, they have inherent limitations as a complete solution as Herlitz (2019, 2020) and Herlitz and Sadek (2021) argue. This initial elimination alone cannot satisfy the *deeper justificatory demands* in high-stakes decisions, and this is especially true according to us, when involving trustworthy AI in public decision making. Even after irrational options have been excluded, multiple eligible alternatives often remain, requiring more substantial justification than random selection can provide. As emphasized by Herlitz and Sadek, arbitrary selection fails to address genuine disagreements and the need for justifications acceptable to stakeholders affected by significant decisions. Andreou (2016) also emphasizes that random selection among non-rankable options can lead to suboptimal outcomes over a sequence of decisions, reinforcing the inadequacy of purely arbitrary solutions.

To address these challenges, Herlitz and Sadek propose a hybrid procedural approach combining deliberative and aggregative mechanisms. This model starts with deliberation, allowing stakeholders to articulate and engage with various perspectives, thereby generating substantive reasons to support specific alternatives. Following deliberation, aggregative methods such as voting finalize the choice. This hybrid approach recognizes stakeholders not merely as preference-holders but as reasoning participants whose viewpoints deserve meaningful engagement. This emphasis on procedures aligns with Rawls's concept of pure procedural justice (1971: 73–78), which asserts that a just outcome depends primarily on the fairness of the process leading to it. Thus, the focus shifts from outcomes themselves to the quality of the procedures that produce them. Since justice and trustworthiness are closely related, adopting a just procedure in AI development is likely to yield trustworthy AI systems, provided technical aspects function correctly. Many contemporary philosophers and political theorists advocate procedural solutions as the preferred means of addressing genuine value conflicts (Anderson, 1999; Andersson & Herlitz, 2022; Chang, 2002; Daniels & Sabin, 2002; Nussbaum, 2011; Pettit, 2012; Tyler, 2006). Herlitz (2024) further argues that the inherent nondeterminacy of value conflicts, rather than being problematic, actually underscores the necessity of procedural solutions. Given the broad applicability of procedural approaches across diverse public institutions, this paper focuses on exploring these solutions.

Herlitz and Sadek don't provide a detailed view on what conditions the deliberation should fulfill, but many accounts of procedural justice can serve as frameworks to mediate between differing values or normative principles. Despite the broad spectrum of procedural justice

theories, there is a notable similarity in their core content, albeit with divergent perspectives on what precisely procedural justice entails. I am going to pick the most uncontroversial criteria here, with the greatest overlap between theories and institutional practices, such that the framework becomes as feasible as possible while still theoretically grounded.

First is the *publicity condition*, as highlighted by Daniels and Sabin (2002), Pettit (2012), Fraser (2009), and Nussbaum (2011), among others. This principal mandates that any process leading to a decision must ensure that the decisions themselves, along with their rationales, are accessible to the public. Such transparency guarantees that all stakeholders, including the general public, have the opportunity to examine the decisions and the logic behind them.

Second is the *relevance condition* (see, e.g., Habermas, 1985; Daniels & Sabin, 2002; Brandstedt & Brülde, 2019). The reasons and rationales guiding decision-making processes must be pertinent and grounded in evidence, principles, and justifications that all parties have reasons to accept (see, e.g., Habermas, 1985, on the universalization principle; Scanlon, 2000, on the principle of reasonable rejection). When it comes to normative principles and values, reasons should be formulated in alignment with reasonable normative theories or ethical considerations, which is what should be understood as "relevant reasons" (see, e.g., Brandstedt & Brülde, 2019). Typically, this involves providing reasons that e.g., promote the common good rather than merely advancing self-interest (Pettit, 2012). Stakeholders may also adjust the weighting of already considered reasons based on their lived experiences (e.g., Herlitz, 2024). For example, if a group has experienced oppression, their testimony might shift the emphasis toward their needs, even if they are not introducing a fundamentally new argument. Instead, their contribution adds context and urgency to existing considerations, influencing the prioritization of certain values or principles.

Third, the *inclusion condition*, as articulated by Anderson (1999), Fraser (2008), Allen (2004), Pettit (2012) on participatory inclusiveness, and Srinivasan (2021), emphasizes the critical importance of engaging all stakeholders in the deliberation process. This principle aims to ensure that the decision-making process is not only transparent and relevant but also representative. Identifying who precisely constitutes 'all stakeholders' is a nuanced endeavor. Ideally, it encompasses not just those directly impacted by the decision but also individuals and groups whose interests might be indirectly affected. Moreover, the inclusion condition mandates a proactive approach to identifying and engaging marginalized or traditionally excluded groups. When it comes to public institutions and public decision-making, we need to strike a balance among those who are directly affected by these institutions, i.e., the recipients; those who contribute to them but are not directly affected; and those who are employed by the institutions.

Fourth is the condition of *fair terms of cooperation and a cooperative spirit* (cf., e.g., Habermas 1985 on the ideal speech situation, Anderson, 1999 on equal opportunity and the empowerment of individuals, Pettit, 2012 on non-domination, Fraser 2008 on participatory parity, Nussbaum, 2011, Brandstedt & Brülde, 2019, Srinivasan, 2021). This cooperative spirit is characterized by mutual respect, reciprocity, and a willingness to seek common ground. It also entails a commitment to advocate for equitable terms and to adhere to them, provided others do likewise. Furthermore, for the conditions to be fair, we probably need to support individuals who are less accustomed to the type of reasoning described here, where one cannot merely argue for something because one believes it to be true or because it is personally beneficial. We will likely also need to educate people about AI: its capabilities and limitations, the alternatives available within different AI technologies such as classic rule-based AI or machine learning, and the advantages and disadvantages of these technologies.

Fifth and sixth are the principles of *appeal and revision*, and *regulation and enforcement*. As emphasized by Allen (2004), Nussbaum (2011), Tyler (2006), and Pettit (2012) among others: There must exist a robust mechanism for challenge and dispute resolution concerning the decisions made. This framework should facilitate the revising and appealing of decisions in the wake of new evidence or compelling arguments. Such a criterion is crucial for ensuring that the decision-making process remains dynamic, adaptable, and responsive to evolving insights and circumstances, thereby allowing for necessary corrections and adjustments over time. As our values change, this also allows public institutions to change with them.

There must also be some form of regulation or enforcement mechanism to ensure the aforementioned criteria are met. This involves oversight by a body or mechanism capable of holding the decision-making process accountable to its standards, ensuring that the process remains fair, transparent, and consistent with the stated principles. This creates a framework of mutual expectations and accountability that enhances the overall fairness and effectiveness of the negotiation process. Incidentally, utilizing these norms has the greatest chance of success in terms of legitimacy and fairness perceptions among the public (see e.g., Tyler, 2006, de Fine Licht and de Fine Licht 2020). For an overview of the principles, see Table 2 below.

2.3. Framework for decision-making in public AI governance

Building upon the theoretical foundations established in previous sections, I propose a structured framework for public institutions to navigate the complex value trade-offs inherent in trustworthy AI governance. (See Table 3 for an overview.) This framework combines analytical approaches for identifying and resolving non-fundamental conflicts with procedural mechanisms for addressing fundamental value incommensurabilities. It is designed to be both normatively robust and practically applicable within public administration contexts.

The framework consists of three distinct phases, each with specific steps for implementation. Phase 1 focuses on conflict *identification and classification*. In Step 1 of this initial phase, institutions must define and clarify values by precisely articulating the relevant trustworthy AI principles such as performance, calibration, explainability, and fairness. Each principle should be operationalized with concrete metrics and evaluation criteria, and the specific manifestation of potential conflicts in the given context must be thoroughly documented. Moving to Step 2, institutions determine the conflict type by analyzing whether conflicts occur within trustworthy AI conditions (such as accuracy versus explainability), between trustworthy AI conditions and external values

Table 2
Procedural Justice Principles for Resolving AI-Related Fundamental Value Conflicts.

No.	Principle	Description
1	Publicity Condition	Ensures decision-making processes and outcomes are open and transparent to all stakeholders, fostering trust through openness.
2	Relevance Condition	Decisions are based on relevant and universally acceptable reasons and evidence, aligning with the common good over individual self-interests.
3	Inclusion Condition	Highlights the need for inclusivity and participatory engagement in the decision-making process, especially for marginalized and excluded groups.
4	Fair Terms of Cooperation and Cooperative Spirit	Promotes a collaborative environment with mutual respect and reciprocity, advocating for equitable engagement terms and educating participants about AI.
5	Principles of Appeal and Revision	Establishes mechanisms for challenging, revising, and appealing decisions, ensuring adaptability and responsiveness to new insights.
6	Regulation and Enforcement	Requires regulatory oversight to maintain fairness, transparency, and accountability in decision-making, implementing checks and balances for adherence to justice principles.

(such as transparency versus security), or between derivative or fundamental values. This step also involves identifying the underlying fundamental values at stake in each conflict.

Phase 2 addresses the *resolution of non-fundamental conflicts*. During Step 3 of this phase, institutions apply elimination principles to narrow down the range of acceptable AI systems. The Dominance Principle eliminates AI systems clearly inferior on at least one dimension without compensating improvements elsewhere. Supervaluationism eliminates options that perform worse across all plausible precisifications of vague concepts. The Maximality Criterion eliminates any AI system determinately inferior to at least one alternative. Proceeding to Step 4, technical optimization explores solutions that might mitigate apparent trade-offs. Institutions investigate whether systems can be modified to better satisfy multiple values simultaneously and document unavoidable fundamental conflicts that remain after technical optimization attempts.

Phase 3 focuses on the *procedural resolution of fundamental conflicts* that could not be resolved through earlier phases. Step 5 implements structured deliberation applying the principles of procedural justice. Publicity ensures transparency of the decision-making process and rationales by publishing detailed documentation of the AI system, its intended use, and potential impacts; disclosing evaluation criteria, trade-offs, and decision frameworks; and making deliberation processes accessible to stakeholders and the public. The relevance principle ensures focus on evidence and reasons acceptable to all reasonable stakeholders by requiring that arguments reference commonly accepted values or constitutional principles, prioritizing evidence-based reasoning over purely self-interested claims, and structuring deliberation around impact assessments and risk evaluations. The inclusion principle ensures participation of all affected stakeholders by identifying and engaging both directly and indirectly affected groups, implementing mechanisms to include traditionally marginalized voices, and balancing representation among system users, contributors, and administrators. Fair cooperation establishes equitable deliberative conditions by providing education about AI technologies and their implications, supporting stakeholders less familiar with technical or deliberative reasoning, and ensuring balanced speaking opportunities and influence.

In Step 6, following structured deliberation, decision aggregation and implementation applies appropriate voting or preference aggregation mechanisms, documents the decision with clear rationales for chosen trade-offs, and develops implementation plans with specified monitoring and evaluation processes. Step 7 establishes appeal and revision mechanisms for challenging AI-related decisions, creates regular review cycles to incorporate new evidence or changing values, and documents how feedback is incorporated into system improvements. Finally, Step 8 develops regulation and enforcement systems that determine oversight bodies responsible for monitoring compliance, establish clear enforcement mechanisms and consequences, and implement auditing protocols and accountability structures.

This framework adapts to various public sector contexts through three implementation modes. The Strategic Mode is designed for high-level policy decisions about AI adoption and governance, emphasizing extensive stakeholder engagement and thorough value analysis. It typically involves elected officials, senior administrators, and public consultation, resulting in policy frameworks and governance structures. The Tactical Mode addresses organizational decisions about specific AI systems, focusing on applying the framework within existing policy constraints. Typically led by department heads and program managers, this mode results in procurement specifications and operational guidelines. The Operational Mode handles day-to-day administration of AI systems, addressing emergent conflicts and routine trade-offs. Implemented by frontline managers and technical staff, this mode results in system adjustments and procedural adaptations.

The framework's staged approach allows public administrators to address value conflicts systematically while maintaining trustworthiness through procedural justice. By combining rational elimination criteria with inclusive deliberative processes, it offers a practical path through the complex landscape of value incommensurability in AI governance, aligning with both democratic principles and administrative pragmatism. Each phase should also include documentation requirements to ensure transparency and accountability, creating an "audit trail" of decisions that can be reviewed by oversight bodies, stakeholders, and the public. This documentation supports organizational learning and helps build institutional knowledge about effective AI governance over time.

Finally, for the governance body or auditors to "know" or effectively apply the philosophically grounded principles in Tables 1 and 2, institutional learning mechanisms must be established. This includes (1) formal integration of the principles into training programs, evaluation guidelines, and governance protocols; (2) iterative procedural application, where repeated involvement in structured decision-making fosters internalization of normative reasoning; and (3) reflexive auditing, wherein principles serve as criteria in post-decision evaluations, supporting learning through systematic reflection. Together, these mechanisms enable decision-makers not only to reference but also to practically apply and refine their understanding of abstract normative principles in concrete contexts. To support this, the ethical audit trail should explicitly document the governance team's articulation of core values deemed foundational—those not subject to compromise—as well as any value conflicts encountered and the justification for their resolution. This enhances procedural transparency and facilitates learning across governance cycles. (See Table 3.)

Table 3
Three-Phase Framework for Resolving Value Conflicts in Public AI Governance.

Phase	Step	Description
1: Conflict Identification and Classification	Step 1: Define and Clarify Values	Identify and articulate relevant trustworthy AI principles, operationalize with metrics and evaluation criteria, document potential conflicts
	Step 2: Determine Conflict Type	Analyze conflicts within trustworthy AI conditions, between AI conditions and external values, or between fundamental and derivative values
2: Resolution of Non-Fundamental Conflicts	Step 3: Apply Elimination Principles	Apply structured decision-making principles to eliminate clearly inferior AI systems
	Step 4: Technical Optimization	Explore system modifications to mitigate trade-offs between values, optimize AI solutions, document unresolved fundamental conflicts
3: Procedural Resolution of Fundamental Conflicts	Step 5: Structured Deliberation	Follow procedural justice principles: Publicity, Relevance, Inclusion, Fair Cooperation
	Step 6: Decision Aggregation and Implementation	Select decision-making mechanisms, document rationale behind trade-offs, include monitoring and evaluation processes
	Step 7: Appeal and Revision Mechanisms	Include mechanisms for challenging decisions, conducting periodic reviews integrating feedback to refine AI systems
	Regulation and Enforcement	Establish oversight bodies, define enforcement mechanisms, implement auditing protocols, create accountability structures

3. Application of the framework: the Swedish public employment service

3.1. Background

The Swedish Public Employment Service (PES) implemented an AI-driven decision support system, known as the BÄR tool, as part of its "Prepare and Match" initiative. This system was designed to classify jobseekers and determine appropriate support levels based on their proximity to the job market.

As mandated by the government, the PES deployed this statistical profiling tool with the dual aims of enhancing consistency in labor market assessments and improving resource allocation efficiency. The system uses a neural network trained on historical data to estimate employment probabilities, categorizing jobseekers into three groups: those too near the job market (requiring minimal support), those suitable for the Prepare and Match program, and those too far from the job market (requiring more intensive support). While caseworkers formally make the final decisions, they are instructed to primarily adhere to the automated recommendations, with limited discretion to override the system.

This implementation represents a compelling case study not primarily for evaluating technical trustworthiness (for technical details, see Berman et al., 2024), but rather for analyzing how value conflicts emerge and evolve within public sector AI governance. The PES case exemplifies the tension between competing normative commitments faced by democratic institutions when implementing AI systems. Unlike traditional evaluations that assess AI systems against a fixed set of criteria, my analysis focuses on the dynamic interplay between values such as efficiency, transparency, professional discretion, and democratic participation. These tensions manifest at multiple governance levels—from strategic decisions about system design to operational choices in individual cases—illustrating the need for a structured approach to value conflict resolution that extends beyond technical optimization. For an overview, see Table 5 below.

3.2. Application

In Phase 1 we identify and classify value conflict where the first step is to define and clarify values. The relevant values in this case include the trustworthy AI Principles performance (accuracy in predicting employment outcomes), calibration (reliability of confidence estimates), explainability (understanding of decision logic), intelligibility (comprehensibility of explanations), and fairness (consistent treatment across demographic groups).

The core governmental values as outlined in 'The State's Core Values - Common Principles for Good Administration' (Statskontoret, 2019) encompass six principles of particular importance. Since these values partially overlap, we focus on their unique aspects while acknowledging shared elements.

Democracy emphasizes that authorities operate on behalf of citizens and actively promote democratic values throughout society (Statskontoret, 2019: 8, 10), defined by basic freedoms of opinion, expression, and religion. Legality requires that public authorities' activities have solid backing within the legal framework (Statskontoret 2019: 12). Decisions concerning individuals must be motivated with clear reference to applicable rules, as evidenced by cases where the Supreme Administrative Court of Sweden has annulled decisions exceeding legal mandates. Decision-makers must provide justifications in either contrastive or non-contrastive forms. Objectivity requires authorities and employees to act factually and impartially, ensuring decisions are based on merits rather than personal preferences or biases, while promoting consistent treatment across similar cases.

Free formation of opinion protects employees' rights to freedom of speech, information, assembly, and religion, with whistleblower protections against reprisals (Statskontoret 2019: 22). It ensures public access to examine authorities' actions and fosters transparency. Respect mandates public sector employees to serve with deep regard for individuals, fulfilling requirements for non-discrimination and personal integrity (Statskontoret 2019: 27). This extends to GDPR compliance, ensuring personal data is handled transparently and lawfully (Statskontoret 2019: 28f). Efficiency and service requires authorities to operate cost-effectively with minimal processing times while maintaining quality (Statskontoret 2019: 30). Authorities must be accessible, provide clear information, make expedient decisions, and use plain language (klarspråk) in communications (Statskontoret 2019: 33). See Table 4 for an overview.

There was also a PES-Specific Value from government directive to PES (Arbetsförmedlingen, 2023) that stated that focus on those furthest from the labor market (supporting the worst-off) should be of priority.

In Step 2, Phase 1, we should determine the conflict type, and here we find multiple conflicts of which we will discuss. First, there are conflicts *within* Trustworthy AI conditions, such as Performance vs. Explainability: The neural network (68 % accuracy) outperformed a more interpretable linear regressor (66 % accuracy), presenting a trade-off between prediction quality and explainability. Now, even though the advantage of the neural network seems small, there is still a trade-off to be made, and for this, we need a separate criterion to decide what to do. Furthermore, 2 % might also be seen as substantial since there are hundreds of thousands of decisions being made each year utilizing this system. Hence, even a small increase in accuracy could lead to significant impact for many individuals.

There is also a conflict within fairness, as mathematical impossibility theorems show that different fairness metrics cannot be simultaneously satisfied when base rates differ across groups (Kleinberg et al., 2016; Loi et al., 2023). Since reemployment rates vary—e.g., immigrants, people with disabilities, and low-income residents face lower prospects than native-born, non-disabled individuals in affluent areas—it is impossible to meet all fairness criteria at once (Kleinberg et al., 2016, Loi et al., 2023). This creates a trade-off between fairness definitions, such as equal false positive and false negative rates, requiring a choice between them.

Second, there are conflicts between Trustworthy AI Conditions and Core Governmental Values where efficiency (AI accuracy) stands against legality (decision justification): The neural network improved efficiency but its "black box" nature complicated legal requirements for justified decisions. There is also a conflict between Performance vs. Democracy: Higher accuracy came at the cost of limited transparency, potentially

Table 4
A description of The Governmental Core Values in a non-hierarchical order.

No.	Principle	Description
1	Democracy	Authorities work on behalf of the citizens, and their employees are tasked with remembering that and promoting democratic values.
2	Legality	The activities of the authorities must be supported by the legal system, employees should be familiar with and follow the rules applicable to their authority's operations, and they must justify their decisions with the ground in specific rules, and they must justify their decisions in the individual case.
3	Objectivity	Authorities and their employees must act factually, impartially, and consistently.
4	Free Formation of Opinion	Employees have rights to freedom of speech, information, assembly, demonstration, association, religion, and protection against investigation and reprisals. The principle of public access means everyone has the right to scrutinize the authorities and their employees critically.
5	Respect	Employees should serve with respect for the individual, fulfilling requirements for non-discrimination and consideration of personal integrity, and act with humanity.
6	Efficiency and Service	Authorities should strive to provide good treatment, be accessible, and serve the citizens efficiently.

undermining respect for individual autonomy, which in turn is a democratic virtue. These conflicts are fundamental rather than merely apparent because they are all part of the core of what is required from Swedish institutions, requiring structured resolution approaches.

A third conflict arises between efficiency and legality or democracy. While the most efficient way to make decisions is to use an AI system, this creates a trade-off between caseworkers' ability to understand the basis of a given decision and the need for the process to be efficient and expedient. Utilizing AI systems in complex cases significantly speeds up decision-making, but these systems vary in their degree of opacity. As a result, it becomes more difficult for caseworkers to provide a genuine justification for the decisions made.

Moving to Phase 2, Step 3, we attempt to find a resolution of non-fundamental conflicts through applying elimination principles. According to the framework, we should first utilize the Dominance Principle. When comparing available AI systems, neither the neural network nor the linear regressor was strictly dominated, as each excelled on different dimensions (accuracy vs. explainability). Both systems significantly outperformed the random baseline (57 %) and traditional caseworker methods for predicting long-term unemployment (10 % accuracy).

For the efficiency vs. legality conflict, neither approach was strictly dominated either. The neural network provided greater efficiency but created challenges for legal justification, while more transparent approaches better satisfied legal requirements but potentially at the cost of efficiency. Similarly, when examining the performance vs. democracy tension, no clear dominance emerged, as increased performance came with decreased transparency and potential impacts on respect for individual autonomy.

Applying Supervaluationism, we analyzed different interpretations of "transparency." The neural network with LIME explanations satisfied minimal transparency requirements, though not ideal transparency. Under all precisifications of "accuracy," the neural network performed better than the linear regressor. However, when considering different precisifications of "legal justification," the linear regressor consistently performed better across all interpretations, as its decision-making process was inherently more explainable to caseworkers and citizens. Similarly, under varying interpretations of democratic values, more transparent models consistently aligned better with democratic principles of accountability and citizen understanding.

According to the Maximality Criterion, when considering solely accuracy vs. explainability, neither system was determinately inferior to the other across these dimensions. However, when expanding the analysis to include the fundamental values of legality and democracy, the neural network appeared determinately inferior to the linear regressor on these dimensions without sufficient compensating advantages in accuracy (the 2 % difference). This suggests that when considering the full range of governmental values, the linear regressor might be preferred according to Sen's maximality principle, though the decision is not clear-cut given the conflicting prioritizations of various criteria.

The fairness conflicts presented a different challenge. Due to the mathematical impossibility theorems, no system could satisfy all fairness criteria simultaneously when base rates differ across groups. Applying our elimination principles revealed that different fairness definitions created a situation where no approach dominated others across all fairness metrics, no approach was consistently better under all precisifications of "fairness," and no approach was determinately inferior to all others. This confirmed that fairness conflicts would require procedural resolution.

This takes us to Step 4 Phase 2 and technical optimization. PES attempted technical optimization through using LIME to provide post-hoc explanations for the neural network's decisions and also experimenting with alternative models, including a linear regressor (66 % accuracy) and a more sophisticated combination of a decision tree and 6 linear regressors (74 % accuracy). However, LIME explanations proved unstable and potentially misleading, as different explanations could be generated for the same prediction. The special treatment of unemployment duration (always listed first in explanations) further complicated matters, as its importance varies between cases. The simplest alternative model tested by PES achieved 66 % accuracy compared to the neural network's 68 %, while a more sophisticated interpretable model achieved 74 % accuracy, suggesting that simpler models could potentially fulfill stated goals of consistency and accuracy equally well or better than the more opaque neural network model.

However, given the recognized potential of neural networks to outperform linear regressors—a capability demonstrated in various contexts—and the Swedish government's directive for agencies to adopt AI solutions (Sweden, 2021), there is a question about whether linear regressors and simple decision trees can be strictly classified as AI. Therefore, even if advanced linear regressors prove more accurate than current neural networks, as found in one of the PES's own studies, there could still be compelling reasons to implement the less accurate neural network over the more accurate linear regressor.

Moving to Phase 3, we implement procedural resolution of fundamental conflicts that remain after technical optimization, such as tensions between different types of predictive fairness and conflicts between efficiency and transparency. At this stage, the Swedish PES must shift from analytical approaches to deliberative ones to address value conflicts that cannot be resolved through elimination or technical means alone.

At the strategic level, PES leadership should establish a formal AI Governance Committee with explicit responsibility for developing foundational policies governing AI use across employment services. This committee should include diverse stakeholder representation from multiple perspectives: jobseeker representatives reflecting different demographic groups (youth, older workers, immigrants, those with disabilities); caseworkers with varying levels of experience and from different regional offices; labor market experts from academia and industry; technical specialists with expertise in AI ethics and explainable AI; legal advisors specializing in administrative law and privacy; representatives from labor unions and employer organizations; and experts on public administration ethics. This diverse composition ensures that all relevant perspectives are represented in strategic deliberations.

The committee would implement a structured deliberation process featuring regular full-committee meetings with published agendas, working group sessions on specific issues like fairness metrics and transparency standards, public hearings to gather broader input, and an online consultation platform for continuous stakeholder input. All deliberations would be thoroughly documented with minutes published on e.g., the PES website, ensuring transparency and accountability throughout the process. Clear publicity mechanisms would support this work, including publication of all committee documentation, annual reports on AI governance decisions and their implementation, an interactive website explaining the AI system's purpose and limitations, and regular press briefings on major decisions and policy changes. These mechanisms ensure that the publicity condition of procedural justice is satisfied, making the decision-making process visible and accessible to all stakeholders.

The committee would tackle fundamental questions such as whether the PES should prioritize a system with slightly higher accuracy (neural network) or one with better explainability (interpretable model), how accuracy should be balanced across different demographic groups, and what level of confidence should trigger human review of AI recommendations. These deliberations must be governed by the procedural justice principles outlined earlier. For example, applying the relevance condition would require all arguments to reference core governmental values like democracy and legality, not merely efficiency or cost-effectiveness. The committee would document how different values were weighed in reaching its conclusions, creating a transparent rationale that could be scrutinized by oversight bodies.

At the tactical level, department heads would translate strategic

principles into practical implementations through cross-functional implementation teams comprising IT specialists, service managers, legal advisors, and caseworker representatives. These teams would maintain regular consultation with the AI Governance Committee to ensure alignment with strategic principles and hold quarterly review meetings to assess implementation challenges. Through this process, they would develop specific operational guidelines addressing when caseworkers can override AI recommendations, what explanation requirements should be satisfied for different stakeholder groups, and how performance metrics should balance accuracy, fairness, and explainability dimensions.

The tactical implementation would also establish robust feedback mechanisms including monthly reviews of cases where caseworkers overrode AI recommendations, quarterly analysis of demographic patterns in AI recommendations and outcomes, structured channels for caseworkers to report concerns about the AI system, and regular surveys of jobseekers regarding their experience with and understanding of AI-assisted decisions. These feedback loops ensure that implementation challenges inform ongoing policy refinement, creating a learning organization that continuously improves its AI governance.

For the "Prepare and Match" program specifically, tactical implementations would address questions such as what specific circumstances justify overriding BÄR recommendations, how explanations should be tailored for jobseekers with different backgrounds and needs, and what metrics should be used to evaluate both the AI system and the human-AI collaboration. This middle layer of implementation bridges theoretical principles and practical constraints, turning abstract values into actionable protocols that guide day-to-day operations.

At the operational level, frontline caseworkers would apply the framework through comprehensive decision-making protocols that provide step-by-step guidance for incorporating AI recommendations into decisions, specific criteria for when to seek additional human review, standardized documentation templates that capture both AI inputs and human reasoning, and regular case review sessions where difficult cases are discussed collegially. These protocols would be complemented by clear appeal and revision processes that create pathways for jobseekers to contest AI-influenced decisions, designate appeal reviewers who are not the original decision-makers, standardize review processes examining both AI recommendations and caseworker judgments, and establish feedback loops to improve both the AI system and operational guidelines based on appeal outcomes.

Continuous training and support would be essential to this operational implementation, including regular training on understanding AI capabilities and limitations, peer support networks for discussing complex cases, decision support tools that help caseworkers understand the factors driving AI recommendations, and refresher sessions on core governmental values and how they apply to AI-assisted decision-making. Regular monitoring and evaluation through audits of decision quality and consistency, analysis of patterns in human overrides, tracking of jobseeker outcomes, and documentation of caseworker experiences would complete the operational framework, ensuring continuous improvement and accountability.

In the specific case of BÄR, implementation would include training caseworkers to critically evaluate its recommendations, particularly for demographic groups where accuracy is lower (e.g., young jobseekers). It would also involve developing standardized processes for incorporating jobseeker input about their own employment prospects, drawing lessons from the successful Danish PES model which found that "the jobseeker's own assessment about their expected duration of unemployment was the most predictive factor" (Styrelsen for Arbejdsmarked og Rekruttering, 2020).

Yet, none of these procedural mechanisms were implemented when BÄR was initially deployed. This oversight is understandable given that the system was implemented several years ago when knowledge about AI governance was less developed. However, implementing this procedural framework is essential for the future if the PES is to achieve trustworthy AI that aligns with Swedish governmental values of democracy, legality, objectivity, respect, and efficiency. By combining rigorous technical evaluation with inclusive deliberative processes, the PES can develop an approach to AI that balances competing values while maintaining trustworthiness through procedural justice (Table 5).

4. Discussion

This paper has argued that trustworthy AI in public decision-making cannot be adequately secured by merely aligning technical systems with predefined normative criteria as it is often portrayed (see e.g., Kaur et al., 2022; Reinhardt, 2023; Zanotti et al., 2023). Rather, trustworthy AI governance must confront and resolve value conflicts—particularly those that are fundamental and incommensurable—through structured, procedurally just mechanisms. This suggests that public institutions should approach AI implementation with the expectation that such conflicts will arise, rather than treating them as anomalies or failures. Importantly, my framework shifts the conversation from whether AI systems can simultaneously satisfy all trustworthiness conditions (they often cannot) to how institutions can make justified trade-offs through just processes. This reconceptualization aligns with democratic governance principles where the legitimacy of decisions depends not only on outcomes but on the procedural justice that produces them (cf. Tyler, 2006, de Fine Licht and de Fine Licht 2020).

Recent research in digital government has emphasized the need for more integrated and operationalizable models of AI governance. Straub et al. (2023) propose a conceptual typology for AI in government based on dimensions like operational fitness, epistemic alignment, and normative divergence, aiming to unify fragmented approaches from technical and social science disciplines. Similarly, Gomes Rêgo et al. (2025) highlight how public organizations are internalizing ethical principles through training, standard-setting, and multi-level governance structures that support institutional learning. Their empirical analysis shows that organizations with structured training and governance practices are significantly more advanced in implementing ethical AI. The procedural justice framework developed in this paper complements these efforts by providing a normative mechanism for resolving the tensions they identify—particularly normative divergence—through structured deliberation and feedback-enabled decision processes. While existing frameworks focus on conceptual integration and institutional capacity, my framework offers a complementary strategy for addressing the ethical and legitimacy challenges that arise in practice when fundamental values conflict.

Empirical studies further underscore that citizens' acceptance of AI in public services is closely tied to procedural features such as transparency, appeals mechanisms, and the level of human involvement (Haesevoets et al., 2024; Horvath et al., 2023). Citizens tend to prefer systems in which AI plays an advisory rather than decisive role, and acceptance increases when systems are accompanied by features that support perceived fairness and legitimacy. The procedural justice framework presented here directly addresses these concerns by embedding principles like publicity, inclusion, and appeal into AI governance design. In doing so, it operationalizes procedural legitimacy in a way that aligns with public expectations and enhances institutional trust. By linking normative theory with these empirical findings, the framework helps clarify how AI systems in public administration can be rendered both reasonable, effective, and publicly acceptable.

The framework's application across strategic, tactical, and operational levels further reveals the importance of coherent governance structures for trustworthy AI. When these levels operate in isolation, fundamental disconnects emerge—as evidenced in the PES case by the gap between efficiency goals at the strategic level and the lack of appropriate discretionary authority at the operational level. Successful AI implementation requires alignment across governance levels, with clear mechanisms for feedback and adaptation. The analysis offered in this paper suggests that institutional learning should be a central

Table 5
Application of the Framework to the PES Case.

Phase	Step	Application to PES Case
Phase 1: Conflict Identification and Classification	**Step 1: Define and Clarify Values**	Identified key AI trustworthiness principles: performance (accuracy in predicting employment outcomes), calibration (confidence estimates), explainability (decision logic), intelligibility (comprehensibility), and fairness (equal treatment across demographic groups). Also considered core governmental values, such as legality, democracy, objectivity, and efficiency.
	Step 2: Determine Conflict Type	Identified major value conflicts: performance vs. explainability (neural network 68 % accuracy vs. interpretable linear model 66 %), fairness trade-offs (mathematical impossibility of meeting all fairness metrics simultaneously), efficiency vs. legality (black-box model limiting decision justification), and democracy vs. transparency (high accuracy model reducing individual autonomy).
Phase 2: Resolution of Non-Fundamental Conflicts	**Step 3: Apply Elimination Principles**	Applied Dominance, Supervaluationism, and Maximality criteria to eliminate clearly inferior AI systems. Neither the neural network nor the linear regressor was strictly dominated, requiring further analysis. Supervaluationism confirmed that the linear regressor was superior in transparency and justification, but the neural network was better in accuracy.
	Step 4: Technical Optimization	Explored technical adjustments, including explainability-enhancing tools like LIME and alternative models (decision tree + linear regressor achieving 74 % accuracy). Found that LIME explanations were unreliable and misleading, reinforcing the need for a more interpretable model.
Phase 3: Procedural Resolution of Fundamental Conflicts	**Step 5: Structured Deliberation**	Established an AI Governance Committee with diverse stakeholders (jobseekers, caseworkers, labor market experts, legal advisors, technical specialists). Implemented structured deliberation with transparency (published documents), relevance (core governmental values prioritized), inclusion (affected stakeholders engaged), and fair cooperation (balanced participation and training on AI).
	Step 6: Decision Aggregation and Implementation	Developed mechanisms for aggregating stakeholder input, documenting decisions, and establishing monitoring/ evaluation criteria. Tactical teams translated strategic principles into operational guidelines, including structured caseworker discretion and demographic fairness monitoring.
	Step 7: Appeal, Revision, and Regulation	Created an appeal system allowing jobseekers to contest AI-based decisions. Established periodic policy reviews and external audits to assess AI system performance. Ensured ongoing adaptation to align with evolving legal, ethical, and governance standards.

component of public AI governance. The documentation requirements within this framework create what I have termed an "ethical audit trail" that enables institutions to learn from implementation experiences (cf. Horvath et al., 2023). This iterative approach acknowledges that initial implementations will rarely achieve perfect balance among competing values, but can evolve toward greater trustworthiness through structured reflection and adaptation.

The framework has significant implications for emerging AI regulation efforts. Rather than prescribing rigid standards that all AI systems must simultaneously satisfy (European Parliament and Council, 2024), regulations might more productively establish procedural requirements for how public institutions navigate value conflicts. The European Union's AI Act, while establishing important baseline requirements, could be complemented by procedural governance frameworks that guide implementation within specific contexts. Rather than contradicting the EU AI Act's requirements, the approach presented in this paper offers a complementary implementation methodology that recognizes contextual implementation challenges through its risk-based approach and proportionality principles.

For instance, where the EU AI Act mandates transparency, the framework doesn't suggest abandoning this requirement, but rather provides a deliberative process for determining what level and form of transparency is appropriate in specific contexts, particularly when transparency might conflict with other mandated requirements like accuracy or robustness. This approach also aligns with the EU AI Act's emphasis on human oversight and accountability, providing concrete mechanisms for implementing these principles in practice. Additionally, the documentation requirements in my proposed framework directly support the compliance documentation mandated by the EU AI Act, creating an ethical audit trail that demonstrates both regulatory compliance and the reasoned basis for implementation choices.

The proposed framework moreover suggests that AI implementation differs significantly from traditional IT projects in several key respects. While conventional IT projects certainly involve trade-offs (Straub et al., 2023), the value conflicts in AI implementation are uniquely challenging due to their: (1) normative complexity—involving fundamental ethical and social values rather than merely technical or business considerations; (2) opacity—the 'black box' nature of many AI systems makes value trade-offs less visible and harder to assess; and (3) scalability of impact—automated decision systems can affect thousands or millions of individuals, amplifying the consequences of value prioritization choices. These distinctive characteristics demand a more structured and deliberative approach than conventional IT project management methodologies typically provide.

The aforementioned framework addresses this gap by incorporating both technical optimization (which might be found in standard IT project management) and procedural justice mechanisms (which

typically are not). Nevertheless, business IT governance models such as the COBIT framework illustrate how principle-based approaches can be operationalized through training, process standardization, and auditing (ISACA, 2018a, 2018b). While COBIT is rooted in corporate priorities and regulatory compliance, it demonstrates the feasibility of embedding abstract normative standards into structured organizational practice—a goal the framework shares for the public sector, albeit with different normative content.

The case study furthermore demonstrates that context matters significantly in determining appropriate value balancing. Rather than suggesting that certain values (like transparency) should always take precedence over others (like fairness), the aforementioned framework emphasizes that appropriate prioritization is inherently context-dependent and must emerge from a fair deliberative process. What constitutes an acceptable balance in employment services may differ substantially from what is appropriate in criminal justice or healthcare contexts, which many times is not sufficiently appreciated, such as in cases where it is argued that AI systems should always be x whatever x may be. In the Swedish PES case, for example, my analysis suggests that proper application of the framework might have led to different implementation choices—potentially favoring the more interpretable model that achieved 74 % accuracy over the neural network with 68 % accuracy—but the specific outcome would depend on the deliberative process rather than a predetermined value ranking.

Perhaps the most significant implication of the analysis is the reframing of AI governance as fundamentally a democratic challenge rather than merely a technical one (cf. Eubanks, 2018). The framework positions AI systems not as autonomous decision-makers but as tools embedded within democratic institutions that remain accountable to citizens through structured processes. This perspective suggests that trustworthy AI is not achieved through technical optimization alone but through the integration of AI systems into legitimate governance structures that respect core democratic values. While technical improvements remain important, the analysis in this paper suggests that procedural justice mechanisms—transparency, inclusion, relevance, fair cooperation, appeal, and regulation—provide the essential foundation for trustworthy AI in public services. By conceptualizing AI governance as a structured approach to value conflicts rather than a checklist of simultaneous requirements, our framework offers public institutions a pragmatic yet principled path forward. This approach acknowledges the inherent tensions in trustworthy AI implementation while providing concrete mechanisms to resolve these tensions in ways that maintain democratic legitimacy and institutional effectiveness.

5. Conclusion

This paper has proposed and demonstrated a procedural justice framework for resolving value conflicts in public AI governance. Rather than assuming that trustworthy AI can or should satisfy all normative conditions simultaneously, the framework recognizes that tensions—especially between fundamental values like transparency, fairness, efficiency, and legality—are not only frequent but often irreconcilable in practice. Drawing on philosophical theories of value conflict and procedural justice, the framework offers a structured approach that distinguishes between resolvable technical trade-offs and incommensurable normative tensions. For the former, analytical tools such as dominance, supervaluationism, and maximality help eliminate clearly inferior alternatives. For the latter, the framework invokes deliberative and participatory procedures grounded in widely endorsed principles of procedural justice.

The empirical application to the Swedish Public Employment Service (PES) illustrates the framework's practical relevance and institutional feasibility. By analyzing conflicts between model accuracy and legal justifiability, and between efficiency and democratic accountability, the case study showed that technical methods alone are insufficient to guide legitimate AI deployment in public services. Instead, structured deliberation, transparency, and inclusive decision-making processes are necessary for solving these value conflicts in a justifiable way. The framework enables such processes across strategic, tactical, and operational levels, emphasizing the importance of institutional learning and ethical auditability.

Looking forward, we see several implications for both research and practice. First, AI governance frameworks should shift from prescribing rigid technical standards toward establishing robust procedural requirements that guide implementation within specific contexts. Second, public institutions must recognize AI implementation as an iterative learning process rather than a one-time decision, with structured mechanisms for feedback, evaluation, and adaptation. Third, the development of institutional capacity for structured deliberation around AI should become a priority, ensuring that technical expertise is complemented by normative reasoning capabilities.

Future research should both empirically evaluate how procedural justice frameworks function in practice and further develop their philosophical foundations. Empirical studies could examine implementation across sectors and governance levels, assessing how principles like inclusion, appeal, and publicity affect legitimacy and trust. Philosophically, future work should explore the normative justification of procedural approaches to incommensurable value conflicts, including whether and when procedural fairness can substitute for substantive value agreement. It may also be fruitful to examine how procedural justice interacts with different theories of legitimacy, such as deliberative versus technocratic models, in the context of AI governance.

The integration of AI systems into public services represents not just a technical challenge but a profound democratic one. By conceptualizing trustworthy AI governance as a structured approach to value conflicts rather than a checklist of simultaneous requirements, the proposed framework offers public institutions a path that balances innovation with accountability, efficiency with transparency, and technical optimization with democratic legitimacy. In this way, AI can become a tool that enhances rather than undermines the fundamental values that guide public administration in democratic societies.

CRediT authorship contribution statement

Karl de Fine Licht: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process.Statement: During the preparation of this work the author(s) used ChatGPT in order to improve grammar and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Acknowledgement

I would like to thank Alex Berman and Vanja Carlsson for support and help with the first draft of the paper.

References

Allen, D. (2004). *Talking to strangers: Anxieties of citizenship since brown v. board of education*. University of Chicago Press.
Anderson, E. (1995). *Value in ethics and economics*. Harvard University Press.
Anderson, E. (1999). What is the point of equality? *Ethics, 109*(2), 287–337.
Andersson, H. (2017). *How it all relates: Exploring the space of value relations* (Doctoral dissertation, Lund University). Lund University Publications https://portal.research.

lu.se/files/21475570/How_It_All_Relates_Exploring_the_Space_of_Value_Relations.pdf.
Andersson, H., & Herlitz, A. (2022). *Value incommensurability: Ethics, risk, and decision-making*. Taylor & Francis.
Andreou, C. (2016). In E. Zalta (Ed.), *Dynamic choice*. The Stanford Encyclopedia of Philosophy.
Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.
Arbetsförmedlingen. (2023). *Uppdrag med anledning av en reformerad arbetsmarknadspolitisk verksamhet – Lägesbeskrivning den 13 oktober 2023* (Af-2022/0052 9355).
Berman, A., de Fine Licht, K., & Carlsson, V. (2024). Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system. *Technology in Society, 76*, 102471.
Brandstedt, E., & Brülde, B. (2019). Towards a theory of pure procedural climate justice. *Journal of Applied Philosophy, 36*(5), 785–799.
Broome, J. (1991). *Weighing goods: Equality, uncertainty, and time*. Blackwell.
Broome, J. (2022). Incommensurability is vagueness. In H. Andersson, & A. Herlitz (Eds.), *Value incommensurability: Ethics, risk, and decision-making* (pp. 13–29). Routledge.
Chang, R. (2002). *Making comparisons count*. New York: Routledge.
Chang, R. (2022). Are hard cases vague cases? In H. Andersson, & A. Herlitz (Eds.), *Value incommensurability: Ethics, risk, and decision-making* (pp. 50–70). Routledge.
Daniels, N. (2007). *Just health: meeting health needs fairly*. Cambridge University Press.
Daniels, N., & Sabin, J. E. (2002). *Setting limits fairly: Can we learn to share medical resources?* Oxford University Press.
Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
European Parliament and Council. (2024). 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, 310*, 1–60. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
Fine, K. (1975). Vagueness, truth, and logic. *Synthese, 30*(3–4), 265–300. https://doi.org/10.1007/BF00485047
Fraser, N. (2009). *Scales of justice: Reimagining political space in a globalizing world, (Vol. 31)*.. Columbia university press.
Habermas, J. (1985). *The Philosophical Discourse of Modernity: Twelve Lectures*. Polity Press.
Haesevoets, T., Verschuere, B., Van Severen, R., & Roets, A. (2024). How do citizens perceive the use of artificial intelligence in public sector decisions? *Government Information Quarterly, 41*(1), Article 101906.
Herlitz, A. (2024). Incommensurability and healthcare priority setting. *Philosophical Studies, 181*(12), 3347–3365. https://doi.org/10.1007/s11098-024-02160-4
Herlitz, A. (2019). Nondeterminacy, two-step models and justified choice. *Ethics, 129*, 284–308.
Herlitz, A. (2020). Nondeterminacy, cycles and rational choice. *Analysis, 80*, 443–449.
Herlitz, A., & Sadek, K. (2021). Social choice, nondeterminacy, and public reasoning. *Res Philosophica, 98*(3), 377–401.
High-Level Expert Group on AI (AI HLEG). (2019). *Ethics guidelines for trustworthy artificial intelligence*.
Horvath, L., James, O., Banducci, S., & Beduschi, A. (2023). Citizens' acceptance of artificial intelligence in public services: Evidence from a conjoint experiment about processing permit applications. *Government Information Quarterly, 40*(4), Article 101876.
ISACA. (2018a). *COBIT 2019 framework: Introduction and methodology*. Information Systems Audit and Control Association.
ISACA. (2018b). *COBIT 2019 implementation guide: Implementing and optimizing an information and technology governance solution*. Information Systems Audit and Control Association.
Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys, 55*(2), 1–38.
Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Comput. Surv., 55*(2), 1–38.
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *In 8th innovations in theoretical computer science conference (ITCS 2017)*. Schloss: Dagstuhl-Leibniz-Zentrum fuer Informatik.
Loi, M., Herlitz, A., & Heidari, H. (2023). Fair equality of chances for prediction-based decisions. *Economics and Philosophy, 1–24*. https://doi.org/10.1017/S0266267123000342
Nussbaum, M. C. (2011). *Creating Capabilities: The human development approach*. Harvard University Press.
Petersen, T. S. (2021). Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts. *Etikk i Praksis - Nordic Journal of Applied Ethics, 1*, 25–40.
Pettit, P. (2012). *On the people's terms: A republican theory and model of democracy*. Cambridge University Press.
Rabinowicz, W. (2008). Value relations. *Theoria, 74*(1), 18–49.
Rawls, J. (1999). *Collected papers*. Harvard University Press.
Raz, J. (1986). *The morality of freedom*. Clarendon Press.
Rêgo, G., de Almeida, P., & dos Santos Júnior, C. D. (2025). Artificial intelligence governance: Understanding how public organizations implement it. *Government Information Quarterly, 42*(1), Article 101880.
Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics, 3*, 735–744. https://doi.org/10.1007/s43681-022-00200-5
Ryberg, J., & Petersen, T. S. (2022). Sentencing and the conflict between algorithmic accuracy and transparency. *Sentencing and Artificial Intelligence*, 57–73.
Scanlon, T. M. (2000). *What we owe to each other*. Harvard University Press.
Sen, A. (1970). *Collective choice and social welfare*. Holden-Day.
Sen, A. (1997). Maximization and the act of choice. *Econometrica, 65*(4), 745–779.
Srinivasan, A. (2021). *The right to sex: Feminism in the twenty-first century*. New York: Farrar, Straus and Giroux.
Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023). Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly, 40*(4), Article 101881.
Statskontoret. (2019). *Den statliga värdegrunden: Gemensamma principer för en god förvaltning. Ordförrådet*. ISBN 978-91-88865-23-6.
Styrelsen for Arbejdsmarked og Rekruttering. (2020). *Beskrivelse Af Profilafklaringsværktøjet Til Dagpengemodtagere* [Description of the Profiling Tool for Unemployment Benefit Recipients].
Sweden. (2021). Uppdrag att främja offentlig förvaltnings förmåga att använda artificiell intelligens. June 21 https://www.regeringen.se/regeringsuppdrag/2021/06/uppdrag-att-framja-offentlig-forvaltnings-formaga-att-anvanda-artificiell-intelligens/.
Tyler, T. R. (2006). *Why people obey the law*. Princeton: Princeton University Press.
Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2023). Keep trusting! A plea for the notion of trustworthy AI. *AI & SOCIETY*, 1–12.

Karl de Fine Licht is an Associate Professor at Chalmers University of Technology, with a PhD in Practical Philosophy from the University of Gothenburg. His research primarily explores trust, legitimacy, and justice in public decision-making, with a substantial focus on the ethical and societal aspects of artificial intelligence. Over the years, de Fine Licht has been actively involved in integrating these critical discussions into academic curricula and engaging in public dialogues on the responsible application of AI in governance and societal systems.