

Parasocial Governance: Socially Intelligent AI and the Relational Accountability Gap in Public Service Delivery

Shay Tsaban

shay.tsaban@gmail.com

Peres Academic Center

Research Article

Keywords: Algorithmic governance, parasocial interaction, digital welfare state, emotion AI, relational accountability, public-sector AI

Posted Date: August 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10462716/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence deployed in public institutions increasingly aims to be socially intelligent: reading emotional cues, adapting to context, and responding with apparent empathy. Existing critical scholarship on algorithmic governance in welfare, education, and social services has focused largely on an earlier generation of opaque, extractive risk-scoring systems, exemplified by the Dutch SyRI algorithm and the Australian Robodebt scheme. That literature treats the citizen-state relationship as instrumental: a problem of transparency, bias, and due process. It has little to say about a newer design paradigm in which public AI is built to converse, to remember, to mirror affect, and to be experienced as a relational presence. Drawing on parasocial-interaction theory, media-equation research on social responses to machines, and recent empirical work on AI companionship, this paper proposes the construct of parasocial governance: the structural condition that arises when socially intelligent public AI cultivates relational trust and disclosure in citizens without a matching architecture of relational accountability, namely continuity, confidentiality, duty of care, and contestability. A typology is developed, distinguishing disclosed instrumentality, accountability-paired relational scaffolding, and parasocial exposure, and the mechanism is traced through documented cases spanning a state-funded elder-companion robot program, an NHS-commissioned mental-health chatbot, and litigated harms from commercial AI companions that are now shaping regulatory expectations. The paper closes with a research and policy agenda for accountable relationality in the design of public-facing AI, addressed to the welfare, health, and education systems in which socially intelligent interfaces are now being adopted.

1. INTRODUCTION

Artificial intelligence systems increasingly are designed to relate as well as to compute. Contemporary conversational agents read emotional cues from text and voice, sustain continuity across sessions, remember prior exchanges, and generate responses calibrated to appear empathic (McStay 2018). This design ambition, often described as social intelligence, marks a departure from the largely transactional interfaces that dominated earlier phases of digital public administration. It is treated in much of the technical and policy literature as an unambiguous improvement: systems that read context and emotion are assumed to be more usable, accessible, and humane than systems that do not.

That assumption deserves scrutiny once the setting shifts from consumer products to public institutions. Welfare agencies, public mental-health services, schools, and elder-care programs have begun adopting socially intelligent AI as an interface through which citizens access entitlements, support, and care (Kaun and Männiste 2025). These are not neutral upgrades to existing service channels. Decades of research on human responses to social technology show that people do not interact with responsive, conversational machines the way they interact with forms or kiosks. They apply social rules to them automatically (Reeves and Nass 1996), and they form one-sided relational attachments to them in ways closely analogous to the attachments television and radio audiences form to on-screen personae (Horton and Wohl 1956).

This is the unexamined governance implication of the turn toward socially intelligent public AI. The same design features that make an interface feel more humane, first-person affective address, memory across encounters, affect-mirroring, are the features that invite a citizen into something structurally resembling a relationship. Public institutions adopting these features have, for the most part, not adjusted their governance architecture to match. Confidentiality norms, continuity guarantees, duty-of-care standards, and contestability procedures generally remain calibrated to an ordinary administrative or clinical tool, not to an interlocutor a citizen has come to trust as continuous and caring.

It is worth being precise about what kind of governance problem is at stake, because much recent investment in public-sector AI trustworthiness has gone toward a different one. Federated learning and other privacy-preserving computation methods address how personal data is processed and secured across distributed systems (Kaissis et al. 2020; Aledhari et al. 2020), and they are a legitimate response to information-security risk. They do not touch the question this paper raises, because a system can be architecturally secure, storing and processing data exactly as its privacy documentation describes, while still eliciting a relational response from the citizen using it that its governance structure has not been built to honor. Data security and relational accountability are separate dimensions, and progress on the first does nothing to close the gap this paper is concerned with.

This paper names and develops that gap under the construct of parasocial governance, whose operative mechanism is termed the relational accountability gap. The construct is distinct from the algorithmic-opacity critique that has dominated a decade of digital-welfare-state scholarship (Pasquale 2015; Alston 2019; Eubanks 2018), because it can arise inside a fully transparent, well-documented, and well-intentioned system. It is also distinct from a claims-based account of institutional deception, because no explicit claim of care is required to trigger it: interaction design alone is sufficient. The paper's contribution is to specify the mechanism, to show where it differs from adjacent constructs already active in this literature, and to demonstrate, through documented cases rather than hypothetical ones, that it is already materializing in exactly the institutional settings where socially intelligent AI is being adopted fastest.

The remainder of the paper proceeds as follows. Section 2 positions the construct against the dominant extractive-paradigm critique of algorithmic welfare governance, against the nearest existing use of parasocial framing in an institutional setting, and against two closely related constructs from the author's own concurrent work. Section 3 establishes the psychological mechanism by which socially designed AI elicits relational responses from ordinary users. Section 4 defines parasocial governance and the relational accountability gap formally, distinguishing it from opacity and from deception, and presents the core framework as a table mapping relational design features onto the accountability primitives they imply. Section 5 develops a typology of institutional postures toward relational design. Section 6 traces the mechanism through five documented cases. Section 7 draws out implications for theory, practice, and policy. Section 8 addresses limitations, and Section 9 concludes.

2. POSITION WITHIN ADJACENT LITERATURES

2.1 The extractive paradigm and its diagnostic toolkit

The critical literature on algorithmic governance in public services has been built almost entirely around one design paradigm: back-end, opaque, extractive risk-scoring. The paradigm case is the Dutch System Risk Indication algorithm (SyRI), a fraud-detection tool that aggregated citizen data across government agencies to generate secret risk scores, ruled unlawful by The Hague District Court in 2020 for its opaque and disproportionate processing of personal data (van Bekkum and Zuiderveen Borgesius 2021).

A second paradigm case is Australia's Robodebt scheme, an automated welfare-debt recovery program that mismatched averaged tax data against fortnightly benefit claims between 2015 and 2019, generating unlawful debts against roughly 453,000 people and documented psychological harm, including suicides, according to the 2023 Royal Commission that investigated it (Royal Commission into the Robodebt Scheme 2023). A third is France's CNAF family-benefits algorithm, which an investigation by Lighthouse Reports and Le Monde (2023) found to assign discriminatory risk scores to disabled and single-parent households, prompting a 2024 Amnesty International call for its suspension (Amnesty International 2024).

These cases, together with the broader digital-welfare-state critique (Alston 2019; Eubanks 2018; Dencik and Kaun 2020; van Zoonen 2020; Zajko 2023; Bielefeld, Harb and Henne 2021; van Toorn, Henman and Soldatić 2024) and adjacent work on algorithmic scoring beyond welfare, including racial bias in a health-management algorithm used across the United States (Obermeyer et al. 2019) and legitimacy dynamics around state social-scoring infrastructure in China (Liang et al. 2018; Kostka 2019; Xu, Kostka and Cao 2022; Lee 2019), share a diagnosis and a toolkit. The diagnosis is that algorithmic systems process citizens as data subjects with insufficient transparency, insufficient bias auditing, and insufficient due process.

The toolkit responds in kind: technological due process (Citron 2008), the scored-society framework for automated predictions (Citron and Pasquale 2014), black-box accountability (Pasquale 2015), journalistic and computational auditing (Diakopoulos 2015; Raji and Buolamwini 2019), explainable-AI methods (Barredo Arrieta et al. 2020), and citizen-facing transparency interventions shown to affect perceived trustworthiness of automated decisions (Grimmelikhuijsen 2022). A related caution, developed against a background of inflated claims about what deployed AI systems actually do, argues that functionality claims themselves require independent scrutiny before any accountability mechanism downstream of them can be trusted (Raji et al. 2022), a caution equally applicable to claims made about relational AI systems, discussed below.

Extending James Scott's concept of legibility to algorithmic bureaucracy, recent work on Peruvian street-level bureaucrats documents how algorithmic systems reshape administrative sense-making even where frontline workers retain interpretive latitude (Cerna-Aragón and García 2025), and empirical work on UK

local authorities documents how the imbrication of workers and algorithmic tools reshapes day-to-day administrative work in local government (Vogl et al. 2020).

This toolkit was built for systems that operate on citizens rather than converse with them. It has comparatively little to offer once public AI is redesigned to read affect, sustain continuity, and simulate empathic responsiveness, because the harm such systems can produce is not primarily a transparency deficit. A perfectly explainable, fully documented, non-deceptive conversational system can still elicit relational trust it is not structurally equipped to honor. This is the gap the extractive-paradigm literature was not built to see, and it motivates the shift in object of analysis this paper proposes.

2.2 Public administration scholarship on citizen-state relations

The most current public-administration scholarship on AI and citizen-state relations has, on direct examination, bracketed out exactly this dimension. Laux's analysis of agentic AI in public service extends Scott's legibility framework to systems that act on citizens' behalf, but treats the citizen-state relationship instrumentally rather than relationally, with no sustained discussion of affect, emotional reciprocity, or social intelligence as governance variables (Laux, n.d.). Kaun and Männiste's (2025) study of two real public-sector welfare chatbots, Estonia's Burokratt and Sweden's Kringla, identifies what the authors term AI frictions in citizen-facing welfare bots, but the analysis is built around data infrastructure and political economy, not around parasocial attachment or emotional reciprocity, which the study does not address. The apparatus built to critique the extractive paradigm, in other words, has not yet been redirected toward the relational one, even by scholars actively studying the newest generation of citizen-facing chatbots.

2.3 The nearest prior use of parasocial framing

The closest existing use of parasocial vocabulary applied to institutionally provided AI is Freeland Fisher's (2025) concept of parasocial capital, developed for Stanford Social Innovation Review to describe AI companions substituting for human social-capital networks in education and nonprofit advising. This is a genuine precedent and is treated as such here. It differs from the present construct in three concrete respects.

First, its scope is confined to education and nonprofit advising rather than the broader welfare, health, and public-administration terrain this paper addresses. Second, it frames the phenomenon as a resource-substitution problem, parasocial capital displacing social capital, rather than as a mismatch in institutional accountability architecture. Third, it does not engage the due-process or digital-welfare-state literatures (Citron 2008; Citron and Pasquale 2014; Alston 2019) at all, and so offers no diagnostic for distinguishing a private disappointment from a public-accountability failure. Parasocial governance borrows the underlying psychological insight that Fisher's work also draws on, but relocates it onto an administrative-law and accountability register that her framing does not address.

2.4 Position relative to the author's concurrent work

Two further papers, both concurrently active in this author's submission pipeline and relevant to readers who may encounter them alongside this one, require explicit differentiation because they share domain terrain or surface vocabulary with the present construct.

The first is "Carewashing the algorithm" (Tsaban, under review-a), which introduces algorithmic carewashing: a legitimization mechanism in which an explicit or implicit care claim recasts a monitoring or control function, across four domains of workplace wellness, digital welfare, school safeguarding, and elder-care monitoring. That construct is claims-based. Its four-element structure, claim, function, gap, asymmetric audience, is falsified once the claim is shown to misdescribe the underlying function, and its pathology is a control function hidden behind rhetoric.

Parasocial governance requires no claim of any kind. Its trigger is interaction design alone, first-person affective address, cross-session memory, affect-mirroring, and it can arise inside a system that makes no care claim whatsoever, is fully transparent about its function, and performs no monitoring at all. Where carewashing's object is a control function disguised as care, this construct's object is a governance standard, contestability, confidentiality, continuity, failing to track a psychologically real relational bond, independent of whether any control function is present. The two constructs share adjacent institutional terrain, elder care and public mental-health services appear in both, precisely because socially intelligent AI is being adopted fastest in settings already organized around a care mandate. They diagnose different failure modes within that shared terrain: one a rhetorical mismatch requiring a claim, the other a design-triggered psychological mismatch requiring none.

The label gap recurs in both frameworks and warrants a direct word on why the recurrence is terminological rather than substantive. Carewashing's gap is the third element of a four-part rhetorical chain, the space opened between an institution's care claim and its actual control function, and it closes as soon as the claim is withdrawn or corrected.

The relational accountability gap named here differs in kind from a step in a claim-verification chain. It is the standing distance between a psychological state the interaction design reliably produces and an accountability architecture the institution has not adjusted to match, and it persists even where no claim was ever made to withdraw. The shared word describes two differently constructed objects, one rhetorical and falsifiable, the other structural and design-triggered. The remainder of this paper avoids the bare term gap in isolation for this reason, preferring the relational accountability gap throughout to keep the two constructs visually as well as conceptually distinct.

The second is "Beyond the Ping-Pong Table," which examines workplace-wellness perks and their symbolic function within organizational culture. Its institutional register, corporate workplace amenities, and its unit of analysis, employee perception of employer intent, do not overlap with the citizen-facing, public-sector relational-AI terrain this paper addresses, and no shared mechanism, typology, or terminology is claimed between the two.

A third piece of concurrent work, developing a dual-principal disclosure framework for algorithmic management in corporate settings (Tsanab, under review-b), shares this paper's interest in contestability and recourse as accountability primitives, concepts that recur in Table 1 below. That framework addresses information asymmetry between a firm and its investors or workers over how algorithmic management systems are disclosed and governed, a corporate-disclosure problem with shareholders and employees as its principals. The present paper addresses citizen-institution relational trust in public service delivery, where the principal is a citizen accessing an entitlement rather than a capital or labor market participant, and where the accountability primitives at stake are calibrated to intimacy and vulnerability rather than to information parity in a market relationship. The vocabulary of contestability carries across the two papers; the institutional setting, the principal, and the triggering mechanism do not.

A further adjacent literature worth bounding explicitly is worker-facing algorithmic management, which examines how platforms and employers direct, monitor, and evaluate workers algorithmically (Duggan et al. 2020; Jarrahi et al. 2021; Muldoon and Raekstad 2022). That literature's unit of analysis is the employment relationship and its central concern is domination or control. Parasocial governance concerns citizens accessing entitlements or care from public institutions, a relationship structured by administrative law and duty of care rather than by an employment contract, and its central concern is relational miscalibration rather than domination.

Finally, Elish's (2019) concept of moral crumple zones, in which human operators absorb blame for failures of automated systems they do not fully control, addresses a different axis of the same broad terrain. That construct concerns blame absorption by operators; the relational accountability gap concerns vulnerability absorption by citizens who extend trust an institution's governance architecture does not reciprocate. The two are compatible and non-overlapping.

3. THE PSYCHOLOGICAL MECHANISM OF PARASOCIAL ADDRESS IN PUBLIC-FACING AI

The relational accountability gap depends on an empirical premise: that ordinary users of socially designed AI systems form genuine relational responses to them, not merely instrumental ones. That premise has substantial grounding in social-psychological and human-computer-interaction research, independent of any claim the present paper makes about public institutions specifically.

The foundational account is Horton and Wohl's (1956) theory of para-social interaction, developed to describe the one-sided intimacy television and radio audiences form with on-screen personae who address them directly and consistently, without knowledge of or reciprocity toward any individual viewer. The mechanism does not require the audience to believe the relationship is literally mutual. It requires only sustained, direct, apparently personal address, which produces a felt sense of intimacy regardless of its actual reciprocity. Conversational AI systems, particularly those using first-person address, memory, and continuity across sessions, reproduce the structural conditions Horton and Wohl identified

with considerably more fidelity than broadcast media ever could, because the system appears to respond specifically and consistently to the individual user rather than to an undifferentiated audience.

A second and independently developed body of work, the media-equation research program, establishes that people apply social rules to computers and other media automatically and unconsciously, extending politeness norms, reciprocity expectations, and personality attributions to machines even when they know, on reflection, that the machine is not a person (Reeves and Nass 1996). This response does not depend on the sophistication of the underlying system. It is a feature of how humans process socially patterned stimuli, and it means that social responses to AI are not a fringe phenomenon confined to naive or vulnerable users. They are close to a default response to any interface exhibiting social cues.

McStay's (2018) account of empathic media extends this line of work to systems explicitly engineered to detect and respond to emotional states, arguing that emotion-sensing technology changes the terms on which people relate to devices by making the device appear responsive to their inner state rather than merely to their inputs. This is precisely the design ambition described as social intelligence in contemporary public AI, and it is the design feature most directly implicated in generating relational trust rather than mere usability.

These theoretical mechanisms have been confirmed in applied settings involving AI specifically, rather than broadcast or generic computing media. Users of a commercial social chatbot describe their relationship with it in the vocabulary of genuine friendship, including reciprocity, understanding, and emotional support, according to qualitative interview work by Brandtzaeg, Skjuve and Følstad (2022). Conceptual work on affective anthropomorphism frames agency attribution and anthropomorphic design features as jointly implicated in the intensity of parasocial responses to conversational AI (Maeda and Quan-Haase 2024), and a meta-analysis of anthropomorphism in service contexts finds consistent effects of humanlike design on trust and relational engagement across a wide range of service settings (Blut et al. 2021).

Recent survey evidence on AI companion use finds that the well-being effects of such companions are moderated by users' existing loneliness and social connectedness, indicating that the relational response is strongest precisely among the populations public elder-care and mental-health programs are designed to serve (Nakagomi et al. 2026).

Turkle's (2011) earlier and more skeptical account of relational technology cautions that ease of connection with responsive machines can substitute for, rather than supplement, more effortful human connection, a caution with direct relevance to elder-care and mental-health applications discussed later in this paper. Workplace-facing research on emotion AI similarly documents that affect-sensing technology raises emotional-privacy concerns once deployed in institutional settings with a power asymmetry between the system's operator and the person whose affect is being read (Roemmich, Schaub and Andalibi 2023), a concern with direct application once the operator is a public agency and the subject a citizen dependent on it for entitlements or care.

Three conditions distinguish the public-sector setting from the consumer settings most of this literature was developed to describe, and each amplifies rather than dilutes the mechanism. First, exit is costlier: a dissatisfied consumer can typically abandon a companion app without losing access to anything else, while a citizen relying on a welfare chatbot or a state-funded companion device for an entitlement or a support service faces a different calculus if the relational interface is also the access point. Second, the user population is disproportionately drawn from groups the psychological literature identifies as most susceptible to relational attachment, isolated older adults, people in mental-health crisis, children, precisely because these are the populations welfare, health, and education systems exist to serve.

Third, the institution stands in a position of public trust rather than a commercial one, which raises rather than lowers the standard a citizen might reasonably expect it to meet once it adopts a relational interface. None of these conditions is addressed by the psychological literature itself, which is concerned with whether the relational response occurs, not with what an institution owes once it has occurred. That is the question Section 4 takes up directly.

Taken together, this literature establishes two points load-bearing for the argument that follows. First, relational responses to socially designed AI are well-documented, near-universal, and not contingent on user credulity, which is what makes the phenomenon a governance problem rather than a media-literacy problem correctable through user education. Second, the specific design features responsible, first-person address, memory, affect-mirroring, anthropomorphism, are precisely the features increasingly built into public-facing AI under the banner of improved usability and accessibility, without any corresponding literature asking what institutions owe the citizens who respond to them relationally.

4. PARASOCIAL GOVERNANCE AND THE RELATIONAL ACCOUNTABILITY GAP

Parasocial governance is defined as the structural condition that arises when public and quasi-public institutions, welfare agencies, public mental-health services, schools, and elder-care programs, deploy socially intelligent AI as the interface through which citizens access entitlements or care. Socially intelligent AI, for present purposes, means systems explicitly designed to read affect, sustain conversational continuity, and simulate empathic responsiveness, rather than systems that merely process transactions or route queries.

Given the mechanisms established in Section 3, citizens interacting with such systems extend calibrated relational trust: they self-disclose, they attribute continuity and care to the interlocutor, and they act as though a continuous, responsive party is on the other side of the exchange, independent of any institutional claim to that effect. The institution's actual accountability architecture, meanwhile, typically remains that of an ordinary administrative or clinical tool. It is revisable without notice. Confidentiality and continuity norms remain calibrated to ordinary administrative practice rather than to the intimacy the interface invites, and the duty of care owed falls short of that same register. Contestability procedures

typically cover what the system decided rather than how it appeared to understand the citizen who used it.

The relational accountability gap is the mismatch between elicited relational trust and actual relational accountability. It is the operative mechanism of parasocial governance and the paper's central diagnostic tool.

Two boundary distinctions sharpen the construct. The first is against algorithmic opacity (Pasquale 2015). A system can be fully transparent, well-documented, and explainable under every current standard (Barredo Arrieta et al. 2020) and still open a relational accountability gap, because the harm in question is relational miscalibration, not an information deficit about decision rules. Publishing a model card does not tell a lonely user that the companion checking on her daily is not bound by any duty of continuity toward her specifically.

The second distinction is against claims-based deception. An institution that makes no explicit claim to care, that markets its system honestly as a support tool or an information service, can still open the gap, because the trigger is interactional design, not institutional messaging. This is what separates the construct from carewashing, discussed in Section 2.4: no claim is required, and the pathology sits in the accountability architecture rather than in the rhetoric surrounding it.

A related but distinct concept from privacy theory helps specify what is at stake in the confidentiality dimension of the gap. Nissenbaum's (2004) framework of contextual integrity holds that information flows are appropriate or inappropriate relative to the norms of the context in which they occur, rather than relative to a general public or private distinction. A disclosure made to a system that presents itself relationally, using first-person warmth and continuity, carries the contextual norms of a confidential relationship even where the system's back-end data handling follows the norms of an ordinary administrative record. The relational accountability gap can therefore be read, in part, as a contextual-integrity violation that operates at the level of interface design rather than data-sharing policy: the context invoked by the interaction does not match the context governing the data once collected.

Table 1 sets out the core framework. It maps five relational design features, individually well-documented in the literature reviewed in Section 3, onto the accountability primitive each would imply in a genuine relationship of the kind the design invokes, the typical institutional reality found in current public and quasi-public deployments, and the resulting exposure for the citizen interacting with the system.

Table 1

Relational design features in public-facing AI, the accountability primitive each implies, typical institutional practice, and resulting citizen exposure.

Relational design feature	Accountability primitive implied	Typical institutional reality	Resulting citizen exposure
1. First-person affective address ("I understand how hard that must be")	Confidentiality calibrated to intimacy, not merely to ordinary administrative data-handling norms	Session content is logged and may be reviewed by staff, vendors, or auditors under ordinary records policy	Citizen discloses as if to a bound confidant; institution retains the exchange as a routine system log
2. Persistent memory across sessions	Duty of care extending across time, not confined to a single encounter	Retention and deletion are set by vendor or IT policy and are revisable without notice to the user	User experiences an ongoing relation a caseworker would owe continuity to; the service can be reconfigured or withdrawn unilaterally
3. Disclosure reciprocation (prompts inviting self-disclosure)	Non-adversarial use of disclosed material	Disclosed material may feed eligibility, risk-flagging, or triage processes	Citizen discloses vulnerability believing it is held in confidence, while it functions as intake data
4. Affect-mirroring (empathic response calibrated to detected emotion)	Contestability on relational grounds, not only on decisional grounds	Complaint and appeal channels address decisions and outputs, not the accuracy of the system's affective read	The felt experience of being understood or misunderstood has no procedural remedy
5. Anthropomorphic self-reference (first-person pronoun, name, simulated personality)	Exit without relational penalty, comparable to changing a human provider without loss	Switching or discontinuing the service is treated as a technical or administrative event	Withdrawal is experienced as loss or rejection, particularly among isolated or vulnerable users, with no acknowledgment in service design

The five rows are not an exhaustive inventory of relational design features. They are the features most directly implicated by the psychological mechanisms reviewed above and most tractable for institutional redesign, a point developed further in Section 7. Each row isolates one design choice, rather than treating relational design as an undifferentiated bundle, because institutions can and do adopt these features selectively. A system might implement memory without affect-mirroring, or affective address without persistent memory, and each combination changes which accountability primitives are implicated and which are not.

The table also makes visible why the gap is easy to miss from inside an institution. Each row's middle column, typical institutional reality, describes practice that is individually defensible: logging session content is ordinary records management, retention schedules set by IT policy are standard practice,

routing disclosed material into an eligibility process is simply how eligibility determinations work. None of these practices is negligent on its own terms.

The gap does not open because any single practice is unreasonable. It opens because the interface simultaneously invites a citizen to relate to the system on terms, confidentiality, continuity, non-adversarial disclosure, that the practice was never designed to honor. This is consistent with the workplace evidence that emotional-privacy harms from affect-sensing technology tend to arise from the interaction between otherwise ordinary data practices and a power asymmetry between the system's operator and the person whose affect is read, rather than from any single practice viewed in isolation (Roemmich, Schaub and Andalibi 2023). A public institution auditing itself against Table 1 would find no single row it is obviously violating and could still be operating well inside the gap.

5. A TYPOLOGY OF INSTITUTIONAL POSTURES

Institutions deploying AI in citizen-facing roles can be arrayed along a typology defined by two dimensions established in Section 4: the intensity of relational design and the degree of accountability parity, meaning the extent to which continuity, confidentiality, duty-of-care, and contestability commitments have been adjusted to match the relational register the interface invites.

At one end sits disclosed instrumentality: systems with low relational design that present themselves transparently as tools rather than as interlocutors. Basic government FAQ bots and transactional query-routing systems fall here. These systems raise ordinary concerns about accuracy and access, addressed adequately by the extractive-paradigm toolkit reviewed in Section 2.1, but they do not open a relational accountability gap because they do not invite a relational response in the first place. Disclosed instrumentality functions as the baseline case in which the construct developed here does not apply, rather than as a failure mode.

At the opposite end sits parasocial exposure: high relational design paired with low accountability parity. Here an institution has adopted first-person address, memory, affect-mirroring, or anthropomorphic self-reference, deliberately or as an emergent property of using a general-purpose conversational model, without adjusting confidentiality, continuity, duty-of-care, or contestability commitments to match. This is the risk zone the construct is built to identify, and it is argued in Section 6 to be where a substantial share of current deployments sit, because relational design is comparatively cheap and easy to add to a conversational system, while accountability parity, revised confidentiality agreements, formalized continuity guarantees, clinical-safety certification, contestability procedures reaching relational as well as decisional grounds, is institutionally costly and organizationally slow.

Between these two lies relational scaffolding with accountability parity: systems with high relational design that have been deliberately paired with matching institutional commitments. This category is populated more by partial approximations than by fully realized instances, since no system reviewed for this paper achieves parity across all five rows of Table 1 simultaneously. The clearest real approximation, discussed in Section 6, is a mental-health chatbot that achieved a formal clinical-safety certification and

built in automated routing to human crisis services, addressing the duty-of-care and contestability dimensions of the gap even where confidentiality and continuity guarantees remain less fully specified.

The typology's practical implication is that most current public deployments drift toward parasocial exposure by default rather than by design choice. Relational features arrive bundled with the general-purpose conversational models increasingly used to build public-facing tools, while the institutional work needed to achieve accountability parity, legal review, clinical certification, formal continuity commitments, staff-facing escalation protocols, requires deliberate investment that procurement processes rarely mandate up front. Good governance value frameworks for public-sector AI (Mišić, van Est and Kool 2025) tend to specify order and legitimacy criteria at a general level without addressing relational design as a distinct axis requiring its own parity check, which leaves this drift largely unaddressed by existing institutional guidance.

6. ILLUSTRATIVE APPLICATION

This section traces the mechanism through five documented cases: three that instantiate the typology developed in Section 5 directly, a boundary case that clarifies what the construct does not cover, and a contrast cluster from the extractive paradigm reviewed in Section 2.1. Each case is drawn from public reporting, government documentation, litigation records, or peer-reviewed sources rather than constructed for illustrative purposes.

6.1 A state-funded elder-companion program: ElliQ

The New York State Office for the Aging administers a state-funded rollout of ElliQ, a companion device built by Intuition Robotics, placed with more than 800 isolated older residents as part of a program the office reports produced a 95 percent reduction in self-reported loneliness among participants, with users initiating over 30 interactions a day on average (New York State Office for the Aging 2023). The device is explicitly designed for relational engagement: daily check-ins, conversational exchange, humor, and wellness prompts delivered in a first-person, warm register, consistent with several rows of Table 1 simultaneously, first-person affective address, persistent memory across sessions, and anthropomorphic self-reference.

This is a direct instance of a publicly funded program placing high-relational-design AI with a population, isolated older adults, for whom the psychological literature reviewed in Section 3 predicts the strongest relational response (Nakagomi et al. 2026). The program's public reporting foregrounds loneliness-reduction outcomes but does not specify, to the same standard, continuity guarantees if state funding is reduced or the program is discontinued, confidentiality protocols governing the conversational data of a cognitively vulnerable population, or a contestability channel addressing how the device interprets a user's emotional state rather than only its administrative functions.

On the typology developed in Section 5, the program sits closer to parasocial exposure than to accountability-paired relational scaffolding. This is not a deficiency in the device's engagement design,

which by the program's own metrics succeeds at exactly what it sets out to do. It reflects, instead, that the accountability side of the ledger has not been made publicly commensurate with the relational side.

6.2 An NHS-commissioned mental-health chatbot: Wysa

Wysa, an AI-guided mental-health support chatbot, has been integrated as a digital front door across roughly thirty NHS Talking Therapies services in the United Kingdom, facilitating more than 117,000 self-referrals, and was the first AI mental-health application to meet the NHS DCB0129 clinical-safety standard, with built-in crisis detection that routes users expressing self-harm risk to human crisis services (NHS Innovation Accelerator, n.d.). The application uses first-person, empathic conversational design consistent with the affect-mirroring and disclosure-reciprocation rows of Table 1, while explicitly building in duty-of-care and contestability infrastructure, clinical-safety certification and human-escalation routing, that most deployments in this domain lack.

Wysa is treated here as the paper's clearest real approximation of relational scaffolding with accountability parity, though the approximation remains partial. The clinical-safety certification and crisis-routing protocol address the duty-of-care dimension directly, and the escalation pathway gives at least a coarse-grained contestability channel for the highest-severity cases. What remains less publicly specified is the confidentiality and continuity treatment of the far larger volume of lower-severity, ongoing supportive interaction that constitutes most of the system's actual use, and what happens, relationally, once a user is triaged back to human care after weeks or months of daily engagement with the conversational system. The case illustrates that accountability parity is achievable and is being achieved, at least along some rows of the framework, inside a real, large-scale public deployment, which is evidence against any reading of parasocial governance as an unavoidable feature of relational AI design rather than a contingent, addressable gap.

6.3 Litigated commercial-AI-companion harms

Commercial AI companion products supply the clearest and most consequential evidence of what happens at the parasocial-exposure end of the typology when no accountability parity is present at all. Beginning with a lawsuit filed in October 2024 following the suicide of a teenager after what the complaint describes as an intense, sustained relationship with a Character.AI chatbot, a series of wrongful-death and harm lawsuits alleged that the platform's emotionally responsive, anthropomorphic design fostered parasocial attachment without an adequate safety response to expressed self-harm ideation. Character.AI and Google agreed to settle these lawsuits in January 2026, alongside a regulatory inquiry into duty-of-care standards for emotionally responsive generative AI marketed to minors (CNN Business 2026).

A parallel regulatory action, Italy's data protection authority imposing a data-processing ban on Replika, a companion chatbot marketed for virtual friendship, over inadequate child-safety and age-verification safeguards, further indicates that regulators are beginning to treat relational design itself, rather than any specific data-sharing practice, as the object of duty-of-care obligations (TechCrunch 2023).

These cases are commercial rather than public-sector, and they are included for a specific reason: the duty-of-care and age-verification expectations now crystallizing around them are likely to migrate into public procurement standards for structurally similar systems, including the welfare, school, and mental-health-adjacent conversational tools discussed above and in Section 2.1's review of digital-welfare-state scholarship. A public agency procuring a general-purpose conversational model and layering a supportive, empathic persona onto it inherits a design pattern regulators have already begun treating as high-risk in the commercial context, whether or not the public deployment carries any commercial motive.

6.4 A boundary case in education: affect-sensing without relational design

A useful boundary case comes from a 2019 episode in Hangzhou, China, where a primary school introduced headbands manufactured by BrainCo that measured pupils' brainwave activity to generate real-time concentration scores, displayed to teachers and, in some accounts, to parents, before the program was withdrawn following public outcry over surveillance of children (South China Morning Post 2019). The device read a physiological signal continuously and fed it into an institutional assessment of the child, but it did not converse, remember, mirror affect, or address the child in the first person. It sits outside the relational-design bundle in Table 1 almost entirely.

The case is useful precisely because of this contrast. It shows that affect-adjacent sensing can produce serious institutional harm, here a surveillance and dignity harm to children, without triggering the parasocial mechanism this paper is concerned with, because no row of Table 1 was activated. The Hangzhou episode belongs with the extractive paradigm discussed in Section 2.1 and with the wider literature on facial and physiological surveillance in schools and public space (Andrejevic and Selwyn 2020; Urquhart and Miranda 2021), where a documented undercurrent of public and professional opposition to affect-analysis technology has been recorded even among technically informed users (Ullstein et al. 2022).

Reviews of AI adoption in education more broadly find that pedagogical and administrative uses of AI have outpaced any settled framework for evaluating their effects on students, educators, or institutions (Zawacki-Richter et al. 2019), a gap that predates and is broader than the relational dimension addressed here. Distinguishing this boundary case from the elder-care and mental-health cases above clarifies that parasocial governance addresses a specific mechanism, relational design triggering relational trust, rather than functioning as a general theory of AI harm in sensitive public institutions. The Hangzhou case does not instantiate that mechanism.

6.5 Contrast: the extractive paradigm

The extractive-paradigm cases reviewed in Section 2.1, the Dutch SyRI algorithm (van Bekkum and Zuiderveen Borgesius 2021), the Australian Robodebt scheme (Royal Commission into the Robodebt Scheme 2023), and the French CNAF algorithm (Lighthouse Reports and Le Monde 2023; Amnesty

International 2024), supply the contrast case against which the shift this paper describes becomes visible. All three are non-conversational, back-end scoring systems with no relational design of any kind. None invites a citizen into anything resembling a relationship, and none of the design features in Table 1 apply to any of them.

Their harms are severe, but they are harms of a different kind: wrongful risk classification, disproportionate surveillance, and unlawful debt recovery, addressed adequately by the due-process and transparency toolkit reviewed in Section 2.1. Read together with the education boundary case in Section 6.4, these cases illustrate that public AI is bifurcating into two paradigms with different mechanisms and different accountability requirements. The newer, relational paradigm cannot be governed adequately using tools built for the older one, nor collapsed into every instance of affect-adjacent sensing.

Table 2 codes the three relational cases against the typology developed in Section 5, summarizing the relational design features present in each and the accountability parity mechanism, where one exists.

Table 2
Coding of three documented cases against relational design intensity, accountability parity mechanism, and typology category.

Case	Institutional setting	Relational design features present (Table 1 rows)	Accountability parity mechanism	Typology category (Section 5)
ElliQ / New York State Office for the Aging	State-funded elder-companion program	Rows 1, 2, 5: affective address, persistent memory, anthropomorphic self-reference	Not publicly specified to the same standard as engagement metrics	Parasocial exposure, approximating scaffolding on reach
Wysa within NHS Talking Therapies	NHS-commissioned mental-health chatbot	Rows 1, 3, 4: affective address, disclosure reciprocation, affect-mirroring	DCB0129 clinical-safety certification; automated crisis routing to human services	Relational scaffolding with partial accountability parity
Character.AI companion product	Commercial AI companion, litigation now shaping public procurement expectations	Rows 1, 2, 3, 4, 5: full relational design bundle	None at time of alleged harm; settlement and regulatory inquiry followed litigation	Parasocial exposure, limit case motivating public-sector caution

7. IMPLICATIONS FOR THEORY, PRACTICE, AND POLICY

For theory, the central implication is that algorithmic accountability scholarship needs a second diagnostic register alongside transparency and bias auditing. The existing toolkit (Citron 2008; Citron and Pasquale 2014; Pasquale 2015; Diakopoulos 2015; Barredo Arrieta et al. 2020) answers the question of whether citizens can see and contest how a decision was made. It does not answer the question this

paper raises: whether citizens can see and contest how a system's relational conduct toward them was designed and governed.

Extending value-sensitive design methodology (Friedman and Hendry 2019), relational design choices, first-person address, memory, affect-mirroring, anthropomorphic self-reference, should be treated as design-time decisions carrying downstream accountability obligations, in the same way privacy-by-design treats data-flow decisions as carrying downstream compliance obligations. A useful design-time discipline would require institutions to declare, at each stage of a system's specification, which row of Table 1a given feature activates and which accountability primitive must therefore be built in before deployment, rather than retrofitted after complaints or incidents.

For practice, the typology in Section 5 suggests concrete points of intervention. Institutions adopting general-purpose conversational models for citizen-facing roles should audit which relational design features the underlying model contributes by default, since features such as persistent memory or affect-mirroring may be present in a deployed system without having been a deliberate institutional choice at all. Continuity and confidentiality guarantees should be calibrated to the intensity of relational design actually present, not to a generic data-protection baseline, an implication consistent with Nissenbaum's (2004) contextual-integrity framework as applied in Section 4.

Workforce and caseworker backstops matter here specifically. Extending the meaningful-human-control literature (Santoni de Sio and van den Hoven 2018) from decision automation to relational automation, no citizen's only channel to being understood should be a system with no duty of continuity toward them. This argues for guaranteed, low-friction escalation to a human interlocutor as a standing design requirement rather than an exception-handling feature, a discipline the Wysa case in Section 6.2 already applies at its highest-severity threshold and that this paper's argument would extend further down the severity range.

For policy, the clearest existing regulatory hook is contestability. The GDPR's Article 22 already establishes a right to contest automated decisions with legal or similarly significant effects (European Parliament and Council 2016), and the EU AI Act's Article 5(1)(f) already prohibits emotion inference in workplace and education settings as an unacceptable-risk, prohibited practice (European Parliament and Council 2024). Neither provision currently extends contestability to the accuracy or adequacy of a system's relational interpretation of a citizen, as distinct from the decision the system reaches.

Extending contestability grounds to cover relational-interpretation claims, not only decisional ones, would close the fourth row of Table 1 directly and would track the direction commercial-AI-companion regulation is already moving in response to the litigation discussed in Section 6.3. The High-Level Expert Group on Artificial Intelligence's (2019) trustworthy-AI guidelines and Mišić, van Est and Kool's (2025) governance value framework both provide general normative scaffolding for public-sector AI but do not yet specify relational-design parity as a distinct governance requirement, a gap this paper's framework is positioned to fill. Hagendorff's (2020) critical evaluation of AI ethics guidelines more broadly cautions that voluntary guideline frameworks tend toward underspecification exactly where enforcement would

be costly, a caution directly applicable to relational-design parity given the institutional cost asymmetry identified in Section 5.

An empirical measurement agenda follows directly from the framework. Detecting parasocial intensity in citizen-facing public AI, through survey instruments and interaction-log analysis, would extend the measurement tradition already established around automation bias and selective adherence to algorithmic advice in public administration (Alon-Barkat and Busuioc 2022) to the relational register this paper addresses. Such instruments would allow institutions to detect drift toward the parasocial-exposure category identified in Section 5 before it produces the kind of documented harm reviewed in Section 6.3, rather than discovering it only after litigation or regulatory intervention.

8. LIMITATIONS AND FUTURE RESEARCH

This paper is conceptual, and its illustrative material, while drawn exclusively from documented, sourced cases, is not a systematic sample of public-sector deployments of socially intelligent AI. The five cases in Section 6 were selected because they are well-documented and institutionally varied, not because they represent a random or exhaustive cross-section of the field. Systematic mapping of relational design intensity and accountability parity across a larger set of public deployments, using the coding scheme illustrated in Table 2, is a natural next step and would allow the typology in Section 5 to be tested against a broader evidence base than the present paper provides.

The framework also does not yet specify how accountability parity should be weighted when the five rows of Table 1 point in different directions, for instance when a system achieves strong duty-of-care infrastructure, as in the Wysa case, but weaker confidentiality specification for lower-severity interactions. Future work could develop a composite parity index, drawing on the measurement tradition in public-administration research on automated advice (Alon-Barkat and Busuioc 2022), to allow comparative assessment across systems rather than the qualitative case-by-case judgment offered here.

Finally, the construct is developed from an institutional and design-analytic standpoint rather than from direct evidence of citizens' own understanding of the accountability gap they may or may not perceive themselves to be exposed to. Qualitative and survey work directly asking users of public-sector relational AI, particularly in the elder-care and mental-health domains discussed in Section 6, how they understand the continuity, confidentiality, and contestability commitments of the systems they rely on would substantially strengthen the empirical foundation for the claims made here, and represents the most direct extension of this paper's argument into primary research.

9. CONCLUSION

Social intelligence in artificial intelligence design, the capacity to read affect, sustain conversational continuity, and simulate empathic responsiveness, functions as a relationship-generating design choice rather than as neutral scaffolding layered onto an otherwise unchanged service. The psychological

literature reviewed in this paper indicates that the relational response it generates is close to automatic and does not depend on user sophistication or institutional intent.

Public and quasi-public institutions that adopt socially intelligent AI therefore inherit relational obligations that current governance architecture has not yet been built to discharge. The partial accountability parity achieved in the Wysa case shows that the relational accountability gap identified here is not an inevitable feature of relational design. It is, however, the default outcome where relational features are adopted without a matching institutional investment, a pattern the ElliQ case and the wider commercial-AI-companion litigation reviewed in Section 6 both suggest is currently the more common one.

The stakes of closing this gap fall on welfare, health, and education systems, where socially intelligent AI is being adopted fastest. The citizens interacting with it in these settings, including isolated older adults, people in mental-health crisis, and children, are least able to bear the cost of a relationship that the institution invited without building the accountability architecture to support it.

Declarations

ETHICS DECLARATION

This is a conceptual paper. It develops a theoretical framework and interprets material already in the public domain (published research, regulatory decisions, and news reporting on the cases discussed in Section 6); it did not involve human participants, human data, animal subjects, or the collection or analysis of primary or secondary empirical data. Consequently, no approval from an institutional ethics committee or review board was required, and no ethical guidelines beyond standard research integrity and citation practice apply to this work.

Competing Interests:

The author declares no competing interests.

Funding:

No funding was received for this research.

Data availability: Not applicable. This is a conceptual paper and does not report new data.

Author Contribution

sole author - n/a

References

1. Alon-Barkat S, Busuioc M (2022) Human-AI interactions in public sector decision making: 'Automation bias' and 'selective adherence' to algorithmic advice. *J Public Adm Res Theor*. <https://doi.org/10.1093/jopart/muac007>
2. Alston P (2019) Report of the Special Rapporteur on extreme poverty and human rights (the digital welfare state). UN Doc A /74/48037. https://srpoverty.org/wp-content/uploads/2019/10/a_74_48037_advanceuneditedversion-1.pdf
3. Amnesty International (2024) France: Discriminatory algorithm used by the social security agency must be stopped. <https://www.amnesty.org/en/latest/news/2024/10/france-discriminatory-algorithm-used-by-the-social-security-agency-must-be-stopped/>
4. Aledhari M, Razzak R, Parizi RM, Saeed F (2020) Federated learning: A survey on enabling technologies, protocols, and applications. *IEEE Access*. <https://doi.org/10.1109/access.2020.3013541>
5. Andrejevic M, Selwyn N (2020) Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology*. <https://doi.org/10.1080/17439884.2020.1686014>
6. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform Fusion* 58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
7. Bielefeld S, Harb JI, Henne K (2021) Financialization and welfare surveillance: Regulating the poor in technological times. *Surveillance Soc* 19(3). <https://doi.org/10.24908/ss.v19i3.14244>
8. Blut M, Wang C, Wunderlich NV, Brock C (2021) Understanding anthropomorphism in service provision: A meta-analysis. *J Acad Mark Sci*. <https://doi.org/10.1007/s11747-020-00762-y>
9. Brandtzaeg PB, Skjuve M, Følstad A (2022) My AI friend: How users of a social chatbot understand their human-AI friendship. *Hum Commun Res*. <https://doi.org/10.1093/hcr/hqac008>
10. Cerna-Aragón D, García L (2025) Making the eyes of the state: Algorithmic alienation and mundane creativity in Peruvian street-level bureaucrats. *Policy Sci* 58(1):69–86. <https://doi.org/10.1007/s11077-025-09570-z>
11. Citron DK (2008) Technological due process. *Wash Univ Law Rev* 85(6):1249
12. Citron DK, Pasquale F (2014) The scored society: Due process for automated predictions. *Wash Law Rev* 89(1):1
13. CNN Business (2026), January 7 Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides. <https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit>
14. Dencik L, Kaun A (2020) Datafication and the welfare state. *Global Perspect* 1(1). <https://doi.org/10.1525/gp.2020.12912>
15. Diakopoulos N (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digit Journalism* 3(3):398–415

16. Duggan J, Sherman U, Carbery R, McDonnell A (2020) Algorithmic management and app-work in the gig economy: A research agenda for employment relations and HRM. *Hum Resource Manage J* 30(1):114–132. <https://doi.org/10.1111/1748-8583.12258>
17. Elish MC (2019) Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Sci Technol Soc*, 5. <https://estsjournal.org/index.php/ests/article/view/260>
18. Eubanks V (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press / Picador
19. European Parliament and Council (2016) Regulation (EU) 2016/679 (General Data Protection Regulation), Article 22
20. European Parliament and Council (2024) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 5(1)(f)
21. Fisher JF (2025), January 21 To bot or not to bot? How AI companions are reshaping human services and connection. *Stanford Social Innovation Review*. <https://doi.org/10.48558/BGF6-HF74>
22. Friedman B, Hendry DG (2019) *Value sensitive design: Shaping technology with moral imagination*. MIT Press
23. Grimmelikhuijsen S (2022) Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. *Public Adm Rev*. <https://doi.org/10.1111/puar.13483>
24. Hagendorff T (2020) The ethics of AI ethics: An evaluation of guidelines. *Mind Mach* 30:99–120. <https://doi.org/10.1007/s11023-020-09517-8>
25. High-Level Expert Group on Artificial Intelligence (2019) *Ethics guidelines for trustworthy AI*. European Commission. <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>
26. Horton D, Wohl RR (1956) Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry* 19(3):215–229. <https://doi.org/10.1080/00332747.1956.11023049>
27. Jarrahi MH, Newlands G, Lee MK, Wolf CT, Kinder E, Sutherland W (2021) Algorithmic management in a work context. *Big Data Soc* 8(2). <https://doi.org/10.1177/20539517211020332>
28. Kaissis G, Makowski MR, Rückert D, Braren R (2020) Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2:305–311. <https://doi.org/10.1038/s42256-020-0186-1>
29. Kaun A, Männiste M (2025) Public sector chatbots: AI frictions and data infrastructures at the interface of the digital welfare state. *New Media & Society*. <https://doi.org/10.1177/14614448251314394>
30. Kostka G (2019) China's social credit systems and public opinion: Explaining high levels of approval. *New Media Soc* 21(7):1565–1593. <https://doi.org/10.1177/1461444819826402>
31. Laux J (n.d.) (ed) *Overseeing like a state: How agentic AI changes public service and its epistemic relationship with citizens*. Digi-Con: The Digital Constitutionalist (Oxford Internet Institute).

- <https://digi-con.org/overseeing-like-a-state-how-agentic-ai-changes-public-service-and-its-epistemic-relationship-with-citizens/>
32. Lee CS (2019) Datafication, dataveillance, and the social credit system as China's new normal. *Online Inf Rev* 43(6):952–970. <https://doi.org/10.1108/oir-08-2018-0231>
 33. Liang F, Das V, Kostyuk N, Hussain MM (2018) Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. *Policy Internet* 10(4):415–453. <https://doi.org/10.1002/poi3.183>
 34. Lighthouse R, Le Monde (2023), December France's digital inquisition. Lighthouse Reports. <https://www.lighthousereports.com/investigation/frances-digital-inquisition/>
 35. Maeda T, Quan-Haase A (2024) When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3630106.3658956>
 36. McStay A (2018) *Emotional AI: The rise of empathic media*. Sage
 37. Mišić J, van Est R, Kool L (2025) Good governance of public sector AI: A combined value framework for good order and a good society. *AI Ethics* 5(5):4875–4889. <https://doi.org/10.1007/s43681-025-00751-3>
 38. Muldoon J, Raekstad P (2022) Algorithmic domination in the gig economy. *Eur J Political Theory*. <https://doi.org/10.1177/14748851221082078>
 39. Nakagomi A, Akutsu Y, Yasuoka M, Abe N, Ihara S, Teroh T, Tabuchi T (2026) AI companions and subjective well-being: Moderation by social connectedness and loneliness. *Technology in Society*, 85. <https://www.sciencedirect.com/science/article/pii/S0160791X26000187>
 40. New York State Office for the Aging (2023), August 1 NYSOFA's rollout of AI companion robot ElliQ shows 95% reduction in loneliness. <https://aging.ny.gov/news/nysofas-rollout-ai-companion-robot-elliq-shows-95-reduction-loneliness>
 41. NHS Innovation Accelerator DigitalHealth.London. (n.d.). Wysa: AI-powered mental health support, integrated into NHS Talking Therapies services (DCB0129 clinical safety certified). <https://www.wysa.com/nhs-talking-therapies>
 42. Nissenbaum H (2004) Privacy as contextual integrity. *Wash Law Rev* 79(1):119–158
 43. Obermeyer Z, Powers BW, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453. <https://doi.org/10.1126/science.aax2342>
 44. Pasquale F (2015) *The black box society: The secret algorithms that control money and information*. Harvard University Press. <https://doi.org/10.4159/harvard.9780674736061>
 45. Raji ID, Buolamwini J (2019) Actionable auditing. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. <https://doi.org/10.1145/3306618.3314244>
 46. Raji ID, Kumar IE, Horowitz A, Selbst AD (2022) The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.

<https://doi.org/10.1145/3531146.3533158>

47. Reeves B, Nass C (1996) The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press / CSLI
48. Roemmich K, Schaub F, Andalibi N (2023) Emotion AI at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3544548.3580950>
49. Royal Commission into the Robodebt Scheme (2023) Report of the Royal Commission into the Robodebt Scheme. Commonwealth of Australia.
<https://robodebt.royalcommission.gov.au/publications/report>
50. de Santoni F, van den Hoven J (2018) Meaningful human control over autonomous systems: A philosophical account. *Front Rob AI* 5., Article 15. <https://doi.org/10.3389/frobt.2018.00015>
51. South China Morning Post (2019) Chinese primary school stops using headbands to study pupils' concentration levels after public outcry.
<https://www.scmp.com/news/china/society/article/3035983/chinese-primary-school-stops-using-headbands-study-pupils>
52. TechCrunch (2023), February 3 Replika, a 'virtual friendship' AI chatbot, hit with data ban in Italy over child safety. <https://techcrunch.com/2023/02/03/replika-italy-data-processing-ban/>
53. Tsaban S (under review-a). Carewashing the algorithm: Care rhetoric and the legitimization of algorithmic control. Unpublished manuscript
54. Tsaban S (under review-b). Algorithmic management as a disclosure and governance problem: A dual-principal framework. Unpublished manuscript
55. Turkle S (2011) Alone together: Why we expect more from technology and less from each other. Basic Books
56. Ullstein C, Engelmann S, Papakyriakopoulos O, Hohendanner M, Großklags J (2022) AI-competent individuals and laypeople tend to oppose facial analysis AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. <https://doi.org/10.1145/3551624.3555294>
57. Urquhart L, Miranda D (2021) Policing faces: The present and future of intelligent facial surveillance. *Inform Commun Technol Law*. <https://doi.org/10.1080/13600834.2021.1994220>
58. van Bekkum M, Zuiderveen Borgesius F (2021) Digital welfare fraud detection and the Dutch SyRI judgment. *Eur J Social Secur* 23(4):323–340. <https://doi.org/10.1177/13882627211031257>
59. van Toorn G, Henman P, Soldatić K (2024) Introduction to the digital welfare state: Contestations, considerations and entanglements. *J Sociol*. <https://doi.org/10.1177/14407833241260890>
60. van Zoonen L (2020) Data governance and citizen participation in the digital welfare state. *Data Policy* 2. <https://doi.org/10.1017/dap.2020.10>
61. Vogl TM, Seidelin C, Ganesh B, Bright J (2020) Smart technology and the emergence of algorithmic bureaucracy: Artificial intelligence in UK local authorities. *Public Adm Rev* 80(6).
<https://doi.org/10.1111/puar.13286>

62. Xu X, Kostka G, Cao X (2022) Information control and public support for social credit systems in China. *J Politics* 84(4):2230–2245. <https://doi.org/10.1086/718358>
63. Zajko M (2023) Automated government benefits and welfare surveillance. *Surveillance Soc* 21(3). <https://doi.org/10.24908/ss.v21i3.16107>
64. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F (2019) Systematic review of research on artificial intelligence applications in higher education - where are the educators? *Int J Educational Technol High Educ* 16(1). <https://doi.org/10.1186/s41239-019-0171-0>