Examining AI Governance Assumptions in Social Work: A Triple Mandate Perspective

Eunhye Ahn [1] **Moon Choi** [2 3] **Claire R. McNellan** [4] **Lauri Goldkind** [5]

Abstract

AI systems deployed in social work must operate within the field's distinctive characteristics, such as involuntary client engagement, life-altering and often irreversible decisions, relational practice, professional discretion amid moral uncertainty, and work with structurally marginalized populations. Yet whether existing AI governance frameworks, developed largely for private-sector and routine public-sector contexts, can be effectively applied in social work settings remains under-examined. We identify six operating conditions that prevailing AI governance frameworks presuppose, show how each misaligns with social work value in ways that compromise algorithmic accountability, and show how AI deployed without attention to this context can compound these failures. Drawing on Staub-Bernasconi's triple mandate, we propose a minimum accountability layer organized around three mandate domains (client, organizational, and professional) across the accountability requirements of transparency, contestability, and redress. Our central argument is that effective AI governance must preserve the professional mandate as the independent ethical fulcrum mediating between organizational power and client rights. This work contributes to an emerging conversation about adapting AI governance knowledge to relational, high-stakes service contexts.

[1]School of Human Ecology, University of Wisconsin-Madison, Madison, WI, USA [2]Graduate School of Science and Technology Policy, Korea Advanced Institute of Science and Technology, Daejeon, South Korea [3]Graduate School of Data Science, Korea Advanced Institute of Science and Technology, Daejeon, South Korea [4]Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA [5]Graduate School of Social Service, Fordham University, New York, NY, USA. Correspondence to: Eunhye Ahn <eunhye.ahn@wisc.edu>.

Accepted at the Trustworthy AI for Good (AI4GOOD) Workshop, ICML 2026, Seoul, South Korea. Non-archival; copyright held by the author(s).

1. Introduction

AI systems are increasingly deployed across public-sector contexts, informing decisions about risk assessment, resource allocation, and service delivery (Afrouz & Lucas, 2023; Lee et al., 2024; Park et al., 2025). Alongside this expansion, a growing body of work on AI governance has proposed frameworks for ensuring such systems operate in ways that are fair, transparent, and accountable (Jobin et al., 2019; Corrêa et al., 2023; Wirtz et al., 2022). These frameworks reflect considerable effort to translate abstract ethical principles into practical mechanisms for oversight and accountability. As deployment expands into an increasingly diverse set of service domains, however, an open question is how well these frameworks, developed primarily with private-sector or routine public-administration contexts in mind, translate into domains with fundamentally different operating conditions.

Social work is one such domain. It operates across the United States through more than 800,000 positions, with state and local government among the largest employers (Bureau of Labor Statistics, 2025), and it is increasingly a site where AI systems inform consequential decisions: predictive risk modeling in child welfare, automated eligibility determination in benefits programs, and algorithmic triage in service referrals (Báez et al., 2026; Lee et al., 2024; Park et al., 2025). Several features distinguish social work from other public services in ways that matter for governance design. Clients often engage involuntarily, compelled by law, institutional mandate, or the absence of alternatives (Rooney & Mirick, 2018; Redden et al., 2020). Decisions are typically high-stakes and often irreversible, such as removing a child from a family, approving essential benefits, or mandating treatment, with consequences that propagate across a lifetime (Gillingham, 2019; Taylor, 2021). Practice is fundamentally relational, with human relationships serving as the core mechanism through which change occurs rather than a channel for service delivery (Devlieghere et al., 2022; Forenza & Eckert, 2018). Professionals exercise significant discretion as street-level bureaucrats, navigating genuinely irresolvable moral dilemmas under intense public scrutiny (Berrick, 2018; Lipsky, 2010; Ahn et al., 2025b). And the populations served are disproportionately affected

by poverty, racial marginalization, and structural inequality (Fong, 2017; Roberts, 2014).

High-profile deployment failures suggest what is at stake when those conditions are not accounted for. Michigan's MiDAS system wrongfully accused tens of thousands of people for unemployment fraud at an extraordinary error rate, stripping them of benefits without adequate notice or opportunity to respond (Charette, 2018). Australia's Robodebt scheme (Mao, 2023) and the Dutch childcare benefits affair (Amnesty International, 2021) produced comparable patterns: wrongful accusations, housing loss, family separation. These failures do not appear to be simply matters of false positives—an error common to many predictive systems; rather, they suggest that accountability mechanisms suited to general market or routine public-service contexts may not deliver equivalent protections when transplanted into high-stakes, often irreversible service domains. The same error carries far greater weight here, because the populations affected are structurally marginalized and frequently have no alternatives but to remain in the system. Whether existing AI governance frameworks can effectively address these dynamics in domains like social work is a question that has received limited systematic attention (David, 2024; Wang et al., 2023).

In this paper, we examine how existing AI governance knowledge can be adapted to social work, a domain shaped by distinct ethical obligations, marginalized populations, and practitioner discretion. Our aim is not to reject AI in social work nor to propose a comprehensive alternative framework, but to offer a diagnostic and a starting point for adaptation. We first identify six operating conditions that recur across existing AI governance frameworks, and examine where and why these conditions break down in social work's distinctive context. Drawing on Staub-Bernasconi's (Staub-Bernasconi, 2007; 2009; 2016) triple mandate as a professional accountability structure, we then propose a set of reflective guiding questions organized around the three mandate domains (client, organizational, and professional) intersecting with accountability requirements of transparency, contestability, and redress. We offer these questions not as a comprehensive governance framework but as a baseline set of accountability conditions that any approach applied in social work should address, and more broadly as an example of how sector-specific conditions can inform governance design.

2. AI Governance: Key Elements and Operating Conditions

2.1. Definition and key elements

Among many efforts to define AI governance, the most prominent is "a system of rules, practices, processes, and technological tools that are employed to ensure an organization's use of AI technologies aligns with the organization's strategies, objectives, and values; fulfills legal requirements; and meets principles of ethical AI followed by the organization" (Mäntymäki et al., 2022, p. 604). Operationalized through tools, norms, standards, and regulations (Butcher & Beridze, 2019; Corrêa et al., 2023; Jobin et al., 2019), AI governance frameworks commonly converge around four central themes: technology, stakeholders and context, regulation, and processes (Birkstedt et al., 2023; Mäntymäki et al., 2022). Among these, accountability warrants particular attention, as AI governance frameworks are, at their core, mechanisms for ensuring that those who build and deploy AI systems can be held answerable for their outcomes (Raji et al., 2020).

2.2. Operating conditions of AI governance frameworks

While a growing body of literature examines AI governance in the public sector, including human services (Kuziemski & Misuraca, 2020; Robles & Mallinson, 2025; Wang et al., 2023), much of this work implicitly assumes conditions that prevail in private-sector services or routine public administration. Reading across widely cited frameworks (Gasser & Almeida, 2017; Wirtz et al., 2020; Mäntymäki et al., 2022; 2023) and systematic reviews of this literature (Jobin et al., 2019; Corrêa et al., 2023; Birkstedt et al., 2023; Batool et al., 2025; Robles & Mallinson, 2025), we identify six operating assumptions these frameworks repeatedly take for granted, presenting them as recurring rather than as an exhaustive taxonomy.

First, perhaps the most widely shared assumption is a techno-solutionist orientation (Morozov, 2013) that treats AI deployment as intrinsically desirable, centering on optimizing implementation rather than evaluating whether it is warranted (Attard-Frost & Lyons, 2025; Corrêa et al., 2023; Wirtz et al., 2022). Even frameworks that question the extent of human-machine collaboration focus on the degree of automation (Wang et al., 2023), rather than whether algorithmic decision-making is appropriate for the specific domain. This orientation works well where AI benefits are well-supported and automating judgment does not alter the fundamental nature of the service, as in traffic signal optimization, weather forecasting, or routine document processing.

Second, AI governance frameworks often assume that AI risks can be classified into distinct, predefined categories and addressed through structured, sequential mitigation (Wirtz et al., 2020; 2022). Frameworks operationalize this through dimensional risk taxonomies, modular governance layers mapped to technical, ethical, and legal domains (Gasser & Almeida, 2017), and lifecycle-based models (Batool et al., 2025; Mäntymäki et al., 2022; Raji et al., 2020). This works

where risk factors are sufficiently independent to be addressed in isolation, an assumption that holds for standardized tasks but becomes unstable where risks are entangled with human relationships and contextual judgment.

Third, improving predictive accuracy is widely treated as the path to better service outcomes (McCradden et al., 2025; Van Amsterdam et al., 2025). Embedded are three premises: outcomes of interest are accurately captured in available data, more accurate models lead to better decisions, and prediction does not itself shape the outcome being predicted (Kolt et al., 2025; Metcalf et al., 2021). These premises hold where data reliably measure what they claim to measure and where predictions inform decisions without reshaping the conditions they describe. They do not hold where data capture institutional behavior rather than the underlying phenomenon of interest (Farrell, 2025; Hu, 2025; McNellan et al., 2022), or where labeling someone as high-risk triggers responses that worsen the very outcome predicted.

Fourth, human oversight is treated as a primary safeguard, expecting humans to detect errors, exercise independent judgment, and override flawed algorithmic decisions (Mišić et al., 2025; Office of Management and Budget, 2024). This requires that AI systems provide explanations operators can understand and act on, that operators have time, expertise, and authority to evaluate outputs independently rather than defer, and that decision-making pace allows deliberation. These conditions erode where operators face high caseloads and time pressure (Kuziemski & Misuraca, 2020), and where AI systems remain opaque with model internals protected by intellectual property restrictions (Camilleri, 2024; David, 2024).

Fifth, many existing AI governance frameworks presuppose at the organizational level that institutions have, or can readily develop, the infrastructure to execute governance requirements, including dedicated AI governance teams, compliance mechanisms, and clear accountability lines (Mäntymäki et al., 2023). The U.S. OMB memorandum (Office of Management and Budget, 2024), for instance, requires each federal agency to designate a Chief AI Officer and convene agency-wide AI governance bodies. Yet these requirements assume organizational capacity many agencies cannot readily meet, widening disparities across the service landscape and leaving smaller, under-resourced agencies without realistic compliance means (Attard-Frost & Lyons, 2025).

Finally, governance frameworks often assume that affected populations have meaningful agency over engagement with AI systems, such that individuals can consent to or refuse AI-driven services, opt out without penalty, and appeal through institutional channels (Corrêa et al., 2023; Fjeld et al., 2020; Jobin et al., 2019). These conditions hold only where participation is voluntary and alternatives are available.

The above assumptions are not inherently flawed; they reflect the contexts in which AI governance frameworks were developed and serve important functional purposes within those contexts. Yet even the frameworks designed for the general public service context often carry forward the assumptions drawn on the private sector and administrative dimensions of the public sector, without examining whether they hold equally across all public sector domains, including social work.

3. Where Governance Assumptions Meet Social Work's Distinctive Context

Unlike the private sector, the public sector is accountable to citizens and legislators rather than shareholders, obligated to protect rights rather than maximize profit, and required to be transparent rather than permitted to withhold information as proprietary (Kuziemski & Misuraca, 2020; Mišić et al., 2025). Yet the public sector is not monolithic. Social work is distinguished even within the public sector by the involuntary nature of client engagement, life-altering high-stakes decisions, and the structural marginalization of populations served. While no single characteristic is entirely unique to social work, their combination creates a context where the operating conditions assumed by standard AI governance frameworks require re-examination. Figure 1 summarizes the five distinctive conditions we discuss below and the seven points of misalignment they produce with standard governance assumptions.

3.1. Involuntary engagement

Among public sector services, social work is distinguished by the involuntary nature of client engagement. Families under child protection investigation, individuals subject to court-ordered treatment, and people dependent on welfare benefits do not choose to engage but are compelled by law, institutional mandate, or the absence of alternatives, producing relationships defined by profound asymmetries of power (Redden et al., 2020; Rooney & Mirick, 2018). Unlike other public services where voluntariness varies but rarely reaches this extreme, social work clients cannot opt out without penalty, cannot choose alternative providers, and face legal consequences for disengagement. When the state controls access to vital resources, consent to data collection is not a genuine choice, as refusal risks losing essential resources or triggering legal sanctions (Garrett, 2025; Varon & Peña, 2021). This becomes particularly consequential when AI systems are introduced: those who cannot leave the system and lack the ability to remove their data have no means to prevent, or even be informed of, how information collected in the context of seeking help is repurposed for algorithmic risk profiling, transforming the relationship from support into surveillance (Keddell, 2019).

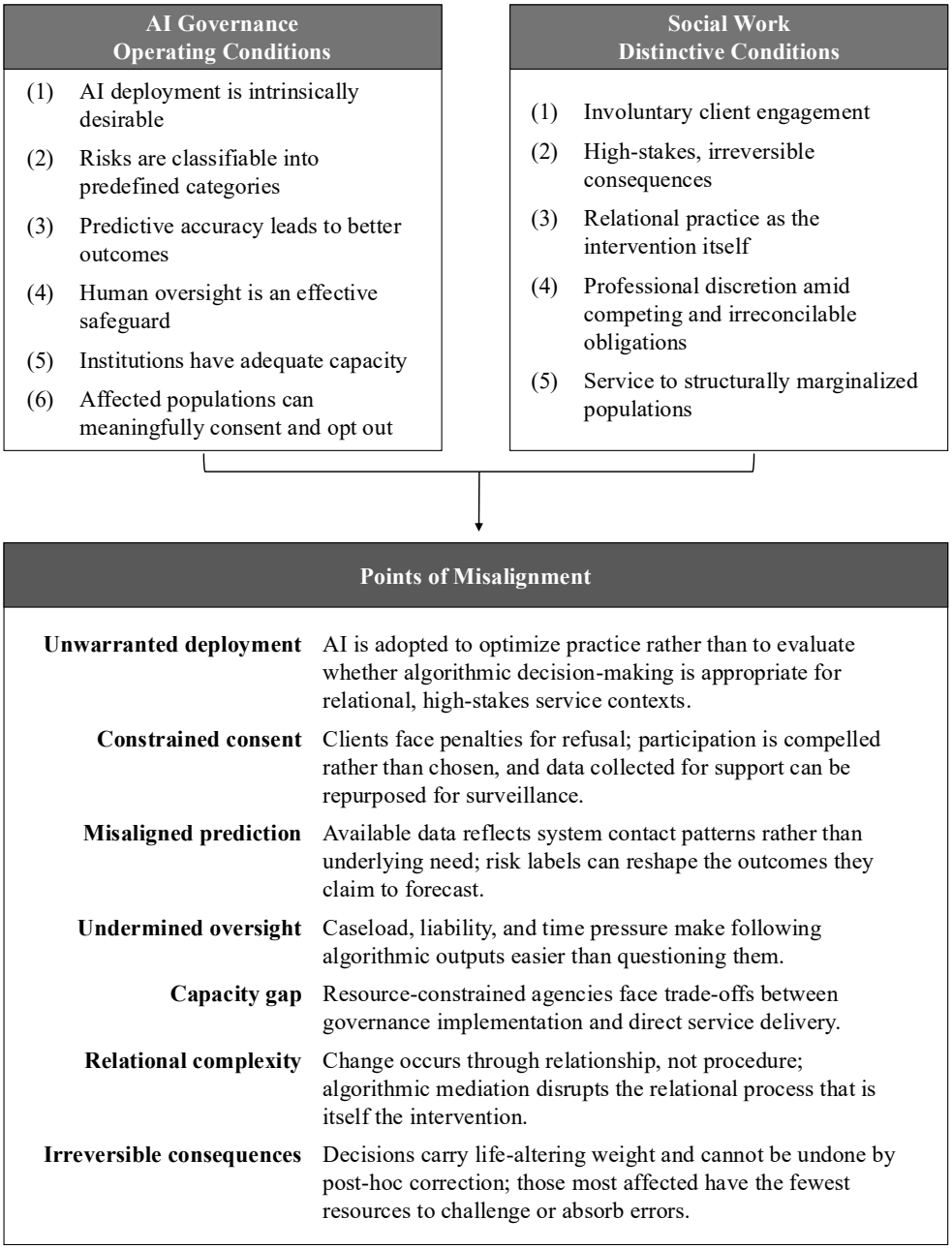

Figure 1. Where standard AI governance operating conditions (left) meet social work's distinctive conditions (right). The bottom panel summarizes seven points of misalignment that emerge when frameworks developed for other contexts are applied to relational, high-stakes service domains.

3.2. High-stakes, life-altering consequences

In social work, decision stakes are often life-altering; removing a child from a family, approving essential benefits, or mandating treatment carries profound consequences, even when warranted (Gillingham, 2019; Taylor, 2021). When a risk assessment errs, the cost is severe and often irreversible. Underestimating the risk of domestic violence can leave a victim exposed until it is too late to intervene, and failing to recognize a family's need can mean withholding the support that might have prevented lasting harm. But the deeper difficulty is that in social work, even a warranted decision can introduce harm of its own. As Berrick (2018) shows, child protection forces a choice between a child's safety and a family's integrity. Removing a child at risk can protect the child yet still inflict lasting trauma on the whole family, while leaving the child in place can preserve the family yet expose the child to harm.

3.3. Relational practice as the intervention itself

While public services vary in how relational they are, social work sits at the far end of this spectrum. The profession is grounded in the principle that human relationships are not merely a vehicle for service delivery but the core mechanism through which change occurs (Devlieghere et al., 2022; Forenza & Eckert, 2018; Mishna et al., 2021). Outcomes depend not only on what decision is reached but on how the client experiences the process of reaching it, which requires engaging with individuals as whole persons, reading nonverbal cues, interpreting emotional states, and drawing on trust built over time (Garrett, 2025; Gillingham & Graham, 2017). Within each interaction, empathy, assessment, intervention, and relationship maintenance occur simultaneously and cannot be decomposed into discrete, documentable actions. These characteristics sit uneasily with AI governance approaches built around classifiable data, standardizable procedures, and measurable accuracy.

3.4. Professional discretion amid uncertainty and public scrutiny

Unlike some public services where decisions follow codifiable rules with limited interpretation (e.g., a postal worker routes a package, a parking enforcement officer issues a citation), social work requires practitioners to exercise significant discretion in every encounter (Berrick, 2018; Ahn et al., 2025b). The problems social workers address, such as poverty, substance use, family crisis, and child safety, are complex, chronic, and rarely have a single correct resolution (Berrick, 2018; Lipsky, 2010). In child welfare, workers must decide whether to remove a child for safety, knowing that removal itself carries harm, under conditions where no solution may be clearly right at the point of decision (Berrick, 2018).

This discretion is exercised under intense public scrutiny. When decisions are associated with severe injury or death of a child, social workers may face not only institutional review and media criticism but even criminal charges, as when child welfare workers in Los Angeles faced prosecution following the death of Gabriel Fernandez (Etehad & Winton, 2017). Such scrutiny has been associated with defensive practice: over-documenting, intervening more than necessary, or refraining from intervening when service is required (Ahn et al., 2025b; Chenot, 2011; Whittaker & Havard, 2016). AI governance frameworks built for conditions where decisions can be standardized and optimized for accuracy offer little guidance where "right" decisions may not exist, and the practitioner bears consequences for getting it wrong.

3.5. Serving populations already affected by structural inequality

Social work disproportionately works with populations affected by poverty, racial marginalization, and structural inequality, whose circumstances are often compounded by heightened state surveillance and the lasting consequences of prior system involvement (Fong, 2017; Hailu et al., 2022; Roberts, 2014). Serving these populations is central to the profession's mission (National Association of Social Workers, 2021), and their concentration of structural disadvantage intensifies every condition above: involuntary engagement carries greater weight when those compelled to participate already lack resources to advocate for themselves; irreversible decisions are more devastating when there is no safety net to absorb their impact; and relational trust is harder to build when clients have been repeatedly failed by the systems meant to help them.

Beyond amplifying these conditions, working with structurally marginalized populations introduces distinct challenges for data-driven governance. Data generated through social work practice reflects not only genuine need but also patterns shaped by historical bias and disproportionate surveillance, including from adjacent systems such as criminal justice, where low-income and racial minority communities are overrepresented (Gillingham, 2019; Neil & Zanger-Tishler, 2025). The populations most affected by algorithmic systems are also the least positioned to participate in the governance processes that shape them or to contest decisions made about their lives (Garkisch & Goldkind, 2024; Ahn, 2026).

4. A Guiding Model for AI Governance in Social Work

4.1. From misalignment to accountability

Ensuring accountability has been recognized as a central purpose of AI governance in organizational and technical contexts (Busuioc, 2021; Raji et al., 2020), yet what this requires in social work practice remains largely underexamined (Brown et al., 2019). When AI systems shape decisions that affect fundamental rights and family integrity, accountability raises questions specific to this context: who is responsible for those decisions, to whom justification is owed, and through what mechanisms that responsibility can be enforced (Novelli et al., 2024). Accountability is a relational concept in which a decision-maker must explain and justify their conduct to those affected or to an oversight authority, which in turn can question, evaluate, and impose consequences (Bovens, 2007). For a meaningful accountability process, three conditions must be met: the decision-maker's conduct must be transparent enough to be examined (information); the reasoning behind decisions must be open to questioning and justification (explanation); and decision-makers must face the possibility of consequences, whether sanctions, required remediation, or redress, when an oversight authority judges their conduct to have fallen short (Bovens, 2007; Busuioc, 2021).

Where AI systems are integrated into social work, each accountability requirement is compromised in ways that follow directly from the misalignments identified above. Information is constrained because involuntary clients who cannot opt out and lack meaningful agency often have no access to how algorithmic outputs shape decisions about their lives. Explanation is undermined because system opacity, compounded by caseload pressure and liability fear, makes it difficult to produce or demand meaningful justification for algorithmic recommendations. Consequences are structurally compromised at two levels: oversight authorities rarely have sufficient access to how algorithmic systems reach their conclusions, and the populations most affected are the least positioned to seek redress through institutional appeals, a structural exclusion that extends to the fundamental question of who has standing to contest decisions made about their lives (Garkisch & Goldkind, 2024; Ahn, 2026).

4.2. The triple mandate as a professional accountability framework

Addressing these accountability failures requires understanding how professional obligation in social work is structured. The triple mandate—simultaneous obligations to the client, the employing organization or state, and the profession itself (Staub-Bernasconi, 2007; 2009; 2016)—is a foundational feature of professional identity that remains essential in the AI era (Garkisch & Goldkind, 2024). What makes it significant is not the complications posed by having three dimensions, but the distinctive role of the third. Without the professional mandate, social work reduces to a double mandate where the practitioner becomes either an instrument of the state or an advocate for the client, with no independent ethical standing to mediate between the two. It is the professional mandate, grounded in ethical principles and disciplinary knowledge, that provides the normative basis to push back against both organizational directives and client preferences when professional judgment demands it (Staub-Bernasconi, 2009; 2016).

In practice, these three obligations frequently pull in opposing directions (Staub-Bernasconi, 2009; 2016), and AI shifts the balance of power within each tension. Between client and organization, organizational mandates demand compliance, efficiency, and risk management, while the client mandate demands self-determination, dignity, and minimal intervention. AI intensifies this tension by equipping organizations with seemingly objective justification for intervention that involuntary clients have virtually no capacity to contest (Garkisch & Goldkind, 2024). Between client and practitioner, professional expertise and client self-determination can already conflict; AI compounds this by layering algorithmic authority onto professional judgment, further reducing the client's space to dispute decisions (Busuioc, 2021). Between practitioner and organization, professionals require discretion, contextual judgment, and time for relationships, while organizations seek standardization and efficiency. AI can create an asymmetric structure where following an algorithmic recommendation requires no justification, while overriding it demands documentation, explanation, and exposure to scrutiny (Busuioc, 2021).

Across all three tensions, the consistent pattern is the systematic weakening of the professional mandate. AI strengthens organizational accountability through compliance metrics and efficiency measures while simultaneously eroding professional discretion, relational judgment, and ethical mediation on which the professional mandate depends (Bovens, 2007; Busuioc, 2021). As that mediating role is undermined, the triple mandate risks collapsing into a double mandate, leaving client and organization in direct, unmediated opposition where the power asymmetry between involuntary clients and the state becomes unmitigated. Standard AI governance does not register this erosion, because the conditions it assumes—effective human oversight, capacity to contest, organizational ability to monitor—are precisely the conditions that fail in social work. The problem is not AI as such, but AI integrated under governance whose operating assumptions do not hold in social work. Preserving the profession's mediating role is therefore not merely a professional interest but a governance imperative that must be addressed deliberately in the design of AI governance for

social work.

4.3. A triple mandate-informed guiding model

Our analysis established that the misalignment between AI governance operating conditions and social work practice represents fundamental accountability failures (Busuioc, 2021; Novelli et al., 2024), and that the triple mandate provides the structural foundation through which these failures can be addressed (Staub-Bernasconi, 2009). While the triple mandate originally describes the competing obligations navigated by individual practitioners, we adapt it here to the governance level, arguing that AI governance in social work must preserve the conditions for fulfilling all three mandates when algorithmic systems are integrated into practice (Garkisch & Goldkind, 2024).

Each mandate defines a core principle that AI governance must protect: client self-determination, dignity, and rights; institutional accountability for responsible AI deployment and its impacts; and professional ethical judgment, discretion, and knowledge-based practice (Staub-Bernasconi, 2009; Garkisch & Goldkind, 2024). Drawing on the misalignment identified above, we propose guiding questions organized at the intersection of the mandate domains and three accountability requirements: transparency, contestability, and redress (Table 1). Each guiding question asks whether the conditions necessary for meaningful accountability are present within a given mandate domain. These questions do not constitute a comprehensive governance framework but a minimum accountability layer that specifies the conditions for sustaining accountability across all three mandates, a floor that any AI governance approach applied in social work must meet regardless of technical or legal compliance.

The model also carries a broader structural argument, illustrated in Figure 2. When AI systems override professional expertise (Figure 2a), the professional mandate is weakened, tensions between mandates are distorted, and the triple mandate risks collapsing into a double mandate in which clients face organizational power without professional mediation. When professional expertise governs AI tools (Figure 2b), tensions remain but are manageable, and the triple mandate is preserved.

This structural argument rests on a fundamental premise that the professional mandate is enacted by people, not delegated to systems, and practitioners can fulfill their mediating role only when organizations and AI tools are designed to support their judgment rather than displace it. As AI becomes more influential in social work, the demands on professional judgment increase rather than diminish. Practitioners must now also discern when to trust, question, or reject algorithmic inputs, which requires sustained investment in training, supervision, and workforce capacity (Ahn et al., 2025a). Efficiency gains from AI must therefore be reinvested in professional capacity rather than used to justify workforce reduction or further automation (Ahn, 2026). AI governance in social work that does not invest in the people who enact the mandate will find it hollowed out, the accountability structure collapsed, and the people social work aims to serve will bear the consequences.

4.4. A virtual scenario: AI in homeless services

A homeless services agency adopts an AI tool to prioritize clients in higher need for scarce housing. Maria, a caseworker, meets a client whose circumstances she reads as acute, yet the tool returns a low score that places him far down the queue. The system is accurate, Maria reviews each score, and procurement met every legal requirement. By conventional AI governance standards, nothing has gone wrong, and the story could end here.

The triple mandate-informed questions guide us to see the other side of this story. Under the client mandate, the client nominally has the right to contest the decision, yet someone who has been homeless for years, exhausted and distrustful of systems that have failed him repeatedly, is in no position to file a formal challenge; the right is real on paper but inaccessible in practice. Under the organizational mandate, the agency has run the general bias evaluation that conventional governance requires and found nothing amiss, yet a generic audit does not ask whether the tool systematically disadvantages the hardest-to-house or quietly displaces caseworker judgment. And most importantly, under the professional mandate, Maria has nominal authority to override the score and mediate between the client and the organization, yet she cannot judge whether the tool is biased when so many decision steps are automated and the process remains opaque to her, and departing from its recommendation invites scrutiny and added documentation. The triple mandate-informed questions bring these failures into view, each a safeguard that conventional governance treats as satisfied but that fails under the conditions of social work.

5. Discussion and Conclusion

In this paper, we have examined how existing AI governance frameworks operate when applied to social work, a context with distinctive features that differ in important ways from the settings where these frameworks were originally developed. Our analysis suggests that the transfer is not straightforward: the populations social work serves are disproportionately involuntarily involved, structurally marginalized, and less able to exercise the kinds of agency that existing frameworks presume; decisions carry consequences that cannot be reversed by post-hoc correction; and the relational, discretionary nature of practice resists the standardization that algorithmic governance often requires.

Table 1. Guiding Model for AI Governance in Social Work: Accountability Requirements Across the Triple Mandate.

Triple Mandate Domain	Guiding Questions	Practical Considerations
Client Mandate *Protecting client self-determination, dignity, and rights in AI-mediated services*	*Transparency:* Can clients receive understandable explanations of how algorithmic outputs influenced decisions about their lives? *Contestability:* Is there an independent pathway for clients to question and contest AI-informed decisions? *Redress:* Are there meaningful mechanisms for redress, including human-led appeals processes, when algorithmic errors cause harm?	Transparency may be technically infeasible for complex algorithms; meaningful appeals slow efficiency and increase costs; involuntary clients face power asymmetries that limit their capacity to exercise these rights even when formally available. Clients should be treated as rights-holders first, with their dignity placed above administrative efficiency.
Organizational Mandate *Institutional accountability for responsible AI deployment and its impacts*	*Transparency:* Is the organization systematically monitoring AI system performance and its effects on the balance among the three mandates? *Contestability:* Can the organization justify its AI adoption and operation in terms of preserving the triple mandate, not only technical performance or legal compliance? *Redress:* Is the organization conducting regular audits for fairness and bias, and taking corrective action when structural inequalities or systematic harms are identified?	Monitoring and auditing require dedicated resources and technical expertise many social work agencies lack; procurement rarely includes social work-specific criteria; transparency about failures risks reputational damage.
Professional Mandate *Protecting ethical judgment, discretion, and knowledge-based practice in AI-mediated decision-making*	*Transparency:* Are the assumptions, limitations, and value trade-offs embedded in the algorithm transparent to practitioners? *Contestability:* Do practitioners have protected authority to override algorithmic recommendations, with AI positioned as a support tool rather than a replacement for professional judgment? *Redress:* Is required AI literacy supported through education, time, supervision, and organizational culture that values professional reasoning over algorithmic compliance?	Override authority is meaningless without supportive organizational culture; proprietary restrictions may limit algorithmic transparency; sustained investment in training is needed to prevent deskilling over time.

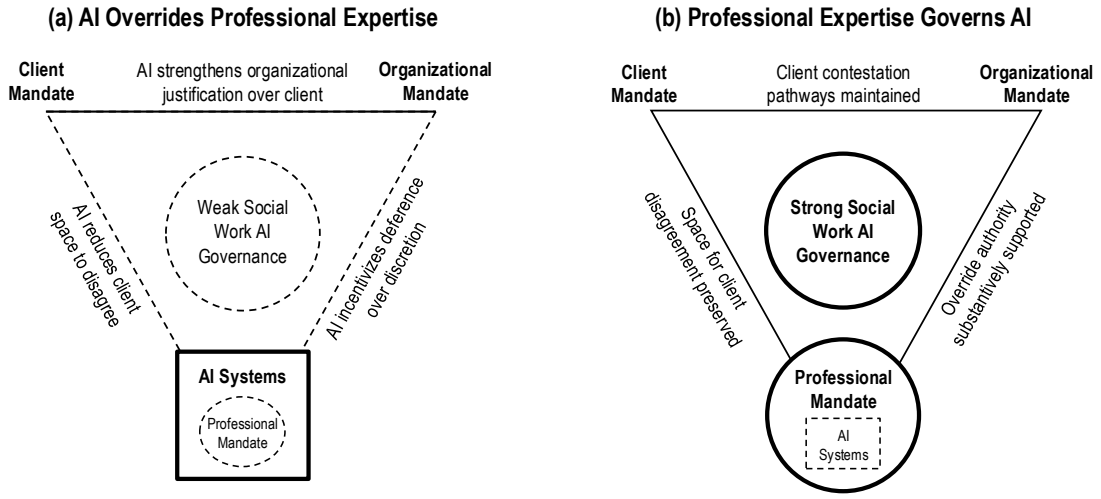

Figure 2. The triple mandate under two AI governance configurations. (a) When AI systems override professional expertise, the professional mandate is weakened (dashed circle), tensions between mandates become unmediated, and the triple mandate collapses into a double mandate in which clients face organizational power without professional mediation. (b) When professional expertise governs AI, the professional mandate remains strong (solid circle), AI tools operate within professional authority, and the triple mandate is preserved with client contestation pathways maintained.

These are not problems of implementation but questions about the conditions under which accountability itself can be sustained.

By adapting Staub-Bernasconi's (Staub-Bernasconi, 2007; 2009; 2016) triple mandate from an individual practice principle to a governance-level framework, we proposed a minimum accountability layer that organizes AI accountability along three mandate domains (client, organizational, and professional) intersecting with three accountability requirements (Bovens, 2007; Busuioc, 2021). The model's central argument is that the professional mandate functions as the independent ethical fulcrum holding the triple mandate together. When AI weakens this fulcrum, the triple mandate risks collapsing into a double mandate, leaving involuntary, marginalized clients and the state in direct opposition without professional mediation.

We offer this work as a contribution to an ongoing dialogue, with several implications worth noting. First, the diagnostic here may generalize beyond social work. Other high-stakes, relational public-sector domains where professionals mediate between vulnerable individuals and institutional power, such as healthcare, disability services, and community corrections, share the structural features that drive the misalignment we identify—involuntary or constrained engagement, irreversible consequences, and reliance on professional judgment that algorithmic systems can erode. Examining to what extent similar adaptations apply in those settings is an open question for future work.

Second, the analysis points to research and practical implications for the machine learning community. As public service agencies increasingly adopt AI, including those that share key features with social work such as involuntary clients and relationship-based services, ML researchers and engineers should recognize that conventional AI governance can carry them only so far. Meeting accountability in these settings calls for a further step informed by the triple mandate. Concretely, this means building transparency that non-expert and involuntary users can actually act on, contestation pathways that are not gated by institutional power, methods for detecting when a deployed model reshapes the outcomes it predicts, and evaluation that captures relational and procedural harms rather than accuracy alone. Most importantly, pursuing these well requires building them with domain practitioners and the people affected, not merely for them, since the conditions that determine whether a safeguard actually functions are visible primarily from within the context of use.

Lastly, and most broadly, we hope this work contributes to an emerging conversation about how AI governance knowledge, substantial and still developing, can be thoughtfully adapted to the diverse contexts in which AI systems are now being deployed. This adaptation is unlikely to succeed as a unilateral effort from any single community; it calls for sustained collaboration between AI researchers, domain practitioners, and those most affected by algorithmic systems.

References

Afrouz, R. and Lucas, J. A systematic review of technology-mediated social work practice: Benefits, uncertainties, and future directions. *Journal of Social Work*, 23(5): 953–974, 2023. doi: 10.1177/14680173231165926.

Ahn, E. Who is the human in human-centered AI? *AI & Society*, 2026. doi: 10.1007/s00146-025-02825-6.

Ahn, E., Choi, M., Fowler, P., and Song, I. H. Artificial intelligence (AI) literacy for social work: Implications for core competencies. *Journal of the Society for Social Work and Research*, 16(1):9–26, 2025a. doi: 10.1086/735187.

Ahn, E., Morstatter, F., Waters-Roman, D., Palmer, L., and McCroskey, J. Qualitative exploration of child welfare workers' decision-making experiences and perspectives on fairness. *Journal of Public Child Welfare*, 19(1):229–252, 2025b. doi: 10.1080/15548732.2024.2312846.

Amnesty International. Xenophobic machines: Discrimination through unregulated use of algorithms in the dutch childcare benefits scandal. Technical Report EUR 35/4686/2021, Amnesty International, 2021.

Attard-Frost, B. and Lyons, K. AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. *AI and Ethics*, 5(3):2557–2604, 2025. doi: 10.1007/s43681-024-00569-5.

Báez, J. C., Ahn, E., Tamietti, A., Victor, B. G., and Goldkind, L. Clinical social workers' perceptions of large language models in practice. *Journal of Evidence-Based Social Work*, 23(1):42–63, 2026.

Batool, A., Zowghi, D., and Bano, M. AI governance: A systematic literature review. *AI and Ethics*, 5(3):3265–3279, 2025.

Berrick, J. D. *The Impossible Imperative: Navigating the Competing Principles of Child Protection*. Oxford University Press, 2018.

Birkstedt, T., Minkkinen, M., Tandon, A., and Mäntymäki, M. AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7):133–167, 2023.

Bovens, M. Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4): 447–468, 2007.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., and Vaithianathan, R. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

Bureau of Labor Statistics. Social workers. U.S. Department of Labor, Occupational Outlook Handbook, 2025. https://www.bls.gov/ooh/community-and-social-service/social-workers.htm.

Busuioc, M. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81 (5):825–836, 2021.

Butcher, J. and Beridze, I. What is the state of artificial intelligence governance globally? *RUSI Journal*, 164 (5–6):88–96, 2019.

Camilleri, M. A. Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems*, 41(7):e13406, 2024.

Charette, R. N. Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold. IEEE Spectrum, January 2018.

Chenot, D. The vicious cycle: Recurrent interactions among the media, politicians, the public, and child welfare services organizations. *Journal of Public Child Welfare*, 5 (2–3):167–184, 2011.

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., and De Oliveira, N. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10):100857, 2023.

David, G. Artificial intelligence: Opportunities and challenges for public administration. *Canadian Public Administration*, 67(3):388–406, 2024.

Devlieghere, J., Gillingham, P., and Roose, R. Dataism versus relationshipism: A social work perspective. *Nordic Social Work Research*, 12(3):328–338, 2022.

Etehad, M. and Winton, R. Gabriel Fernandez case prosecution. Los Angeles Times, 2017.

Farrell, H. AI as governance. *Annual Review of Political Science*, 28(1):375–392, 2025.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *SSRN Electronic Journal*, 2020.

Fong, K. Child welfare involvement and contexts of poverty. *Children and Youth Services Review*, 72:5–13, 2017.

Forenza, B. and Eckert, C. Social worker identity: A profession in context. *Social Work*, 63(1):17–26, 2018.

Garkisch, M. and Goldkind, L. Considering a unified model of artificial intelligence enhanced social work: A systematic review. *Journal of Human Rights and Social Work*, 10(1):23–42, 2024. doi: 10.1007/s41134-024-00326-y.

Garrett, P. M. 'magic moments': AI and the 'disappearance' of social work ethics? *The British Journal of Social Work*, 2025. doi: 10.1093/bjsw/bcaf230.

Gasser, U. and Almeida, V. A. F. A layered model for AI governance. *IEEE Internet Computing*, 21(6):58–62, 2017.

Gillingham, P. Can predictive algorithms assist decision-making in social work with children and families? *Child Abuse Review*, 28(2):114–126, 2019.

Gillingham, P. and Graham, T. Big data in social welfare: The development of a critical perspective on social work's latest 'electronic turn'. *Australian Social Work*, 70(2): 135–147, 2017.

Hailu, E. M., Maddali, S. R., Snowden, J. M., Carmichael, S. L., and Mujahid, M. S. Structural racism and adverse maternal health outcomes: A systematic review. *Health & Place*, 78:102923, 2022.

Hu, L. What is new, and what is old, in fairness and machine learning. *ACM Journal on Responsible Computing*, 2025.

Jobin, A., Ienca, M., and Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1 (9):389–399, 2019.

Keddell, E. Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences*, 8(10):281, 2019.

Kolt, N., Shur-Ofry, M., and Cohen, R. Lessons from complex systems science for AI governance. *Patterns*, 6(8): 101341, 2025.

Kuziemski, M. and Misuraca, G. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6):101976, 2020.

Lee, J. Y., Ahn, E., Xu, A., Yang, Y., Chang, Y., Cha, H., and Ammari, T. Artificial intelligence in applied family research involving families with young children: A scoping review. *Family Relations*, 2024.

Lipsky, M. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. Russell Sage Foundation, 30th anniversary expanded edition, 2010.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., and Viljanen, M. Defining organizational AI governance. *AI and Ethics*, 2(4):603–609, 2022.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., and Viljanen, M. Putting AI ethics into practice: The hourglass model of organizational AI governance. *arXiv preprint arXiv:2206.00335*, 2023.

Mao, F. Robodebt: Illegal Australian welfare scheme investigation. BBC News, 2023.

McCradden, M. D., Mazwi, M. L., and Oakden-Rayner, L. Can an accurate model be bad? *Patterns*, 6(4):101205, 2025.

McNellan, C. R., Gibbs, D. J., Knobel, A. S., and Putnam-Hornstein, E. The evidence base for risk assessment tools used in U.S. child protection investigations: A systematic scoping review. *Child Abuse & Neglect*, 134:105887, 2022.

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., and Elish, M. C. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 735–746, 2021.

Mishna, F., Milne, E., Bogo, M., and Pereira, L. F. Responding to COVID-19: New trends in social workers' use of information and communication technology. *Clinical Social Work Journal*, 49(4):484–494, 2021.

Mišić, J., Van Est, R., and Kool, L. Good governance of public sector AI: A combined value framework for good order and a good society. *AI and Ethics*, 5(5):4875–4889, 2025.

Morozov, E. *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs, 2013.

National Association of Social Workers. Code of ethics of the National Association of Social Workers. NASW Press, 2021.

Neil, R. and Zanger-Tishler, M. Algorithmic bias in criminal risk assessment: The consequences of racial differences in arrest as a measure of crime. *Annual Review of Criminology*, 8(1):97–119, 2025.

Novelli, C., Taddeo, M., and Floridi, L. Accountability in artificial intelligence: What it is and how it works. *AI & Society*, 39(4):1871–1882, 2024.

Office of Management and Budget. Advancing governance, innovation, and risk management for agency use of artificial intelligence. Technical Report OMB Memorandum M-24-10, Executive Office of the President, 2024.

Park, S., Ahn, E., Ahn, T.-H., Ahn, S., Park, S., Kwon, E., Ahn, S., and Yang, Y. Artificial intelligence and aging in place: A scoping review of current applications and future directions. *The Gerontologist*, 65(6), 2025.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Technical report, arXiv, 2020.

Redden, J., Dencik, L., and Warne, H. Datafied child welfare services: Unpacking politics, economics and power. *Policy Studies*, 41(5):507–526, 2020.

Roberts, D. E. Child protection as surveillance of African American families. *Journal of Social Welfare and Family Law*, 36(4):426–437, 2014.

Robles, P. and Mallinson, D. J. Advancing AI governance with a unified theoretical framework: A systematic review. *Perspectives on Public Management and Governance*, 2025.

Rooney, R. H. and Mirick, R. (eds.). *Strategies for Work with Involuntary Clients*. Columbia University Press, third edition, 2018.

Staub-Bernasconi, S. Soziale Arbeit: Dienstleistung oder Menschenrechtsprofession? In Lob-Hüdepohl, A. and Lesch, W. (eds.), *Ethik Sozialer Arbeit – Ein Handbuch*, pp. 20–54. UTB/Schöningh, 2007.

Staub-Bernasconi, S. Social work as a discipline and profession. In Leskošek, V. (ed.), *Theories and Methods of Social Work: Exploring Different Perspectives*, pp. 9–30. Faculty of Social Work, University of Ljubljana, 2009.

Staub-Bernasconi, S. Social work and human rights—linking two traditions of human rights in social work. *Journal of Human Rights and Social Work*, 1(1):40–49, 2016.

Taylor, B. J. Risk-managing decision-making: A psychosocial rationality model. *The British Journal of Social Work*, 51(7):2819–2838, 2021.

Van Amsterdam, W. A. C., Van Geloven, N., Krijthe, J. H., Ranganath, R., and Cinà, G. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 6(4):101229, 2025.

Varon, J. and Peña, P. Artificial intelligence and consent: A feminist anti-colonial critique. *Internet Policy Review*, 10 (4), 2021.

Wang, X., Oussalah, M., Niemilä, M., Ristikari, T., and Virtanen, P. Towards AI-governance in psychosocial care: A systematic literature review analysis. *Journal of Open Innovation: Technology, Market, and Complexity*, 9(4): 100157, 2023.

Whittaker, A. and Havard, T. Defensive practice as 'fear-based' practice: Social work's open secret? *British Journal of Social Work*, 46(5):1158–1174, 2016.

Wirtz, B. W., Weyerer, J. C., and Sturm, B. J. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, 43(9):818–829, 2020.

Wirtz, B. W., Weyerer, J. C., and Kehl, I. Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, 39 (4):101685, 2022.