AI in Public Decision-Making: A Philosophical and Practical Framework for Assessing and Weighing Harm and Benefit

Public Administration

WILEY

public administration

SPECIAL ISSUE ARTICLE OPEN ACCESS

AI in Public Decision-Making: A Philosophical and Practical Framework for Assessing and Weighing Harm and Benefit

Karl de Fine Licht[1] | Anna Folland[2]

[1]Science, Technology, and Society, Chalmers University of Technology, Gothenburg, Sweden | [2]Department of Communication and Arts, Roskilde University, Roskilde, Denmark

Correspondence: Karl de Fine Licht (karl.definelicht@chalmers.se)

Received: 28 March 2025 | **Revised:** 17 September 2025 | **Accepted:** 22 September 2025

Funding: The authors received no specific funding for this work.

ABSTRACT

Artificial intelligence (AI) is increasingly used in public decision-making; yet existing governance tools often lack clear definitions of harm and benefit, practical methods for weighing competing values, and guidance for resolving value conflicts. This paper presents a five-step framework that integrates moral philosophy, trustworthy AI principles, and procedural justice into a coherent decision process for public administrators. The framework operationalizes harm and benefit through multidimensional well-being measures, applies normative principles such as harm–benefit asymmetry, incorporates technical assessment criteria, and offers structured methods for resolving both derivative and fundamental value conflicts. A worked example, based on the Dutch childcare benefits scandal, illustrates its application under real-world constraints. Comparative analysis positions the framework alongside established tools, highlighting its added value in combining normative reasoning with procedural legitimacy. The paper also discusses implementation challenges, including cognitive biases, institutional inertia, and political trade-offs, and suggests empirical approaches for validation. By linking philosophical depth with practical usability, the framework supports transparent, context-sensitive governance of AI in the public sector.

1 | Introduction

In recent years we have seen a great increase in the use of artificial intelligence (AI) in public decision-making (see e.g., O'Neil 2017; Eubanks 2018; de Fine Licht and Fine Licht 2020, Vogl et al. 2020; Kaur et al. 2022; Berman et al. 2024). These systems are used by employment agencies to determine people's employability, by social services to assess whether children are at risk of being harmed, and by the police to predict areas with the highest risk of criminal behavior within their territory of responsibility, just to mention a few examples. Yet, the use of AI systems in public decision-making is not devoid of risks and controversies. Because of this, much has been written about ethical or trustworthy AI in public decision-making (see e.g., Wirtz and Müller 2019; HLEG 2019; Busuioc 2021; Kaur et al. 2022; Reinhardt 2023; Zanotti et al. 2023; Berman et al. 2024). In general, the notion is that we should aim for benefits while avoiding harms, respect people's autonomy, and strive for transparency, interpretability, accuracy, accountability, etc. regarding the decisions the system or the decision-maker takes with the support of these systems.

Despite progress, there is still a lack of research on clearly defining "harm" and "benefit" in AI decision-making, establishing normative foundations, and addressing value conflicts (Petersen 2021; Ryberg and Petersen 2022; Reinhardt 2023). Recent developments in moral philosophy have contributed to clarifying these concepts and managing value conflicts (e.g., Andersson and Herlitz 2022; Folland 2025a). Philosophical work on AI in public decision-making has, for instance, proposed

ways to identify core values and connect them to transparency, interpretability, and explainability (Ryberg and Petersen 2022). However, these efforts often stop short of offering methods for weighing harm against benefit or balancing competing values. Moreover, a gap remains between philosophical theory and practical engagement with AI technologies in public-sector contexts. This disconnect hinders the translation of philosophical insights into workable institutional frameworks. While some promising interdisciplinary work has emerged (e.g., Berman et al. 2024), further research is needed to bridge these domains.

In this paper, we aim to address these gaps by applying recent advancements in moral philosophy to the context of AI decision-making systems. Specifically, we draw on and extend the framework outlined by Johansson and Risberg (2023) and others, which provides a detailed analysis of harm, benefit, and their comparative weights. We explore how the concepts of harm and benefit (Algander 2013; Folland 2025a; Johansson and Risberg 2023), and principles specifying how to measure their extent—such as the "Degree of Harm" (Gardner 2017)—can be operationalized in AI systems. Furthermore, we examine how normative principles invoking harm and benefit can be integrated into AI decision-making processes. Even when we focus solely on harm and benefit, value conflicts may still arise. To address this, we propose a hybrid account for resolving such conflicts (cf. Herlitz and Sadek 2021; de Fine Licht 2025). This approach can also be extended to manage tensions among other core values in AI ethics and trustworthy AI, such as transparency, explainability, interpretability, and accuracy. The paper concludes by presenting a practical framework for decision-making in the context of AI systems.

The paper is structured as follows: Section 2 outlines our methodological approach, combining conceptual analysis with practical application. Section 3 examines foundational concepts of harm and benefit, drawing on well-being theory and establishing normative principles for ethical evaluation. Section 4 integrates these concepts with trustworthy AI principles and addresses value conflicts using analytical and procedural justice approaches. Section 5 presents a unified framework for ethical AI governance in public administration, including concrete tools and implementation guidelines. Section 6 discusses implementation challenges and directions for future research, followed by concluding remarks in Section 7.

2 | Methods and Materials

The framework was developed through systematic philosophical analysis, following established approaches in applied ethics (Beauchamp and Childress 2019; Daniels 1996). The process unfolded in four stages that together link conceptual foundations to practical application. We began with a comprehensive literature mapping, using searches in PhilPapers, Web of Science, Scopus, and Google Scholar with terms such as *harm*, *benefit*, and *well-being*. Sources were selected for theoretical contribution, relevance to public decision-making, and conceptual rigor. Foundational works—including Mill (1859/1991), Feinberg (1984), Bradley (2012), and Johansson and Risberg (2023)—were analyzed for their definitional approaches, normative commitments, and implications for practice.

Building on this base, we conducted a comparative conceptual analysis of counterfactual, causal, and well-being accounts of harm and benefit. This involved testing necessary and sufficient conditions, assessing internal consistency, and evaluating their potential for operationalization. Reflective equilibrium (Rawls 1971) was used to refine emerging principles against canonical philosophical cases. The harm–benefit asymmetry principle, for example, was calibrated by adjusting its weighting coefficient until it yielded consistent results across established examples. From here, we synthesized normative principles by identifying areas of convergence in the literature. Thematic analysis of recurring arguments produced an initial set of 12, which was refined to the seven presented in Section 3 by eliminating redundancies and mapping logical dependencies. Each principle was then tested for coherence against both classical philosophical cases and scenarios drawn from AI governance.

Finally, we operationalized these concepts for use in public decision-making. Validated well-being measures were reviewed against our three-dimensional account of well-being (hedonic, desire-fulfillment, objective list), leading to the selection of the Positive and Negative Affect Schedule (PANAS), Satisfaction with Life Scale (SWLS), and Questionnaire for Eudaimonic Well-Being (QEWB). Public-sector risk assessment methodologies (e.g., ISO 31000; Cox Jr. 2008) informed the design of compatible scoring systems, while existing AI governance instruments such as Data Protection Impact Assessments (DPIAs) and Algorithmic Impact Assessments (AIAs) guided procedural integration. This methodological foundation underpins the following sections. Section 4 integrates the harm–benefit framework with principles of trustworthy AI and approaches to value conflict, while Section 5 presents the operationalized framework through assessment matrices, implementation tools, and an illustrative case: the Dutch childcare benefits scandal and the SyRI welfare-fraud detection system.

3 | Harm and Benefit

Harm and benefit are central to evaluating the ethical acceptability of AI in public decision-making. This section draws on philosophical theories of well-being to clarify their dimensions, normative significance, and role in ethical evaluation. We begin by examining harm and benefit through major well-being theories—hedonism, desire-fulfillment, and objective list accounts—to identify how impacts can be recognized and assessed. We then outline normative principles, including reasons against harming, conditions under which harm may be justified, and the asymmetry between harming and failing to benefit. These concepts form the foundation of the framework developed in Section 5.

3.1 | The Concepts of Harm and Benefit

Identifying harm is a crucial step in any ethical evaluation, particularly in public decision-making involving AI. Most moral philosophers agree that harming someone is *negatively affecting their well-being* (Gardner 2021; Johansson and Risberg 2023), and that harm comes in *degrees*, where the degree of harm depends on the severity of its impact (Gardner 2017). Spelled out in

terms of necessary and sufficient conditions, here is that view: an event harms an individual if, and only if, it negatively affects their well-being. The concept of benefit is generally seen as the mirror opposite of harm, such that an event benefits an individual if, and only if, it positively affects their well-being.

There are two main accounts of what it is to "negatively affect" someone's wellbeing: counterfactual and causal accounts.[1] Counterfactual accounts hold that an event harms an individual if they would have been better off had it not occurred; for example, an AI-based decision that someone is no longer eligible for social benefit counts as harmful if that person would have been better off in the absence of that decision. Supposing the person otherwise would have kept their access to social benefits, the AI-based decision is likely harmful, since access generally improves a person's economic situation. Causal accounts define harm as an outcome that directly reduces well-being, without reference to hypothetical comparisons; for instance, an AI decision leading to the loss of housing or healthcare. This approach avoids the complexities of counterfactual comparisons but requires a clear causal link between the event and the negative outcome. For instance, if an AI system generates a decision that leads to an immediate loss of housing or healthcare access, that decision counts as harmful.

Philosophers distinguish between three main theories of well-being (e.g., Brülde 1998; Crisp 2017). Simply put, hedonism holds that a good life consists of positive feelings and the absence of negative ones. Desire theories equate well-being with the fulfillment of one's desires and the unfulfillment of one's aversions. Objective list theories hold that certain characteristics—such as meaningful work, decent living conditions, strong relationships, recognition, and personal autonomy—are intrinsically valuable regardless of whether they are desired or enjoyed (Brülde 1998, 286–367; Crisp 2017; Hurka 1993).

These theoretical perspectives correspond to well-established measurement instruments. The Positive and Negative Affect Schedule (PANAS) assesses emotional states, capturing the hedonistic dimension (Watson et al. 1988). The Satisfaction with Life Scale (SWLS) measures cognitive evaluations of life satisfaction, aligning with desire theory (Diener et al. 1985). The Questionnaire for Eudaimonic Well-being (QEWB) captures purpose, growth, and engagement, reflecting objective list theory (Waterman et al. 2010). Well-being is thus multidimensional, and AI systems may affect these dimensions differently. For example, an AI that accelerates service delivery may increase life satisfaction (desire fulfillment) while reducing autonomy (eudaimonic well-being). Similarly, AI-generated mental health interventions might improve short-term mood but overlook deeper goals of meaning and self-realization. Such tensions highlight the need for frameworks that anticipate and justify trade-offs between dimensions, rather than privileging one without reason. We are going to examine this more thoroughly in Section 4.

3.2 | Placeholder TextNormative Principles

If the concepts of harm and benefit help determine *what* is at stake, normative principles determine *how* these impacts should be weighed and justified in decision-making. These principles provide a structured basis for moral reasoning, clarifying not only whether an action is harmful or beneficial but also the extent of its impact, whether the harm is avoidable or redundant, and under what circumstances it may be justified.

A core assumption in moral philosophy is that the fact that an action is harmful constitutes a moral reason against performing that action (Algander 2013, 135; Gardner 2017, 73–74; Shafer-Landau 2021; Shiffrin 2012, 361). This is sometimes referred to as the *reason against harming*: if an action is harmful, there is a moral reason to avoid it. Conversely, harmless actions are often judged permissible—an idea central to John Stuart Mill's *Harm Principle*, which holds that the state may justifiably restrict liberty only to prevent harm to others (Mill 1977, for critical discussion, see Gardner 2017; Folland 2025b). Political theorists have applied this principle to defend individual freedom in areas such as reproductive technologies and religious expression (Holtug 2002). In public administration, this distinction suggests that AI systems with negligible impact on well-being—such as those optimizing nonessential service schedules—require less ethical scrutiny, allowing governance attention to focus where citizens' fundamental interests are at stake.

Although there is generally strong moral reason to avoid causing harm, certain conditions can render it at least ethically permissible. Shiffrin (2012, 362) identifies three such conditions: (a) the harm is deserved, (b) the harm is necessary to prevent a greater harm—either to the same individual or to others—or (c) the individual affected has given informed consent. In public administration, these conditions frequently arise because decisions often involve distributing scarce resources, managing risks, or enforcing rules that cannot fully satisfy all interests. For instance, an AI system allocating limited medical resources may delay or reduce treatment for some patients if doing so enables the provision of urgent, life-saving care to others. In other cases, harm may be justified where it results from fair enforcement of legal penalties, such as AI-assisted detection of tax fraud leading to fines or prosecutions. Consent can also play a role in legitimating harm: citizens may agree to participate in AI-driven pilot programs that carry certain risks (e.g., experimental traffic management systems) in exchange for potential long-term benefits to the community.

The strength of the reason against harming is widely understood to be proportional to the degree of harm: the more severe the harm, the stronger the reason to avoid it (Gardner 2017, 83). This distinction is salient for AI systems that vary in stakes—wrongful denial of essential social benefits, with risks of eviction and severe distress, carries greater moral urgency than minor inconveniences caused by rescheduling municipal services. Two further factors shape the moral weight of harm. First, harm is more significant when it is *less inevitable*: if the same type of harm would not occur without the action, the reason against causing it is stronger (Gardner 2017, 85). For example, if a child-neglect prediction model produces a high false-positive rate, the avoidable nature of resulting intrusions strengthens the moral case against its unqualified use. Second, harm is less significant when it is *redundant*: if the harm would occur regardless of the decision, the reason against causing it is weaker (Gardner 2017, 86). In predictive policing, harm from increased surveillance

may be redundant if it would have happened under existing policing strategies as well; it is nonredundant—and thus carries more moral weight—if an AI system does harm that would not have occurred anyway.

Benefit is often defined as the mirror opposite of harm: an event benefits an individual if it positively affects their well-being. The degree of benefit is proportional to the improvement in well-being, ranging from minor (e.g., timely service information) to substantial (e.g., secure housing or access to life-saving treatment). However, many philosophers hold that harm and benefit are not morally equivalent. The *harm–benefit asymmetry* principle states that, when harm and benefit are equal in degree, the moral reason against causing harm is stronger than the reason to provide the benefit (Feit 2019; Folland 2023; Shiffrin 2012). As Bradley (2013, 39) illustrates, failing to rescue a drowning child at no cost is morally worse than failing to provide an equally large but less urgent benefit. In public-sector AI, this implies that, when preventing harms and providing benefits are in tension, preventing significant harm should take precedence over securing proportionate benefits. For example, in unemployment benefit eligibility systems, avoiding stripping people of social benefits wrongfully should take priority over marginal efficiency gains for those receiving benefits.

Having established the conceptual foundations of harm and benefit, we now turn to the methodological steps through which these principles were developed and operationalized, providing the groundwork for their application in AI governance contexts.

4 | Integrating Harm and Benefit Into Trustworthy and Ethical AI

This section outlines how to address value conflicts in AI-driven public decision-making and thereby sets up for the practical framework in Section 5. It begins with identifying key values in trustworthy and ethical AI; it then moves on to discuss conflicts between nonfundamental values, which can often be resolved through appeal to shared normative commitments, deliberative processes, or decision-theoretical methods. It then turns to fundamental value conflicts—cases of genuine incommensurability—where no common evaluative standard exists to adjudicate between competing principles. In such instances, we argue that procedural justice offers the most legitimate and practically viable path to resolution.

4.1 | The Conditions for Trustworthy or Ethical AI

Harm and benefit are integral concepts for understanding what ethical or trustworthy AI amounts to. Yet, there is much more to the ethical evaluation than simply identifying harms and benefits; values such as transparency, justice, and legality also shape how these are perceived and weighed. To develop a *practical* decision-making framework for AI governance, it is necessary to further examine how discussions of benefit and harm can be integrated into such frameworks. Furthermore, when benefits and harms are understood in terms of well-being, tensions may arise both *within* dimensions of well-being (e.g., emotional welfare vs. autonomy) and *between* well-being and other principles, such as justice, transparency, or legality. These conflicts are often unavoidable, given the competing goals and priorities embedded in AI governance frameworks. Hence, for the framework to be feasible, we need to address such conflicts.

It is commonly thought that "trustworthy AI" and "ethical AI" include a broad spectrum of associated values and principles. These include, but are not limited to, performance, calibration, interpretability, explainability, intelligibility, fairness, legality, and accountability. Berman et al. (2024) provide a recent overview and operationalization of trustworthy AI principles. According to Berman and colleagues, trustworthy AI should be understood as part of a broader sociotechnical system. Accordingly, it is not the AI system in isolation that should be evaluated, but the system as a whole, including institutional structures, human actors, and the contexts of deployment. This sociotechnical perspective is prevalent in both academic literature and policy practice. It reflects the recognition that elements such as human oversight, institutional accountability, and stakeholder engagement are integral to the trustworthiness of AI systems—even though they are not strictly technical features of the AI itself.

Berman and colleagues identify several conditions for trustworthy AI (Berman et al. 2024). Performance is assessed by the system's accuracy, its ability to enhance human decision-making, and the communication of its performance to stakeholders. Calibration concerns the accuracy and reliability of confidence estimates provided to users. Interpretability and explainability address whether stakeholders can understand the system's decision logic and whether the explanations faithfully reflect how decisions are made. Intelligibility and availability focus on making these explanations not only theoretically accessible but practically comprehensible. While equal and fair treatment is not elaborated in detail by Berman et al., it should at a minimum include predictive fairness—the principle that prediction errors and accuracy should be distributed equally across demographic groups (Loi et al. 2023). Finally, legality, accountability, appeal, and human oversight pertain to the broader socio-technical system in which AI is embedded. Trustworthy AI must comply with legal standards, assign responsibility for decisions, offer channels for appeal, and include meaningful human involvement.

4.2 | Identifying and Resolving Nonfundamental Value Conflicts

To get a genuinely practical, in the sense of feasible, framework it needs to manage value conflicts in a reasonable way. This is because the conditions of trustworthy or ethical AI often come in conflict, both with one another and with broader normative commitments such as promoting justice or well-being. As demonstrated through case studies (e.g., Berman et al. 2024) and formal analyses (Loi et al. 2023), these conflicts are not merely practical implementation issues but reflect deeper incommensurabilities between values. For example, accuracy and transparency often conflict in AI systems: enhancing transparency may require simplification or disclosure that compromises predictive performance, while optimizing accuracy may depend on opaque, complex models that reduce intelligibility (Kaur et al. 2022). Similar tensions arise between transparency and security, where making system operations more visible can

undermine safeguards against adversarial attacks (Gongora-Salazar et al. 2023).

Some conflicts can be resolved using the harm-benefit framework, making the framework offered here a step toward practical application. For example, the principle of harm-benefit asymmetry helps adjudicate tensions between reducing harm and promoting benefit by prioritizing harm avoidance over delivering an equivalent benefit. In practice, this means that when AI systems face trade-offs, minimizing harm takes ethical precedence. This can clarify cases where an AI system improves efficiency (a benefit) but risks exclusion or distress for some users (a harm).

However, not all conflicts are easily resolved, even within this framework. Harms and benefits can range from rare to frequent, from minor to severe, as well as be unevenly distributed among groups. Harms and benefits may be frequent but minor: AI fines many citizens small amounts in error (harm) or sends frequent minor service updates (benefit). Harms and benefits can also be rare but severe: AI might wrongly evict a few families (harm) or award rare but transformative grants (benefit). When various outcomes of these sorts are at stake, issues of fairness and distributive justice arise. As mentioned, tensions may also arise between dimensions of well-being itself, for instance, when promoting life satisfaction compromises autonomy, or when increasing positive effect undermines long-term purpose. These are normative dilemmas involving potentially incommensurable goods. Beyond well-being, conflicts grow more complex. Clashes between well-being and values such as legality, democratic legitimacy, or procedural fairness may not be resolvable through harm-benefit comparisons alone.

Given these considerations, value conflicts in trustworthy or ethical AI governance typically arise in two forms: (1) between the various conditions that define trustworthy or ethical AI (e.g., accuracy vs. transparency), and (2) between those conditions and values external to them or more fundamental (e.g., justice, democratic legitimacy). For example, a conflict between accuracy and transparency involves two internal conditions, both potentially justified by broader concerns such as preventing harm or promoting benefit. In addressing such conflicts, it is necessary to begin by precisely defining the concepts involved and identifying the core values that underpin them (cf. Ryberg and Petersen 2022). This entails, first, examining whether a genuine conflict exists by clarifying key terms, and second, determining whether the conflict reflects a deeper tension between fundamental values. In some cases, what appears to be a conflict may dissolve once its conceptual or normative basis is clarified. In others, understanding which fundamental values justify the contested principles may reveal ways to resolve the conflict without compromising those underlying commitments.

What counts as a fundamental value is, of course, contestable. However, if the aim is to develop a framework for public agencies that is both principled and practical—neither too rigid nor too permissive—it is helpful to draw on Rawls's notion of a "realistic utopia" (Rawls 1999). This involves setting aspirational yet achievable goals by identifying the ideals a society should pursue within the constraints of political and institutional realities. Accordingly, the framework should be grounded in values that most citizens living in liberal democracies would endorse upon reflection and that public decision-makers are obligated to uphold. Core principles of democratic governance—such as constitutional commitments and entrenched bureaucratic norms—offer a natural foundation (cf. de Fine Licht 2025). Examples include prioritizing the worst-off, promoting transparency, enhancing efficiency, ensuring legality, and maintaining accountability. In the context of trustworthy AI, some of its conditions can thus be viewed as reflecting foundational democratic values, such as accountability and fairness, while others are more contingent or context-specific, such as accuracy and calibration. Identifying and safeguarding these core values is essential to designing a governance framework that is both ethically robust and practically workable.

When nonfundamental or derivative values come into conflict, analytical principles can help resolve these tensions while preserving the fundamental values they support. This is important, as public agencies generally aim to uphold their established norms and values wherever possible. Three principles can help eliminate clearly suboptimal options and narrow the decision space to those alternatives that best align with core commitments: the Dominance Principle, Supervaluationism, and the Maximality Principle. To reduce complexity and improve efficiency, the Dominance Principle is applied first. It excludes AI systems that are clearly inferior—those that perform worse on at least one dimension of trustworthiness without compensating gains elsewhere (Savage 1954; Sen 1970; Broome 1991). This ensures that decision-makers focus only on nondominated options. For example, if one AI system consistently delivers higher accuracy without reducing fairness, an alternative that performs worse on both counts can be excluded from further consideration.

Next, building upon our earlier emphasis on precise definitions, we should apply Supervaluationism when nondeterminacy arises from vague evaluative criteria such as "fairness" or "transparency" (Fine 1975; Andersson 2017). This formal approach systematically implements the definitional clarity we identified as crucial earlier, 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. For example, suppose one system scores moderately well on transparency but poorly on fairness, while another performs consistently well across all reasonable definitions of both. Even if we cannot definitively settle on a single interpretation of fairness, Supervaluationism allows us to rule out the first system because it underperforms on every admissible understanding of the relevant values.

Lastly, Sen's (1970, 1997, 2000) 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 to 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, nondominated choices remain, preserving ethical pluralism while maintaining decision feasibility.

4.3 | Resolving Fundamental Value Conflicts

When derivative values clash, nonderivative values or principles can adjudicate the conflict. But when nonderivative values themselves clash and cannot be reduced to a common measure, we face value incommensurability. In such cases, principles from decision theory—such as Dominance, Supervaluationism, and Maximality—help eliminate clearly suboptimal options and preserve coherence in decision-making.

Consider a city deciding how to allocate scarce public housing using data-driven systems. One option is a highly accurate deep model drawing on both administrative and social media data; another is a more interpretable model using only consented administrative data, offering slightly lower accuracy but better group parity; a third is a transparent points-based rule system with perfect parity but the lowest accuracy; and a fourth is the legacy first-come-first-served process. The Dominance Principle excludes the legacy process since it performs strictly worse on accuracy and parity. Supervaluationism then rules out the deep model once side-constraints against unconsented data are applied, since it fails under all admissible completions. Maximality identifies the interpretable model and the points-based system as undefeated, as each is stronger on one value but weaker on another. Here, decision-theoretic principles narrow the options but cannot determine a final choice when fundamental values—such as privacy, equality, and welfare protection—cannot be meaningfully compared.

The challenge intensifies when broader external values are considered. While elimination methods maintain rigor by excluding irrational options, they are limited as a full solution (Herlitz 2019, 2020; Herlitz and Sadek 2021). High-stakes public decisions require deeper justification than elimination alone can provide. Returning to the housing example, both the interpretable model and the points-based system remain after exclusion, but a random choice between them would not address the substantive disagreements at stake. As Herlitz and Sadek argue, arbitrary selection fails to provide reasons acceptable to stakeholders whose access to housing depends on fair procedures. Andreou (2016) further observes that repeated random selection among nonrankable options can yield systematically suboptimal outcomes, reinforcing the need for more robust justificatory processes.

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 leading up to them. Since justice and trustworthiness are closely related, adopting a just procedure in AI development is likely to yield trustworthy AI systems, provided that technical aspects function correctly.

Many contemporary philosophers and political theorists advocate procedural solutions for addressing value conflicts (Anderson 1999; Chang 2002; Daniels and Sabin 2002; Tyler 2006; Nussbaum 2011; Pettit 2012; Andersson and Herlitz 2022). Herlitz (2024) further argues that the inherent indeterminacy of value conflicts underscores the need for procedural approaches. Yet, as Herlitz and Sadek note, existing accounts often lack detailed conditions for deliberation. Given the applicability of procedural methods across public institutions, this part of the paper develops a framework drawing on theories of procedural justice. Although these theories differ in emphasis, they converge on several core elements. Here, the focus is on the most widely accepted criteria, balancing theoretical grounding with institutional feasibility.

The publicity and relevance conditions ensure transparency and justification. Publicity (Daniels and Sabin 2002; Pettit 2012; Fraser 2009; Nussbaum 2011) requires that decisions and their rationales be accessible to the public, allowing scrutiny and understanding. For example, if a decision favors a highly accurate but opaque machine-learning model over a transparent rule-based system for unemployment benefits—on the grounds that reducing erroneous denials matters more—this reasoning must be publicly communicated. Relevance (Habermas 1985; Daniels and Sabin 2002; Brandstedt and Brülde 2019) adds that justifications must be based on reasons acceptable to those affected, grounded in evidence, shared principles, or normative theories rather than narrow self-interest (Pettit 2012). Stakeholder experience can also shape the weight of reasons, as illustrated by testimony from historically disadvantaged groups influencing the prioritization of values (Herlitz 2024).

The conditions of inclusion, cooperation, and deliberative quality address participation. Inclusion (Anderson 1999; Fraser 2009; Pettit 2012) requires meaningful involvement of all relevant stakeholders, both directly and indirectly affected. It calls for proactive engagement of marginalized groups and attention to representational balance across service recipients, institutional staff, and public contributors. Fair cooperation (Habermas 1985; Anderson 1999; Fraser 2009; Pettit 2012; Nussbaum 2011; Brandstedt and Brülde 2019) emphasizes mutual respect, reciprocity, and equitable deliberation. This includes supporting participants unfamiliar with formal reasoning or procedural norms, discouraging arguments grounded solely in personal interest, and ensuring shared understanding of the technologies under discussion.

Finally, the conditions of appeal, revision, and enforcement safeguard procedural integrity over time. Appeal and revision (Nussbaum 2011; Tyler 2006; Pettit 2012) require mechanisms

to challenge and modify decisions in light of new evidence or arguments, keeping processes responsive to evolving values and circumstances. Enforcement ensures compliance with procedural standards through oversight mechanisms such as review bodies or external auditors. These accountability structures help maintain fairness, transparency, and consistency in AI-related decision-making.

5 | A Framework for Ethical AI in Public Decision-Making

We present a structured approach to guide public administrators in the ethical governance of AI, showing how the seemingly distinct normative domains—harm and benefit, trustworthy AI, and democratic decision-making—converge around shared commitments to human well-being, fairness, and legitimacy. Based on this integration, we develop practical tools, matrices, and implementation guidelines that translate abstract principles into actionable mechanisms. The resulting framework offers a methodology that balances ethical rigor with administrative feasibility, supporting the responsible use of AI in the public sector.

5.1 | Setting up the Framework

So far, we have discussed three interrelated domains of norms and values: harm and benefit, trustworthy AI, and public decision-making in liberal democracies. While analytically distinct, these domains substantially overlap. Trustworthy or ethical AI frameworks routinely include harm and benefit considerations—AI4People, for instance, highlights beneficence and nonmaleficence (Floridi et al. 2018), and the EU AI Act emphasizes risk mitigation and harm prevention. Similarly, democratic norms encompass key principles of trustworthy AI, such as fairness, transparency, and accountability. Harm and benefit considerations are also embedded in foundational democratic thought, exemplified by Mill's Harm Principle.

This convergence suggests the need for a unified framework for ethical AI governance in public administration. Although each domain emphasizes different elements, they share core commitments to human well-being, procedural fairness, and democratic legitimacy. Our proposed framework integrates these perspectives while preserving their distinct contributions. The harm-benefit analysis offers substantive ethical content, guided by well-being theories that provide concrete metrics for evaluating AI's impact. The trustworthy AI principles supply technical and governance standards, ensuring systems function reliably and within acceptable bounds. The democratic norms ensure procedural legitimacy through inclusive deliberation, public justification, and mechanisms for appeal and revision.

This integrated framework functions as a decision procedure composed of five interconnected steps:

1. Identification and Measurement: Administrators begin by assessing potential harms and benefits using the multi-dimensional well-being framework from Section 3.1, operationalized through tools like PANAS, SWLS, and QEWB (see Table A1 in Appendix A).
2. Normative Evaluation: Identified harms and benefits are evaluated using principles outlined in Section 3.2—such as Reason Against Harming, Strength of Reason, and Harm–Benefit Asymmetry—to form initial ethical judgments (see Table A2 in Appendix A).
3. Technical Assessment: In parallel, the system is assessed against trustworthy AI criteria (Section 4.1), including performance, interpretability, fairness, legality, and accountability (see Table A3 in Appendix A).
4. Conflict Resolution: Where value conflicts emerge—between well-being dimensions, trustworthy AI principles, or other values—analytical methods such as the Dominance Principle, Supervaluationism, and Maximality (Section 4.2) are used to eliminate clearly suboptimal options (see Figure A1 in Appendix A).
5. Procedural Deliberation: For unresolved conflicts involving fundamental values, the procedural justice approach outlined in Section 4.3 is applied. This includes ensuring publicity, relevance, inclusion, fair cooperation, appeal mechanisms, and enforcement (see Table A4 in Appendix A).

This approach recognizes that ethical governance of AI cannot rely solely on substantive or technical criteria; procedural legitimacy is equally essential. The framework accommodates persistent value incommensurability while offering practical steps for making ethically defensible decisions. The next section operationalizes this framework through newly developed tables, matrices, and a flowchart (presented in Appendix A), each linked to the framework's five steps, to provide clear, at-a-glance overviews that improve accessibility for both academic and practitioner audiences.

5.2 | Operationalizing the Framework: Tools, Matrices, and Implementation Guidelines

Building on the integrated framework outlined in Section 5.1, we now operationalize theoretical considerations into practical governance mechanisms—introducing new tables, matrices, and a decision flowchart in Appendix A for each stage of the process—to guide public administrators in ethically implementing AI systems. These structured tools correspond to each phase of the five-step decision-making process, ensuring consistency and coherence throughout the framework.

To identify and measure potential harms and benefits (Step 1), we have developed a Multi-dimensional Impact Assessment Matrix (see Table A1 in Appendix A). This matrix evaluates AI impacts across three distinct well-being dimensions: emotional states (hedonic), preference satisfaction (desire fulfillment), and meaningful life elements (objective list). Established instruments such as PANAS, SWLS, and QEWB were selected based on empirical validation and widespread adoption in well-being research. Multiple data collection methodologies—including pre/post surveys, experience sampling, and longitudinal tracking—enable comprehensive assessment, capturing the varied manifestations of AI impacts across different contexts and time frames. Crucially, baseline measurements taken prior to AI implementation facilitate meaningful counterfactual comparisons,

while analyses stratified by demographic groups identify disparities and support equity-oriented governance.

For normative evaluation (Step 2), we introduce the Ethical Weight Assessment Tool (see Table A2 in Appendix A), aligned with the normative principles articulated in Section 3.2. This tool evaluates each identified impact according to three critical dimensions: severity, inevitability, and redundancy. A clearly defined scoring system quantifies these dimensions on a 1–5 scale. The harm–benefit asymmetry principle is operationalized explicitly by assigning greater ethical significance to harms compared to benefits. We recommend initial independent scoring followed by consensus discussions to minimize bias, supplemented by periodic reassessments as new evidence emerges, particularly concerning high-severity harms, consistent with precautionary governance.

The technical assessment phase (Step 3) incorporates a Trustworthy AI Compliance Matrix (see Table A3 in Appendix A) addressing essential criteria outlined in Section 4.1, including performance, explainability, fairness, legality, and accountability.

The matrix (Table A3 in Appendix A) provides targeted assessment questions, specifies required evidence types, and categorizes compliance levels (full, partial, noncompliant) in a tabular format for quick reference and easier application in practice. The matrix prioritizes outcome-focused evaluation over specific technical prescriptions and explicitly acknowledges the inherent trade-offs among different trustworthy AI principles.

In addressing value conflicts (Step 4), the Analytical Resolution Flowchart (see Figure A1 in Appendix A) applies the decision-theoretic methods described in Section 4.2—namely the Dominance Principle, Supervaluationism, and Maximality. This systematic approach supports administrators in identifying and eliminating clearly suboptimal choices while respecting legitimate ethical pluralism.

For conflicts involving fundamental values that remain unresolved through analytical methods (Step 5), the Structured Deliberation Framework (see Table A4 in Appendix A) operationalizes the procedural justice approach described in Section 4.3. Each procedural justice principle—publicity, relevance, inclusion, cooperative engagement, appeal and revision mechanisms, and enforcement—is translated into specific implementation practices, enhancing procedural legitimacy and democratic acceptability.

Finally, the Final Decision Matrix (see Table A5 in Appendix A) synthesizes empirical, normative, and procedural considerations to support justified AI governance decisions (whether implementation, modification, or rejection). Each decision includes comprehensive rationales, mitigation strategies for accepted harms, monitoring and review mechanisms, and clearly defined documentation practices.

5.3 | Illustrative Application: The Dutch Childcare Benefits Scandal

Between 2014 and 2020, the Dutch government used the Systeem Risico Indicatie (SyRI), a statutory framework for welfare-fraud detection that linked data from multiple agencies. In February 2020, the Hague District Court ruled that SyRI violated Article 8 of the European Convention on Human Rights by failing to balance fraud prevention against privacy rights (Van Bekkum and Borgesius 2021). Excessive data use, lack of transparency, weak safeguards, and discrimination risks were central concerns. The later childcare benefits scandal—in which thousands of families were wrongly accused of fraud—demonstrated how such shortcomings can result in severe harms (Lighthouse Reports, 2021). Taken together, these cases show how the five-step framework (Section 5.1) can be applied under time and data constraints.

The first step involves mapping harms and benefits. Expected benefits included improved fraud detection, protection of public funds, and deterrence of abuse. The harms, however, were significant: large-scale privacy intrusions, disproportionate targeting of low-income areas, stigmatization, and—in the childcare scandal—financial ruin, distress, and loss of trust in government. These can be measured with PANAS, SWLS, and QEWB scales and analyzed using the Multi-dimensional Impact Assessment Matrix (Appendix A, Table A1). Baseline well-being and performance data should be compared with post-deployment indicators, stratified by demographic group. Even in urgent contexts, rapid assessments using existing data can identify major harm–benefit patterns.

The second step weighs harms and benefits through the Ethical Weight Assessment Tool (Appendix A, Table A2). The harms were neither inevitable nor redundant, as they stemmed directly from opaque processes. Hypothetical scoring—for example, severity 5, inevitability 4, nonredundancy 5—produces a high ethical concern. According to the Reason Against Harming and Harm–Benefit Asymmetry principles, such predictable and serious harms outweigh modest efficiency gains.

In the third step, compliance is assessed using the Trustworthy AI Compliance Matrix (Appendix A, Table A3). SyRI failed on fairness (e.g., use of nationality proxies), transparency (secret indicators and models), and accountability (absence of project-level DPIAs). Similar flaws—opaque models, proxy variables, lack of appeals—marked the childcare scandal. Missing calibration data and error rates further undermined compliance. Both systems would therefore be rated noncompliant, requiring suspension or redesign.

The fourth step addresses conflict resolution with the Analytical Resolution Flowchart (Appendix A, Figure A1). By the Dominance Principle, the original designs are excluded, as alternatives such as targeted audits or human-in-the-loop screening performed at least as well without discriminatory effects. Supervaluationism reinforces that discriminatory targeting is unacceptable under any reasonable interpretation of fairness. If multiple redesigned options remain—such as models excluding sensitive attributes and incorporating confidence scores—the Maximality Principle guides the choice of an option not clearly inferior to others.

Finally, procedural deliberation ensures that unresolved conflicts are handled justly. The Structured Deliberation Framework (Appendix A, Table A4) requires participation by affected individuals, advocacy groups, experts, and policymakers.

Publicity and relevance conditions mandate that design choices, harms, and rationales be publicly disclosed. Inclusion and fair-cooperation conditions ensure marginalized groups are heard and technical information is explained in accessible terms. Appeal and revision mechanisms allow individuals to contest decisions, while independent oversight enforces compliance with procedural standards.

This case illustrates how the framework translates moral theory into administrative practice. By integrating harm–benefit analysis, compliance assessment, and procedural justice into governance tools, it provides a defensible method for determining when AI systems should be reformed or withdrawn and for guiding the development of fairer, more transparent alternatives.

6 | Discussion

The framework developed here contributes to ethical AI governance in public decision-making by linking philosophical theory with administrative practice. Several considerations warrant discussion: institutional barriers, methodological and conceptual issues, and future development.

Institutional and political barriers can undermine effectiveness if not addressed. Cognitive bias, especially automation bias, threatens Steps 3 (Technical Assessment) and 5 (Procedural Deliberation), where human judgment is essential. In welfare eligibility systems, case workers may accept AI outputs despite contradictory evidence. Mitigation strategies include staff training, structured "challenge points," and human-in-the-loop designs requiring explicit justification for algorithmic decisions. Institutional inertia may also impede Step 4 (Conflict Resolution) when replacing familiar systems is required. Phased implementation, integration with existing tools (e.g., DPIAs), and leadership buy-in can counter this resistance. Resource constraints—limited staff, expertise, and time—may hinder full application of all five steps, especially in small or politically sensitive agencies. Mitigations include prioritizing high-risk systems, using streamlined harm–benefit assessments, and drawing on external expertise. Political feasibility further complicates Step 5: ethical recommendations may conflict with political priorities. Aligning them with broader policy goals, framing in terms of long-term public value, and building consensus can improve adoption.

Positioning the framework relative to existing governance tools clarifies its contribution. It aligns with instruments such as the EU AI Act, DPIAs, AIAs, OECD AI Principles, and the G7 Toolkit, sharing commitments to transparency, accountability, fairness, and risk mitigation. It overlaps with compliance tasks (e.g., legality checks, documenting purposes, identifying discriminatory impacts) and participatory elements like stakeholder consultation. Its distinctiveness lies in four features: (1) anchoring in philosophical theories of harm, benefit, decision-making, and procedural justice; (2) operationalizing harm–benefit asymmetry through multidimensional well-being measures and weighted calculations; (3) embedding conflict-resolution methods (dominance, supervaluationism, maximality); and (4) extending procedural safeguards to enforce publicity, relevance, inclusion, and appeal. Rather than *replacing* existing tools, the framework *strengthens* their normative depth and legitimacy while providing administrators with structured methods for value-laden decisions.

Methodological and conceptual considerations also arise. Operationalizing harm and benefit through measures such as PANAS, SWLS, and QEWB ensures multidimensionality but raises challenges: resource-intensive data collection, adaptation for administrative contexts, and contextual adjustments of the harm–benefit coefficient (1.5). Agencies should refine this parameter through reflective equilibrium. Addressing value conflicts, the framework integrates analytic methods with procedural safeguards, balancing rigor with feasibility. However, in polarized contexts lacking shared democratic commitments, common ground must be established first. On distributive justice, while the framework stratifies impacts and prioritizes harm prevention, it does not fully resolve allocation questions. Further integration of distributive theories such as luck egalitarianism or prioritarianism could strengthen this dimension.

The framework's impact depends on *institutional capacity*, *political will*, and *adaptability*. Administrators need technical expertise and ethical reasoning skills, requiring investment and organizational support. Transparency may face resistance in sensitive areas like national security or commercial domains, necessitating adaptive governance. As AI evolves, unanticipated harms and benefits will emerge. Appeal, revision, and continuous improvement mechanisms enable responsiveness, but theoretical refinement must continue. Despite challenges, the framework advances current approaches by combining philosophical sophistication with practical tools. It operationalizes concepts such as multidimensional well-being, harm–benefit asymmetry, and structured conflict resolution into processes compatible with existing governance. Its modular, step-by-step design offers administrators matrices, decision rules, and safeguards adaptable to varied contexts. Documentation, transparency, and engagement build trust and support iterative improvement.

Finally, the illustrative case in 5.3 demonstrates application but *does not validate* the framework. Validation should follow a mixed-methods program: (1) pilot studies in partner agencies using pre-registered stepped-wedge or difference-in-differences designs with outcome and process metrics; (2) simulations and red-team exercises stress-testing Steps 2–5 under constraints; and (3) comparative case studies across domains using process-tracing. Reliability testing (Cohen's κ, ICC), sensitivity analyses of coefficients, and iterative practitioner input (surveys, focus groups, Delphi panels) will support refinement. External validity requires published protocols, anonymized artifacts, pre-registration, and monitoring dashboards with review triggers. These methods provide a pathway to test, revise, and improve the framework.

7 | Conclusion

This paper addresses a critical gap in ethical AI governance by developing a more comprehensive framework for assessing and weighing harm and benefit in public decision-making. Drawing on advances in moral philosophy and trustworthy AI research,

we translate key normative concepts into tools suitable for administrative use. The framework makes three main contributions. First, it provides a theoretically grounded method for identifying and measuring AI's impacts across multiple dimensions of well-being. Second, it offers a structured approach to normative evaluation. Third, it presents a coherent strategy for resolving value conflicts.

However, this framework is a starting point, not a final solution. As AI technologies evolve and their societal implications become clearer, continued refinement will be necessary. Future research should focus on empirical validation, development of domain-specific applications, and integration with broader regulatory regimes. Ethical AI governance in the public sector requires both philosophical depth and practical feasibility. By combining insights from ethics, public administration, and AI research, this framework supports that ambition. As AI becomes more embedded in public decision-making, such integrated approaches will be essential for ensuring that these systems uphold public values, respect human dignity, and promote collective well-being in democratic societies.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

Endnotes

[1] Proponents of counterfactual views include Boonin (2014), Feit (2023), and Klocksiem (2012, 2022). Proponents of causal views include Harman (2009) and Smuts (2012). Alternative views include Johansson and Risberg's (2023) idea that "negatively affecting" cannot be analyzed further in either counterfactual or causal terms alone; instead, they suggest that it may best be understood particularistically or pluralistically.

References

Algander, P. 2013. Harm, Benefit, and Non-Identity. Doctoral dissertation, Uppsala University.

Anderson, E. 1999. "What Is the Point of Equality?" *Ethics* 109, no. 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., and A. Herlitz. 2022. *Value Incommensurability: Ethics, Risk, and Decision-Making*. Taylor & Francis.

Andreou, C. 2016. "Dynamic Choice." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Stanford University.

Beauchamp, T. L., and J. F. Childress. 2019. *Principles of Biomedical Ethics*. 8th ed. Oxford University Press.

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.

Boonin, D. 2014. *The Non-Identity Problem and the Ethics of Future People*. Oxford University Press.

Bradley, B. 2012. "Doing Away With Harm." *Philosophy and Phenomenological Research* 85, no. 2: 390–412. https://doi.org/10.1111/j.1933-1592.2012.00615.x.

Bradley, B. 2013. "Asymmetries in Benefiting, Harming and Creating." *Journal of Ethics* 17, no. 1–2: 37–49. https://doi.org/10.1007/s10892-012-9134-6.

Brandstedt, E., and B. Brülde. 2019. "Towards a Theory of Pure Procedural Climate Justice." *Journal of Applied Philosophy* 36, no. 5: 785–799.

Broome, J. 1991. *Weighing Goods: Equality, Uncertainty, and Time*. Blackwell.

Brülde, B. 1998. *The Human Good*. University of Gothenburg.

Busuioc, M. 2021. "Accountable Artificial Intelligence: Holding Algorithms to Account." *Public Administration Review* 81, no. 5: 825–836.

Chang, R. 2002. *Making Comparisons Count*. Routledge.

Cox, L. A., Jr. 2008. "What's Wrong With Risk Matrices?" *Risk Analysis* 28, no. 2: 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x.

Crisp, R. 2017. "Well-Being." In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. Stanford University. https://plato.stanford.edu/archives/fall2017/entries/well-being/.

Daniels, N. 1996. *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge University Press.

Daniels, N., and J. E. Sabin. 2002. *Setting Limits Fairly: Can We Learn to Share Medical Resources?* Oxford University Press.

de Fine Licht, K. 2025. "Resolving Value Conflicts in Public AI Governance: A Procedural Justice Framework." *Government Information Quarterly* 42, no. 2: 102033.

de Fine Licht, K., and J. Fine Licht. 2020. "Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy." *AI & Society* 35: 917–926.

Diener, E., R. A. Emmons, R. J. Larsen, and S. Griffin. 1985. "The Satisfaction With Life Scale." *Journal of Personality Assessment* 49, no. 1: 71–75.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Feinberg, J. 1984. *Harm to Others*. Oxford University Press.

Feit, N. 2019. "Harming by Failing to Benefit." *Ethical Theory and Moral Practice* 22, no. 4: 809–823. https://doi.org/10.1007/s10677-017-9838-6.

Feit, N. 2023. *Bad Things: The Nature and Normative Role of Harm*. 1st ed. Oxford University Press.

Fine, K. 1975. "Vagueness, Truth, and Logic." *Synthese* 30, no. 3–4: 265–300. https://doi.org/10.1007/BF00485047.

Floridi, L., J. Cowls, M. Beltrametti, et al. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28, no. 4: 689–707.

Folland, A. 2023. "Feit on the Normative Importance of Harm." *Theoria* 89, no. 2: 176–187. https://doi.org/10.1111/theo.12453.

Folland, A. 2025a. Harm: Essays on its Nature and Normative Significance. Doctoral dissertation, Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-543566.

Folland, A. 2025b. "Doing Away With Skepticism About Harm." *Ethical Theory and Moral Practice* 28: 93–110. https://doi.org/10.1007/s10677-024-10480-x.

Fraser, N. 2009. *Scales of Justice: Reimagining Political Space in a Globalizing World*. Vol. 31. Columbia University Press.

Gardner, M. 2017. "On the Strength of the Reason Against Harming." *Journal of Moral Philosophy* 14, no. 1: 73–87. https://doi.org/10.1163/17455243-46810043.

Gardner, M. 2021. "What Is Harming?" In *Principles and Persons*, edited by J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan. Oxford University Press. https://doi.org/10.1093/oso/9780192893994.003.0018.

Habermas, J. 1985. *The Philosophical Discourse of Modernity: Twelve Lectures*. Polity Press.

Harman, E. 2009. "Harming as Causing Harm." In *Harming Future Persons*, edited by M. A. Roberts and D. T. Wasserman, 137–154. Springer. https://doi.org/10.1007/978-1-4020-5697-0_7.

Herlitz, A. 2019. "Nondeterminacy, Two-Step Models and Justified Choice." *Ethics* 129, no. 2: 284–308.

Herlitz, A. 2020. "Nondeterminacy, Cycles and Rational Choice." *Analysis* 80, no. 3: 443–449.

Herlitz, A. 2024. "Incommensurability and Healthcare Priority Setting." *Philosophical Studies* 181, no. 12: 3347–3365. https://doi.org/10.1007/s11098-024-02160-4.

Herlitz, A., and K. Sadek. 2021. "Social Choice, Nondeterminacy, and Public Reasoning." *Res Philosophica* 98, no. 3: 377–401.

High-Level Expert Group on AI. 2019. *Ethics Guidelines for Trustworthy Artificial Intelligence*. European Commission.

Holtug, N. 2002. "The Harm Principle." *Ethical Theory and Moral Practice* 5, no. 4: 357–389.

Hurka, T. 1993. *Perfectionism*. Oxford University Press.

Johansson, J., and O. Risberg. 2023. "A Simple Analysis of Harm." *Ergo* 9: 509–536. https://doi.org/10.3998/ergo.2275.

Kaur, D., S. Uslu, K. J. Rittichier, and A. Durresi. 2022. "Trustworthy Artificial Intelligence: A Review." *ACM Computing Surveys* 55, no. 2: 1–38.

Klocksiem, J. 2012. "A Defense of the Counterfactual Comparative Account of Harm." *American Philosophical Quarterly* 49, no. 4: 285–300.

Klocksiem, J. 2022. "Harm, Failing to Benefit, and the Counterfactual Comparative Account." *Utilitas* 34, no. 4: 428–444.

Loi, M., A. Herlitz, and H. Heidari. 2023. "Fair Equality of Chances for Prediction-Based Decisions." *Economics and Philosophy* 40: 1–580. https://doi.org/10.1017/S0266267123000342.

Mill, J. S. 1859/1991. *On Liberty*. Cambridge University Press.

Mill, J. S. 1977. "On Liberty." In *Collected Works of John Stuart Mill*, edited by J. M. Robson, vol. XVIII, 213–310. University of Toronto Press. Original work published 1859.

Nussbaum, M. C. 2011. *Creating Capabilities: The Human Development Approach*. Harvard University Press.

O'Neil, C. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

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* 15, no. 1: 25–40.

Pettit, P. 2012. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge University Press.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press.

Rawls, J. 1999. *A Theory of Justice (rev. ed.)*. Harvard University Press.

Reinhardt, K. 2023. "Trust and Trustworthiness in AI Ethics." *AI and Ethics* 3, no. 3: 735–744.

Ryberg, J., and T. S. Petersen. 2022. "Sentencing and the Conflict Between Algorithmic Accuracy and Transparency." In *Sentencing and Artificial Intelligence*, edited by J. Ryberg and J. V. Roberts. Oxford University Press. https://doi.org/10.1093/oso/9780197539538.003.0004.

Savage, L. J. 1954. *The Foundations of Statistics*. Wiley.

Sen, A. 1970. *Collective Choice and Social Welfare*. Holden-Day.

Sen, A. 1997. "Maximization and the Act of Choice." *Econometrica* 65, no. 4: 745–779.

Sen, A. 2000. "Consequential Evaluation and Practical Reason." *Journal of Philosophy* 97, no. 9: 477–502.

Shafer-Landau, R. 2021. *The Fundamentals of Ethics*. 5th ed. Oxford University Press.

Shiffrin, S. V. 2012. "Harm and Its Moral Significance." *Legal Theory* 18, no. 3: 357–398. https://doi.org/10.1017/S1352325212000080.

Smuts, A. 2012. "Less Good but Not Bad: In Defense of Epicureanism About Death." *Pacific Philosophical Quarterly* 93, no. 2: 197–227.

Tyler, T. R. 2006. *Why People Obey the Law*. Princeton University Press.

Van Bekkum, M., and F. Z. Borgesius. 2021. "Digital Welfare Fraud Detection and the Dutch SyRI Judgment." *European Journal of Social Security* 23, no. 4: 323–340.

Vogl, T. M., C. Seidelin, B. Ganesh, and J. Bright. 2020. "Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities." *Public Administration Review* 80, no. 6: 946–961. https://doi.org/10.1111/puar.13286.

Waterman, A. S., S. J. Schwartz, B. L. Zamboanga, et al. 2010. "The Questionnaire for Eudaimonic Well-Being: Psychometric Properties, Demographic Comparisons, and Evidence of Validity." *Journal of Positive Psychology* 5, no. 1: 41–61.

Watson, D., L. A. Clark, and A. Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54, no. 6: 1063–1070.

Wirtz, B. W., and W. M. Müller. 2019. "An Integrated Artificial Intelligence Framework for Public Management." *Public Management Review* 21, no. 7: 1076–1100.

Zanotti, G., M. Petrolo, D. Chiffi, and V. Schiaffonati. 2023. "Keep Trusting! A Plea for the Notion of Trustworthy AI." *AI & Society* 39: 1–2702.

Appendix A

This appendix provides the instruments referenced in Section 5.2. Each table/figure lists the framework step(s) where it applies (see Section 5.1 for the five-step procedure).

Identification and Measurement Tools

TABLE A1 | Multi-dimensional impact assessment matrix.

Well-being dimension	Potential impacts	Measurement tools	Data collection method
Hedonistic (emotional states)	• Stress/relief • Anxiety/comfort • Frustration/satisfaction	PANAS (Watson et al. 1988)	• Pre/post surveys • Experience sampling • Longitudinal tracking
Desire Fulfillment (preferences)	• Goal achievement • Service accessibility • Resource allocation	SWLS (Diener et al. 1985)	• Satisfaction surveys • Usage statistics • Preference matching
Objective List (meaningful life)	• Autonomy/agency • Social connection • Personal development	QEWB (Waterman et al. 2010)	• Structured interviews • Community feedback • Case studies

Note: Where used: Step 1—Identification and Measurement (Section 5.1); introduced in Section 5.2; referenced in Section 5.3 (Illustrative Application).
Note on measures: PANAS (Watson et al. 1988), SWLS (Diener et al. 1985), QEWB (Waterman et al. 2010). Agencies may substitute equivalent validated instruments where appropriate and record any substitutions in the project file.

Implementation Guidelines:

1. Establish baseline measurements before AI implementation
2. Identify both direct impacts (immediate AI decisions) and indirect impacts (systemic effects)
3. Stratify impact analysis by demographic groups to identify disparate effects
4. Document uncertainty through confidence intervals and sensitivity analysis

Normative Evaluation Framework

TABLE A2 | Ethical weight assessment tool (scales and aggregation).

Impact	Severity (1–5)	Inevitability (1–5)	Redundancy (1–5)	Ethical weight	Priority level
Harm A	4 (High)	2 (Mostly avoidable)	1 (Unique to AI)	High	Critical
Benefit B	3 (Moderate)	4 (Likely regardless)	3 (Partially redundant)	Low-moderate	Secondary

Note: Where used: Step 2—Normative Evaluation (Section 5.1); introduced in Section 5.2; sensitivity analyses discussed in Section 6.
Scoring scales: Severity (1–5), Inevitability (1–5), Redundancy (1–5, reverse-scored).
Aggregation: Ethical Weight_Harm = f(Severity, Inevitability, Nonredundancy); Ethical Weight_Benefit = f(Severity, Inevitability, Nonredundancy) ÷ 1.5 (harm–benefit asymmetry coefficient).
Tie-break rule: If final Ethical Weights differ by ≤ 5%, prefer the option with the lower expected harm; if still tied, escalate to Figure A1 (Analytical Resolution Flowchart).

Scoring key:

- Severity: 1 = minimal, 5 = severe
- Inevitability: 1 = certain to occur regardless, 5 = fully avoidable,
- Redundancy: 1 = would happen anyway, 5 = unique to this intervention

Ethical Weight Calculation:

- For harms: (Severity × (6-Inevitability) × (6-Redundancy))
- For benefits: (Severity × (6-Inevitability) × (6-Redundancy)) ÷ 1.5 (applying harm-benefit asymmetry)

Implementation Guidelines:

1. Score each impact independently, then validate through team consensus
2. Document reasoning for each rating with reference to specific evidence
3. Re-evaluate periodically as new information becomes available
4. Pay special attention to high-severity harms, even if they have low probability

Technical Assessment Checklist

TABLE A3 | Trustworthy AI compliance matrix (criteria, evidence, ratings).

Criterion	Assessment question	Evidence required	Compliance level
Performance	Does the system achieve its stated purpose with acceptable accuracy?	• Accuracy metrics • Error rates • Validation studies	◦ Full ◦ Partial ◦ Noncompliant
Explainability	Can decisions be explained in terms understandable to affected individuals?	• Explanation methods • User testing • Stakeholder feedback	◦ Full ◦ Partial ◦ Noncompliant
Fairness	Does the system distribute errors and benefits equitably across groups?	• Disparity metrics • Disaggregated performance • Bias audits	◦ Full ◦ Partial ◦ Noncompliant

Note: Where used: Step 3—Technical Assessment (Section 5.1); introduced in Section 5.2; informs Step 4 trade-offs.
Minimum gates: No deployment if any criterion is rated "Noncompliant." Full compliance is required for legality, fairness (pre-specified disparity thresholds), and appealability. All remaining criteria must be ≥ "Partial" and accompanied by a time-bound remediation plan approved by governance.

Implementation Guidelines:

1. Conduct technical assessment in parallel with impact identification
2. Document methods used to achieve each criterion
3. Identify trade-offs between criteria (e.g., accuracy vs. explainability)
4. Establish minimum compliance thresholds for critical applications

Conflict Resolution Decision Tree

TABLE A4 | Structured deliberation framework (publicity, relevance, inclusion, fair cooperation, appeal and revision, enforcement).

Procedural principle	Implementation mechanism	Documentation requirements
Publicity	• Public hearings • Published impact assessments • Open decision records	• Meeting minutes • Public notices • Accessible documentation
Relevance	• Structured reasoning template • Evidence standards • Expert consultation	• Citations to evidence • Reasoning chains • Expert opinions
Inclusion	• Stakeholder mapping • Diverse representation • Multiple engagement formats	• Participant demographics • Outreach efforts • Accessibility accommodations
Fair cooperation	• Facilitated dialogue • Structured turn-taking • Educational supports	• Process rules • Educational materials • Facilitation protocols
Appeal and revision	• Appeal mechanisms • Regular review cycles • Feedback channels	• Appeal procedures • Review schedules • Decision modification logs
Enforcement	• Oversight bodies • Monitoring protocols • Compliance audits	• Compliance reports • Enforcement actions • Remediation plans

Note: Where used: Step 5 — Procedural Deliberation (Section 5.1); introduced in Section 5.2.
Implementation note: For each principle, record (i) concrete actions taken, (ii) participants and evidence considered, and (iii) outcomes/commitments (including timelines and owners).

Implementation Guidelines:

1. Document each elimination decision with explicit reasoning
2. Consider sensitivity analysis to test robustness of decisions
3. Maintain transparency about the resolution process
4. Apply consistently across similar decision contexts

TABLE A5 | Final decision matrix (synthesis and disposition).

Option	arm-Benefit profile	Technical compliance	Procedural legitimacy	Decision	Justification
Implement as proposed	• Moderate harms • High benefits • Ethical weight positive	• Full: 6/10 criteria • Partial: 3/10 criteria • Noncompliant: 1/10	• Strong stakeholder support • Transparent process • Appeal mechanism in place	Implement with modifications	Significant benefits outweigh moderate harms; technical weaknesses can be addressed through proposed modifications; strong procedural legitimacy
Alternative approach	• Low harms • Low benefits • Ethical weight neutral	• Full: 8/10 criteria • Partial: 2/10 criteria • Noncompliant: 0/10	• Limited stakeholder engagement • Transparent process • Appeal mechanism in place	Reject	Despite technical strengths, limited benefits do not justify even low harms; procedural weaknesses in stakeholder engagement
No AI implementation	• No AI-related harms • No AI-related benefits • Status quo preserved	N/A	• Strong stakeholder opposition to status quo • Transparent process • Appeal mechanism in place	Reject	Status quo produces unacceptable harms; some AI implementation necessary to address current shortcomings

Note: Where used: Decision synthesis after Steps 1–5 (Sections 5.1–5.2); referenced in Section 5.3.
Usage note: Summarize option scores (A1–A3), elimination reasoning (Figure A1), procedural findings (A4), and the final disposition (implement/modify/suspend/withdraw) with mitigation, monitoring, and review triggers.

Implementation Guidelines:

1. Scale deliberation process to risk and significance of decision
2. Document all stakeholder input, including dissenting views
3. Explicitly connect deliberative outputs to final decisions
4. Create institutional memory to ensure consistency across cases

Integration and Decision Support

Implementation Guidelines:

1. Document comprehensive rationale for final decision
2. Develop mitigation strategies for any accepted harms
3. Establish monitoring mechanisms to validate expected outcomes
4. Set review periods to reassess decisions based on observed impacts

Practical Application Guide

To implement this framework effectively, public administrators should:

1. Prepare the organizational context:
   - Develop institutional capacity for ethical assessment
   - Train staff on framework application
   - Establish governance structures (e.g., ethics committees)
2. Scale application appropriately:
   - High-risk applications require complete framework implementation
   - Lower-risk applications may use streamlined assessment
   - Develop criteria for determining required assessment depth
3. Ensure documentation quality:
   - Maintain comprehensive records of the assessment process
   - Document reasoning behind key decisions
   - Create accessible summaries for public communication
4. Implement continuous improvement:
   - Establish feedback mechanisms to refine the framework
   - Conduct periodic reviews of framework effectiveness

1. **Identify Conflict Type:**
    - Between dimensions of well-being
    - Between trustworthy AI criteria
    - Between harm prevention and benefit promotion
2. **Apply Dominance Principle:**
    - Is option A clearly superior to option B on all relevant dimensions?
    - If yes → eliminate dominated option
    - If no → proceed to next step
3. **Apply Supervaluationism:**
    - Does option A perform better than option B across all plausible interpretations of the contested value?
    - If yes → eliminate consistently inferior option
    - If no → proceed to next step
4. **Apply Maximality:**
    - Is option A determinately worse than any available alternative on at least one dimension?
    - If yes → eliminate option A
    - If no → retain option A in consideration set
5. **Remaining Options:**
    - If single option remains → implement
    - If multiple options remain → proceed to procedural deliberation

FIGURE A1 | Analytical resolution flowchart (Dominance → Supervaluationism → Maximality). Where used: Step 4—Conflict Resolution (Section 5.1); introduced in Section 5.2; invoked in Section 5.3 when excluding inferior design options.

- Share lessons learned across public administration contexts

This operationalized framework provides public administrators with concrete tools to translate ethical principles into practice. By systematically addressing identification, evaluation, technical assessment, conflict resolution, and procedural deliberation, the framework enables responsible AI governance that balances harm prevention with benefit promotion while upholding democratic values.