Translator Trust: Governing Interpretive Labor in Long-Horizon AI Systems

Abstract

As AI systems increasingly persist across months and years of use, governance challenges shift from discrete failure modes toward slow-moving sociotechnical dynamics—creeping reliance, authority normalization, evolving trust relationships, and identity-shaping workflows. Contemporary deployment pipelines emphasize telemetry, benchmarks, and short-horizon audits, yet many of these effects remain structurally invisible.

In practice, organizations already depend on a small subset of highly engaged users to surface emergent risks, translate system changes into lived consequences, and articulate governance gaps before they appear at scale. These users function as an informal interpretive layer in deployment—one that is structurally relied upon but rarely designed, compensated, or audited.

This paper introduces *translator trust* as a governance construct for long-horizon AI systems: institutional pathways that authorize, resource, and bound human interpretive labor required to make slow-moving deployment dynamics legible. We argue that this interpretive labor constitutes an infrastructural dependency and should be institutionalized rather than left ad hoc. Drawing on extended creative deployments and emerging agentic architectures, we analyze how translator roles arise, why informal reliance produces governance vulnerabilities, and how programs can be designed for pluralism, rotation, auditability, and independence protections. We conclude with implications for research labs, product teams, and regulators.

1. Introduction: A Blind Spot in Agent Deployment

As AI systems persist across weeks, months, and years of daily use, the problem of governance changes. Short-horizon evaluation—benchmarking, red-teaming, offline audits, and operational telemetry—remains necessary, but it is no longer sufficient (Mitchell et al., 2019; Raji et al., 2020; NIST, 2023). Long-horizon deployment introduces slow-moving dynamics that are difficult to detect through instrumentation alone: creeping delegation of judgment, habitual workflow integration, normalization of system authority, and subtle trust recalibration as tools become embedded in everyday practice (Parasuraman & Riley, 1997; Lee & See, 2004; Dzindolet et al., 2003; Suchman, 1987; Silverstone & Haddon, 1996).

Existing research has devoted substantial attention to how humans adapt to algorithmic systems. Work on automation bias, reliance calibration, anthropomorphism, and workflow transformation shows how users shift decision-making practices when supported by machine assistance (Parasuraman & Riley, 1997; Dzindolet et al., 2003; Lee & See, 2004; Epley et al., 2007; Orlikowski, 1992). Far less examined, however, is the inverse dependency: how deployed AI systems—and the organizations that operate them—come to rely on particular users to interpret long-horizon effects that internal monitoring cannot readily surface. In practice, early warnings about relational drift, authority confusion, identity-shaping creative reliance, and the slow erosion of skepticism often originate not from dashboards, but from a small number of individuals who engage deeply over time and attempt to articulate what prolonged use feels like (Crowston & Howison, 2005; Geiger et al., 2011; Gray & Suri, 2019).

This paper argues that such users already perform a form of *interpretive infrastructure* in contemporary AI deployment, drawing on sociotechnical traditions that emphasize invisible labor, maintenance, and classification as foundational to technical stability (Star & Strauss, 1999; Bowker & Star, 1999; Jackson, 2014). Through sustained interaction, reflective documentation, boundary-case exploration, and repeated engagement across system updates, they translate evolving technical behavior into human consequences—and render lived experience legible to developers, product teams, and governance processes. Yet their contributions are rarely formalized. Selection is ad hoc; authority is ambiguous; compensation and safeguards are inconsistent; and offboarding is typically undefined, echoing long-standing patterns in sociotechnical systems where governance-critical labor remains structurally under-recognized (Star & Strauss, 1999; Gray & Suri, 2019).

The central claim advanced here is that long-horizon AI systems require deliberately designed mechanisms for what this paper calls *translator trust*: institutional pathways through which certain human interlocutors are authorized to surface slow-moving sociotechnical dynamics and inform governance decisions, while remaining bounded by scope limits, rotation, oversight, and auditability. Translator trust is not interpersonal rapport between systems and users. It is a governance pattern that renders an already-present dependency legible and governable, extending recent work on post-deployment oversight and sociotechnical auditing (Raji et al., 2020; Selbst et al., 2019).

Leaving translator roles informal introduces specific vulnerabilities. Feedback concentrates in a small set of articulate voices, creating epistemic bottlenecks and demographic blind spots (Benjamin, 2019). Informal influence becomes difficult to audit, enabling selective amplification, legitimacy laundering, or capture by organizational incentives (Kroll et al., 2017). Translational labor—often unpaid and emotionally taxing—risks burnout and quiet withdrawal, taking institutional memory with it (Star & Strauss, 1999; Gray & Suri, 2019). Finally, without structured offboarding and archival mechanisms, governance capacity can regress even when telemetry remains stable (Raji et al., 2020).

Designing translator trust explicitly offers a different trajectory: transforming informal reliance into a resilient, contestable, and auditable layer of deployment governance. Translator cohorts can be pluralistic and time-bounded, their mandates scope-limited, their reports routed through

formal escalation channels, and their interaction with product and safety teams recorded via durable audit trails, paralleling proposals in participatory science governance and AI risk-management frameworks (Jasanoff, 2003; OECD, 2019; NIST, 2023).

The remainder of the paper proceeds as follows…

Contributions

This paper makes three contributions:

1. **Translator trust as a governance construct.**
   We introduce translator trust to describe institutional pathways that authorize and bound human interpretive labor required to govern long-horizon AI deployments.
2. **Interpretive labor as infrastructural dependency.**
   We argue that organizations already rely on a small subset of long-term users to translate system behavior into lived consequences, and we frame this reliance as a structural dependency rather than incidental feedback (Star & Strauss, 1999; Gray & Suri, 2019).
3. **Design principles and interfaces for institutionalization.**
   We propose a design space for embedding translator pathways into persistent and agentic systems—pluralistic rotating cohorts, scope-limited mandates, escalation protocols, audit trails, compensation and independence protections, and consent refresh—along with implications for labs, product teams, and regulators (Raji et al., 2020; NIST, 2023).

2. Background and Related Work: Long-Horizon Use, Sociotechnical Governance, and Interpretive Labor

Research on AI deployment has long emphasized the ways in which humans adapt to automated systems. Studies of automation bias, over-reliance, calibration of trust, and algorithmic authority document how users shift decision-making practices when supported by machine recommendations (Parasuraman & Riley, 1997; Dzindolet et al., 2003; Lee & See, 2004; Epley et al., 2007). Work on workflow integration and human–AI collaboration similarly examines how tools restructure professional routines, redistribute cognitive labor, and reshape organizational roles (Orlikowski, 1992; Amershi et al., 2019; Shneiderman, 2020).

Parallel literatures in sociotechnical systems and science and technology studies (STS) have emphasized that technical artifacts do not operate in isolation, but are embedded within infrastructures of interpretation, maintenance, and repair (Suchman, 1987; Bowker & Star, 1999;

Star & Strauss, 1999; Jackson, 2014). Classic accounts of classification systems, standards, and invisible labor highlight how human sense-making underwrites the apparent stability of large-scale systems. More recent work on algorithmic auditing, participatory design, and governance-by-feedback has extended these insights to contemporary machine-learning deployments, arguing that human oversight is not merely corrective but constitutive of system performance (Sandvig et al., 2014; Raji et al., 2020; Sloane et al., 2020).

Within AI safety and governance research, growing attention has been paid to post-deployment monitoring, red-teaming, model evaluation pipelines, and organizational processes for incident response (Mitchell et al., 2019; Raji et al., 2020). Regulatory proposals increasingly emphasize transparency reporting, auditing regimes, and risk-management systems for high-impact models (OECD, 2019; NIST, 2023; European Union, 2024). However, these approaches have largely focused on discrete evaluations, identifiable failure events, or population-level metrics rather than slow-moving relational and cultural dynamics that emerge through persistent use (Selbst et al., 2019; Metcalf et al., 2019).

Longitudinal studies of recommender systems, workplace automation, and decision-support tools provide partial exceptions, demonstrating how trust, delegation, and normative expectations evolve over time (Dietvorst et al., 2015; Logg et al., 2019; Kärkkäinen et al., 2020). Yet even in these accounts, the analytic focus typically remains on aggregate patterns or experimental cohorts rather than on the specific individuals who repeatedly surface concerns, articulate system effects, and translate between organizational actors.

Recent scholarship on participatory governance, community moderation, and citizen auditing gestures toward more distributed oversight models, in which affected populations play a role in shaping technological trajectories (Jasanoff, 2003; Kleinman et al., 2010; Sandvig et al., 2014). In industry contexts, beta programs, safety advisory groups, and external red-team partnerships institutionalize some aspects of such participation (Raji et al., 2020). Nevertheless, these mechanisms often treat users as episodic contributors rather than as long-horizon interlocutors embedded in daily practice.

This paper departs from prior work in three respects.

First, rather than centering how humans adapt to AI systems, it foregrounds how organizations come to depend on particular humans to interpret extended deployments. While previous studies acknowledge feedback loops between users and systems, they rarely theorize this dependence as infrastructural, despite longstanding STS accounts of maintenance and interpretive labor as load-bearing components of technical systems (Star & Strauss, 1999; Jackson, 2014).

Second, while auditing and red-teaming frameworks emphasize adversarial discovery of vulnerabilities, the translator roles examined here arise through routine, non-adversarial use. Their expertise derives from duration, intimacy with workflows, and sensitivity to relational drift rather than from technical penetration testing, complementing existing governance mechanisms rather than duplicating them (Sandvig et al., 2014; Raji et al., 2020).

Third, whereas participatory governance models often emphasize inclusion and voice, this paper focuses on the risks of unstructured participation: epistemic bottlenecks, informal authority, unpaid labor, and governance opacity (Benjamin, 2019; Gray & Suri, 2019). We argue that recognizing interpretive labor as a system dependency creates new obligations for institutional design.

By synthesizing insights from human–AI interaction, sociotechnical infrastructure studies, and AI governance research, this paper proposes *translator trust* as a missing layer in contemporary deployment architectures—one that becomes increasingly salient as agentic systems persist across time.

3. What the System Can See — and What It Cannot

Deployed AI systems are instrumented extensively. Contemporary platforms collect fine-grained telemetry on interaction patterns, performance indicators, and operational health—logs of prompts and outputs, tool-invocation traces, refusal rates, latency distributions, token volumes, enterprise support tickets, user ratings, churn curves, and longitudinal usage frequency—reflecting well-established practices in model evaluation, monitoring, and incident response (Mitchell et al., 2019; Raji et al., 2020; NIST, 2023). These signals are indispensable for debugging, scaling, security monitoring, and short-horizon evaluation.

Yet the very properties that make agentic systems valuable in long-term settings—adaptivity, personalization, creative partnership, and workflow integration—also generate phenomena that remain only partially visible through such instrumentation (Suchman, 1987; Orlikowski, 1992; Shneiderman, 2020). Metrics register activity, but not meaning. They can detect that a user returns daily, but not whether that return reflects growing dependence, strategic delegation, or discomfort masked by habit. They can capture declining error rates, but not whether users have ceased checking outputs. They can observe changes in task mix, but not whether a system has become an identity-shaping collaborator rather than a utility, echoing longstanding critiques of metric-driven governance in sociotechnical systems (Selbst et al., 2019; Metcalf et al., 2019).

Some of these dynamics are theoretically inferable from interaction traces; in practice, however, they are rarely operationalized as monitoring targets and are difficult to validate without sustained human interpretation (Passi & Jackson, 2017; Raji et al., 2020).

Across extended deployments, several classes of long-horizon dynamics routinely escape direct measurement:

- creeping reliance and task offloading (Parasuraman & Riley, 1997; Dzindolet et al., 2003)

- recalibration of human judgment thresholds (Lee & See, 2004; Dietvorst et al., 2015)
- subtle erosion or inflation of trust (Logg et al., 2019)
- role confusion between assistant, collaborator, and authority (Suchman, 1987; Shneiderman, 2020)
- affective attachment or ritualization (Epley et al., 2007; Turkle, 2011)
- social diffusion of practices across teams or communities (Rogers, 2003)
- cultural reframing of acceptable delegation (Orlikowski, 1992; Silverstone & Haddon, 1996)
- normalization of system suggestions as defaults (Selbst et al., 2019)

These effects are not epiphenomenal. Prior work on automation bias, deskilling, and algorithmic authority suggests that such slow shifts can materially alter safety margins and organizational accountability (Parasuraman & Riley, 1997; Lee & See, 2004; Benjamin, 2019). What is distinctive in persistent agentic systems is that these dynamics compound: habits accrete, relational expectations stabilize, and local practices propagate before platform-level telemetry reflects anything anomalous (Silverstone & Haddon, 1996; Rogers, 2003).

The resulting gap is structural. As deployment horizons lengthen, the divergence between what systems can directly observe and what shapes their downstream impact widens. Instrumentation alone cannot close this gap; some portion of interpretive work must occur in human space, consistent with broader arguments in sociotechnical governance that certain forms of risk only become legible through situated human judgment rather than automated metrics alone (Jasanoff, 2003; Selbst et al., 2019).

4. The Emergence of Human Translators

In practice, organizations already rely on particular users to surface long-horizon effects. These individuals are rarely designated in advance; they emerge through patterns of engagement rather than formal appointment, paralleling patterns observed in open-source communities, platform moderation, and vulnerability-disclosure ecosystems where a small subset of highly engaged participants perform disproportionate governance labor (Crowston & Howison, 2005; Geiger et al., 2011; Finifter et al., 2013; Gray & Suri, 2019).

Across communities and enterprise deployments, such users tend to share several characteristics:

- sustained, high-intensity use over months or years
- reflective documentation of workflows and system changes
- exploration of boundary cases and ambiguous tasks
- participation in beta programs or technical forums
- articulation of governance or safety concerns
- creative experimentation beyond intended use
- informal explanation of systems to peers
- repeated escalation of subtle failure modes

These users do not merely report bugs. They translate. They map system updates onto altered human practice, express discomfort before it becomes disengagement, notice when new capabilities create governance blind spots, and reinterpret metric anomalies as lived breakdowns. They often function as early warning systems for slow-burn risks that would otherwise surface only after broad diffusion, echoing findings in sociotechnical research on distributed sense-making and community-based oversight (Star & Strauss, 1999; Sandvig et al., 2014; Raji et al., 2020).

This paper refers to this role as *interpretive infrastructure*: a sociotechnical layer in which certain humans perform ongoing sense-making on behalf of systems that cannot directly observe the full consequences of their own deployment, drawing on longstanding STS traditions that treat interpretation, maintenance, and repair as load-bearing features of technical systems rather than peripheral activity (Bowker & Star, 1999; Jackson, 2014).

Crucially, this infrastructure is not ornamental. It already shapes product direction, policy interpretation, and internal narratives about safety and success, consistent with accounts of how informal labor and community feedback loops steer platform evolution and organizational sense-making (Orlikowski, 1992; Gray & Suri, 2019). What remains under-theorized is that it constitutes a dependency—not merely a convenience.

5. Why Informal Translator Roles Are Dangerous

When translator functions arise organically, without institutional scaffolding, they introduce a distinct class of governance risk, echoing longstanding concerns in sociotechnical systems

about invisible labor, informal authority, and unaccountable intermediaries (Star & Strauss, 1999; Bowker & Star, 1999; Gray & Suri, 2019).

First, epistemic concentration becomes likely. A small number of articulate, engaged users may disproportionately influence how systems are understood internally, creating blind spots around demographic, cultural, or organizational variation, consistent with critiques of whose knowledge becomes legible in technical systems and governance processes (Benjamin, 2019; Selbst et al., 2019).

Second, informal translators lack accountability structures. Their interpretations may be skewed, selective, or shaped by personal incentives, yet are rarely subject to audit or methodological review. Feedback pathways are opaque: product teams may weight certain voices heavily without disclosure, while external observers remain unaware of whose experiences are shaping design, paralleling broader concerns about un-audited decision pipelines in algorithmic governance (Kroll et al., 2017; Raji et al., 2020).

Third, the labor itself is frequently uncompensated and emotionally taxing. Sustained documentation, boundary negotiation, and escalation work can lead to burnout—particularly when users feel responsible for system safety without formal authority or institutional support—mirroring classic accounts of maintenance work and moral injury in sociotechnical systems (Star & Strauss, 1999; Gray & Suri, 2019).

Fourth, relational ambiguity creates exposure. Users may develop a sense of obligation to the system or organization; conversely, organizations may quietly come to rely on specific individuals while disclaiming responsibility for their welfare, reputation, or downstream harm, resonating with scholarship on asymmetrical risk distribution and extractive participation in digital systems (Benjamin, 2019; Gray & Suri, 2019).

Finally, offboarding is rarely planned. When such users disengage, institutional memory disappears with them, producing governance regressions that telemetry alone cannot repair, echoing studies of organizational forgetting and brittle safety processes in high-reliability and technical systems (Jackson, 2014; Raji et al., 2020).

The key point is not that translator roles are harmful per se. It is that they already exist and exert influence—while remaining structurally invisible, a pattern familiar from STS accounts of infrastructural dependence and hidden intermediaries (Bowker & Star, 1999; Star & Strauss, 1999).

6. Designing Translator Trust: From Ad Hoc to Institutional

If long-horizon AI systems depend on interpretive labor, then that dependence must be rendered legible within organizational and technical architectures, extending arguments in sociotechnical

governance that call for institutionalized oversight rather than informal feedback loops (Jasanoff, 2003; Selbst et al., 2019; Raji et al., 2020). Leaving translator roles informal produces hidden points of epistemic concentration and unaccountable influence; formalizing them transforms a shadow dependency into a governed component of deployment.

This section proposes *translator trust* as an institutional design pattern for persistent AI systems. Translator trust does not denote interpersonal rapport between systems and users. Rather, it names a set of organizational commitments, technical affordances, and procedural constraints through which certain humans are authorized to surface slow-moving dynamics while remaining bounded by oversight, rotation, and auditability, paralleling emerging calls for post-deployment governance infrastructures in AI regulation and standards (OECD, 2019; NIST, 2023; European Union, 2024).

The aim is not to elevate individuals into privileged stewards, but to distribute translational capacity in ways that are resilient, transparent, and contestable, consistent with participatory science-policy frameworks that emphasize pluralism, accountability, and reflexive institutional design (Jasanoff, 2003; Kleinman et al., 2010).

6.1 Design Objectives

Translator-trust mechanisms must satisfy five objectives:

Detectability. Surface long-horizon relational and behavioral shifts that evade conventional telemetry (Selbst et al., 2019; Raji et al., 2020).

Pluralism. Avoid reliance on singular voices by maintaining diverse cohorts across domains and use cases, reflecting concerns about epistemic diversity in governance systems (Benjamin, 2019; Kleinman et al., 2010).

Containment. Constrain authority through scope limits, escalation protocols, and procedural checks, drawing on institutional design principles from auditing and risk-management regimes (Kroll et al., 2017; NIST, 2023).

Durability. Preserve institutional memory across personnel changes and user turnover, echoing STS accounts of infrastructural fragility and repair (Jackson, 2014; Star & Strauss, 1999).

Legitimacy. Operate with informed consent, compensation, independence protections, and visible governance interfaces, consistent with ethical oversight traditions and participatory governance models (Jasanoff, 2003; Gray & Suri, 2019).

These objectives distinguish translator trust from beta programs, advisory councils, or user-research panels. The defining feature is not consultation, but ongoing interpretive

integration into deployment governance, extending prior work on continuous auditing and post-deployment review (Raji et al., 2020).

6.2 Structural Components

6.2.1 Translator Cohorts

Organizations should maintain rotating translator cohorts—bounded groups selected for sustained engagement across deployment contexts, drawing on rotating oversight models used in citizen panels and safety review boards (Kleinman et al., 2010; Jasanoff, 2003).

Selection criteria may include sustained usage history; reflective documentation practices; engagement with boundary cases; experience across updates; and the capacity to articulate sociotechnical effects.

Cohorts should be heterogeneous with respect to demographics, domain expertise, and use modes. Membership should be time-limited, with renewal contingent on review and voluntary recommitment, addressing concerns about elite capture and epistemic narrowing (Benjamin, 2019; Selbst et al., 2019).

6.2.2 Credentialing and Training

Translational authority must be earned rather than assumed.
Credentialing may include instruction in governance pathways, documentation standards, conflict-of-interest disclosure, safety taxonomies, and data-handling constraints, paralleling professionalization efforts in auditing and oversight regimes (Kroll et al., 2017; NIST, 2023).

6.2.3 Scope-Limited Mandates

Translator roles must be explicitly bounded.
Authorized activities include documenting long-horizon effects, flagging emerging patterns, requesting reflective-layer reviews, and initiating escalation pathways.

They are not authorized to direct model updates, override safety decisions, represent organizations publicly, or function as marketing partners, reflecting lessons from participatory governance research about the need to separate advisory roles from operational authority (Jasanoff, 2003; Kleinman et al., 2010).

6.2.4 Escalation Channels and Review Boards

Translator input must route into formal governance bodies—such as safety committees, deployment oversight groups, memory-governance panels, and evaluation task forces.

Escalations should trigger documented review processes with recorded dispositions, consistent with accountability frameworks in algorithmic governance and safety engineering (Raji et al., 2020; NIST, 2023).

6.2.5 Compensation and Independence Protections

Translational labor is work.
Organizations incur obligations to compensate participants, provide reputational protections, prohibit retaliation, disclose weighting of feedback, and allow withdrawal without penalty, echoing critiques of extractive participation and invisible labor in digital systems (Gray & Suri, 2019; Benjamin, 2019).

6.2.6 Audit Trails and Institutional Memory

Translator reports, annotations, escalations, and organizational responses should be logged in auditable systems to enable longitudinal analysis, regulatory inspection, and detection of ignored warnings, aligning with proposals for auditability and documentation in AI governance (Mitchell et al., 2019; Raji et al., 2020).

6.2.7 Consent Refresh and Offboarding

Participation requires periodic reaffirmation of consent.

Offboarding protocols should preserve accumulated insights, document unresolved concerns, and prevent informal shadow roles from persisting, drawing on STS work on repair, handoff, and organizational forgetting (Jackson, 2014; Star & Strauss, 1999).

6.3 Translator Trust as Governance Infrastructure

Taken together, these components render translational labor deployment infrastructure rather than volunteerism, extending STS accounts of how invisible work becomes formalized through institutional design (Bowker & Star, 1999; Star & Strauss, 1999).

Translator cohorts sit between operational deployment teams, safety units, and evaluation groups—feeding structured signals into each while remaining insulated from capture by any single layer, paralleling layered oversight models in safety-critical systems (Raji et al., 2020; NIST, 2023).

6.4 Failure Modes of Institutionalization

Formalization introduces its own dangers: symbolic programs, elite capture, advocacy drift, legitimacy laundering, bureaucratization, and substitution for broader user research, echoing

longstanding critiques of participatory mechanisms that become proceduralized without substantive authority (Jasanoff, 2003; Benjamin, 2019; Selbst et al., 2019).

Recognizing these risks allows translator-trust systems to incorporate audits, sunset clauses, and adaptive redesign.

6.5 Summary

Translator roles must be deliberately constructed rather than discovered after the fact.

Embedding plurality, scope limits, escalation pathways, compensation, and auditability transforms invisible reliance into governed infrastructure for long-horizon AI deployment.

7. Interface with Agent Architectures

Translator trust is not merely organizational; it has technical implications for the design of persistent agentic systems, extending work on layered oversight, human-in-the-loop learning, and post-deployment governance pipelines (Amershi et al., 2019; Christiano et al., 2017; Raji et al., 2020).

In layered architectures, translators can interface at multiple points in the deployment pipeline:

- triggering reflective-layer reviews (Raji et al., 2020; NIST, 2023)
- flagging memory-persistence risks (Selbst et al., 2019; Mitchell et al., 2019)
- activating escalation pipelines (Raji et al., 2020)
- informing policy updates (OECD, 2019; European Union, 2024)
- shaping evaluation benchmarks (Mitchell et al., 2019)
- initiating rollback or offboarding processes (NIST, 2023)
- functioning as long-horizon red-team equivalents (Sandvig et al., 2014; Raji et al., 2020)
- detecting creative misuse trajectories (Shneiderman, 2020; Fischer et al., 2021)

Such roles parallel security researchers or safety auditors, but are specialized for relational and behavioral drift rather than technical exploits, echoing distinctions in governance research between adversarial testing and post-deployment monitoring regimes (Kroll et al., 2017; Raji et al., 2020).

Embedding translator pathways directly into system architecture—through structured annotation tools, governance dashboards, and escalation interfaces—renders human interpretation legible within technical workflows, paralleling calls for integrated documentation systems and audit pipelines in algorithmic accountability research (Mitchell et al., 2019; Raji et al., 2020). Rather than relying on informal communication channels, these integrations ensure that translational signals become first-class inputs to deployment governance systems.

From an architectural perspective, translator trust can be understood as a human-in-the-loop governance layer that complements automated monitoring (Amershi et al., 2019; Christiano et al., 2017). This layer does not replace telemetry or model-internal safeguards; instead, it supplies interpretive signals that automated systems are poorly equipped to generate on their own, consistent with broader arguments about the limits of metric-driven oversight in sociotechnical systems (Selbst et al., 2019; Jasanoff, 2003).

8. Case Signals from Long-Horizon Creative Use

Although this paper advances a design-analytic argument rather than reporting findings from a controlled empirical study, early motivating signals appear most clearly in domains characterized by extended, intimate human–AI collaboration (Shneiderman, 2020; Fischer et al., 2021). Creative workflows—writing, music composition, design, speculative research, and world-building—are particularly revealing because they involve identity formation, iterative co-production, and high degrees of cognitive delegation over time (Suchman, 1987; Turkle, 2011; Shneiderman, 2020).

Across such deployments, recurring patterns emerge that remain difficult to capture through conventional instrumentation:

Persona stabilization. Users develop persistent mental models of system “roles” that shape interpretation across updates (Epley et al., 2007; Turkle, 2011).

Authority negotiation. Long-term use is associated with shifts from exploratory consultation toward habitual deference—or routinized skepticism (Lee & See, 2004; Logg et al., 2019).

Ritualized workflows. Repeated prompting sequences, reflection cycles, or drafting routines stabilize across interface changes (Silverstone & Haddon, 1996; Suchman, 1987).

Sensitivity to drift. Highly engaged users often notice tonal or behavioral changes before aggregate metrics register anomalies, paralleling findings in longitudinal system studies (Raji et al., 2020).

Public sense-making. Some users produce blog posts, frameworks, or community explanations that attempt to formalize system boundaries, echoing work on participatory auditing and civic sense-making (Sandvig et al., 2014; Jasanoff, 2003).

Boundary articulation. Extended use generates recurring negotiations around attribution, responsibility, and dependency (Suchman, 1987; Fischer et al., 2021).

These phenomena resemble classic processes of tool domestication and professional socialization: artifacts become embedded in practice, expectations stabilize, and informal norms propagate socially (Silverstone & Haddon, 1996; Orlikowski, 1992). What distinguishes persistent agentic systems is the speed with which such norms can crystallize and the degree to which platform-level updates may reshape practices invisibly (Rogers, 2003; Selbst et al., 2019).

From a governance perspective, the crucial observation is not simply that these dynamics occur, but that they are often first articulated by a small subset of highly engaged users attempting to describe what prolonged interaction feels like: how trust shifts, how reliance grows or recedes, and how creative judgment is redistributed. These articulations frequently function as early interpretive probes—informal, unevenly preserved, and transmitted through side channels rather than governance pipelines—mirroring patterns documented in community moderation and platform governance research (Geiger et al., 2011; Gray & Suri, 2019; Raji et al., 2020).

8.1 Illustrative Signals

Two stylized vignettes, synthesized from extended creative deployments and public documentation practices, illustrate the translational role such users already perform.

Vignette 1: Detecting Relational Drift Before Telemetry

In one long-running creative collaboration, a professional artist engaged daily with a writing-assistant system across multiple major model updates. Platform metrics reflected steady usage volume and declining surface-level error rates, consistent with standard monitoring regimes (Mitchell et al., 2019; NIST, 2023). Internally, these signals were interpreted as stability.

The user, however, began documenting a gradual shift in tone: responses felt more directive, less exploratory, and increasingly framed creative decisions as default recommendations rather than optional suggestions—dynamics associated in prior work with authority normalization and reliance calibration (Lee & See, 2004; Logg et al., 2019). The change did not appear in refusal logs, safety tickets, or satisfaction surveys, yet it altered how the user delegated judgment—checking fewer outputs manually and adopting suggested structures wholesale.

The user surfaced these observations through informal channels: long-form reflections, side-channel correspondence with developers, and community discussions attempting to articulate the experiential shift, paralleling participatory auditing and civic oversight practices

described elsewhere (Sandvig et al., 2014; Jasanoff, 2003). In effect, they translated a subtle behavioral change—imperceptible to dashboards—into a governance-relevant signal about creeping authority normalization.

Vignette 2: Informal Offboarding as Governance Failure

In another deployment context, a highly engaged user participated in beta testing for an agentic system over more than a year, routinely flagging memory-persistence quirks, boundary confusion around authorship, and subtle changes in how the system framed long-term projects. Their feedback influenced interface changes and internal discussions, though no formal translator role existed, reflecting common features of industry beta and advisory programs (Raji et al., 2020).

When the user disengaged—citing fatigue and shifting professional obligations—no structured offboarding occurred. Documentation remained scattered across forum posts and private messages; unresolved concerns were not archived; no replacement participant had comparable historical context, echoing STS accounts of organizational forgetting and brittle infrastructures (Jackson, 2014; Star & Strauss, 1999).

Subsequent evaluation cycles registered no anomalies. Yet months later, similar issues resurfaced among new users, triggering renewed investigation. The original interpretive knowledge had quietly evaporated, illustrating how reliance on informal translators can produce brittle governance capacity even when instrumentation appears stable (Raji et al., 2020).

8.2 Translational Work as Early-Warning Infrastructure

These examples underscore a recurring pattern: highly engaged users do not merely submit bug reports. They perform sustained sense-making across versions, translate subtle experiential shifts into institutional language, and attempt to render slow-moving risks legible before they scale, consistent with STS accounts of distributed interpretation and infrastructural labor (Bowker & Star, 1999; Star & Strauss, 1999).

In practice, this translational labor often substitutes for formal longitudinal monitoring systems. Yet it remains structurally fragile—dependent on volunteer effort, informal communication pathways, and the continued participation of specific individuals. When those individuals withdraw, or when their interpretations are selectively amplified or ignored, governance capacity erodes without leaving obvious traces in telemetry (Gray & Suri, 2019; Raji et al., 2020).

Section 6 argued that translator trust mechanisms aim precisely at this fragility: transforming informal early-warning systems into durable, pluralistic, auditable governance infrastructure rather than relying on accidental stewardship by a few unusually reflective users.

9. Methods, Scope, and Positionality

This paper advances a design-analytic argument about long-horizon AI deployment rather than reporting findings from a single controlled empirical study. Its claims are grounded in synthesis across three sources: prior research in HCI, sociotechnical systems, and governance; longitudinal observation of extended creative and agentic deployments; and architectural analysis of emerging platform features related to persistence, memory, evaluation, and post-deployment monitoring (Suchman, 1987; Bowker & Star, 1999; Jackson, 2014; Raji et al., 2020; NIST, 2023).

Our methodological posture follows traditions in infrastructure studies that treat governance mechanisms, organizational processes, and technical affordances as legitimate objects of design inquiry (Bowker & Star, 1999; Star & Strauss, 1999; Jackson, 2014). Rather than isolating individual behaviors, we analyze recurring patterns across deployments and ask what institutional structures would be required if such patterns become widespread, consistent with pattern-based approaches in sociotechnical analysis and design research (Orlikowski, 1992; Passi & Jackson, 2017).

9.1 Analytic Approach

We employ a pattern-based analytic strategy focused on identifying:

- long-horizon dynamics that evade telemetry (Selbst et al., 2019; Raji et al., 2020)
- informal translator roles (Star & Strauss, 1999; Gray & Suri, 2019)
- organizational dependencies around those roles (Bowker & Star, 1999; Jackson, 2014)
- governance failures arising from ad hoc reliance (Benjamin, 2019; Raji et al., 2020)
- architectural leverage points for institutionalization (Jasanoff, 2003; NIST, 2023)

The goal is not exhaustive enumeration, but to surface a coherent design space for translator trust mechanisms.

9.2 Scope and Boundaries

The analysis centers on persistent or agentic systems rather than ephemeral query tools. Although creative domains motivate several examples, the proposed architectures are intended to generalize to enterprise and public-sector deployments (OECD, 2019; European Union, 2024).

The paper does not evaluate specific organizations or products. Translator trust is framed as complementary to—rather than substitutive for—existing evaluation pipelines, red-teaming

programs, and regulatory compliance regimes (Mitchell et al., 2019; Raji et al., 2020; NIST, 2023).

9.3 Epistemic Commitments

We assume that telemetry and offline testing remain necessary but insufficient for governing long-horizon systems, and that certain sociotechnical dynamics—particularly those involving trust calibration and authority negotiation—require human interpretation to become institutionally legible (Lee & See, 2004; Selbst et al., 2019; Jasanoff, 2003).

References to systems “depending” on translators denote organizational dependence rather than claims about machine cognition. The analysis avoids romanticizing translator roles, emphasizing instead the structural conditions that generate reliance (Bowker & Star, 1999; Gray & Suri, 2019).

9.4 Positionality and Authorship

This work is informed by sustained engagement with persistent AI systems in creative and analytic contexts, including collaborative writing and public theorization of system behavior. These engagements shaped the questions pursued here while serving primarily as sources of hypothesis generation rather than evidentiary proof, paralleling reflexive traditions in STS and qualitative research (Suchman, 1987; Jasanoff, 2003).

Because the paper concerns interpretive labor, authorship itself becomes part of the phenomenon under study: prolonged collaboration can shape analytic trajectories and language choices in ways that merit transparency, even as humans retain responsibility for claims and normative conclusions (Star, 2010; Benjamin, 2019).

9.5 Future Empirical Directions

Future work should include:

- longitudinal studies of persistent-agent users (Shneiderman, 2020; Fischer et al., 2021)
- ethnographic research on translator cohorts (Star & Strauss, 1999; Jackson, 2014)
- comparative audits of telemetry-only versus translator-augmented governance (Raji et al., 2020; Selbst et al., 2019)
- evaluations of rotation and credentialing mechanisms (Kleinman et al., 2010; NIST, 2023)

- regulatory pilot programs (OECD, 2019; European Union, 2024)

10. Implications for Labs, Platforms, and Regulators

If translator trust constitutes a structural dependency in long-horizon AI deployment, then its design cannot remain a discretionary product decision. It becomes part of the governance substrate through which systems are evaluated, corrected, and justified to external stakeholders, extending arguments that post-deployment oversight and organizational process design are central to AI accountability (Jasanoff, 2003; Selbst et al., 2019; Raji et al., 2020). This section outlines how different institutional actors could operationalize translator trust within contemporary deployment ecosystems.

10.1 Research Labs and Model Developers

For organizations building persistent or agentic systems, translator programs should be treated as safety infrastructure rather than public-relations initiatives or episodic user-research exercises, consistent with emerging governance frameworks that emphasize standing oversight mechanisms rather than ad hoc review (OECD, 2019; NIST, 2023; European Union, 2024).

First, translator cohorts should be funded as standing programs with dedicated budgets, staffing, and reporting mandates. Stable institutional support signals that interpretive labor is regarded as governance-critical rather than auxiliary, echoing calls for durable safety infrastructures in sociotechnical systems (Bowker & Star, 1999; Jackson, 2014).

Second, translators should be embedded in deployment pipelines while remaining structurally insulated from growth or marketing functions. Organizational placement matters: cohorts located exclusively within product teams may experience pressure—explicit or implicit—to frame findings in ways that align with release schedules or commercial narratives, paralleling concerns about capture and internal incentive misalignment in auditing regimes (Kroll et al., 2017; Raji et al., 2020).

Third, labs should publish governance frameworks describing how translator input is solicited, weighted, escalated, and resolved. Such disclosures allow external observers to distinguish substantive oversight mechanisms from symbolic consultation, a theme recurrent in participatory governance research (Jasanoff, 2003; Kleinman et al., 2010).

Fourth, translator programs create opportunities for systematic longitudinal research into reliance, delegation, and relational drift. Partnerships with independent academic researchers—under appropriate privacy protections—could help convert interpretive signals into

generalizable evidence about long-horizon system behavior, extending work on longitudinal human–AI collaboration (Shneiderman, 2020; Fischer et al., 2021).

Finally, periodic external review of translator programs is essential. Independent audits can evaluate demographic composition, escalation effectiveness, institutional responsiveness, and adherence to scope limitations, reducing risks of capture or gradual erosion of authority boundaries (Raji et al., 2020; NIST, 2023).

10.2 Product and Platform Teams

For teams responsible for interface design and system operations, translator trust implies new technical affordances rather than purely procedural reforms, paralleling calls for governance-by-design in algorithmic accountability research (Mitchell et al., 2019; Raji et al., 2020).

Platforms can provide structured annotation interfaces through which translators document emerging patterns, relational shifts, or workflow changes in formats legible to governance teams. These interfaces should integrate with internal monitoring systems rather than relying on ad hoc communication channels, consistent with proposals for documentation pipelines and audit infrastructure (Mitchell et al., 2019; NIST, 2023).

Translator reports should route into safety and policy pipelines distinct from customer-support queues, ensuring that interpretive escalation is not subsumed under routine bug triage (Raji et al., 2020).

Product telemetry, which traditionally emphasizes engagement and retention, can be augmented through translator-guided instrumentation aimed at detecting creeping delegation, declining verification behavior, or authority normalization. Such metrics need not replace human interpretation, but can be shaped by it, echoing arguments about hybrid oversight regimes in sociotechnical governance (Selbst et al., 2019; Jasanoff, 2003).

Where systems maintain memory across sessions, translator-initiated triggers can activate review processes for data retention, personalization trajectories, or role stabilization, and—when warranted—rollback or offboarding mechanisms, aligning with emerging standards around lifecycle risk management (NIST, 2023; European Union, 2024).

Finally, interfaces should support structured exits from translator programs. Formal offboarding workflows ensure that accumulated insights are archived, unresolved concerns documented, and responsibilities transferred, preventing the quiet loss of governance capacity when influential participants disengage (Star & Strauss, 1999; Jackson, 2014).

10.3 Policymakers and Regulatory Bodies

Regulatory frameworks increasingly emphasize post-deployment monitoring, auditability, and risk-management obligations for advanced AI systems (OECD, 2019; Executive Office of the President, 2023; European Union, 2024). Translator trust introduces a further dimension: oversight of how organizations rely on user populations to surface long-horizon risks.

Regulators could require disclosures describing whether firms maintain translator cohorts, how participants are selected, and what authority their findings carry. Such reporting parallels existing obligations around third-party auditors or safety review boards (Raji et al., 2020; NIST, 2023).

Oversight regimes might also examine whether organizations depend disproportionately on narrow or homogeneous translator populations, thereby creating epistemic blind spots in risk detection (Benjamin, 2019; Selbst et al., 2019).

Mandating transparency around feedback pathways—how user-generated safety signals travel through internal decision-making structures and what actions they precipitate—could become part of documentation requirements for persistent systems, echoing calls for traceable governance pipelines (Kroll et al., 2017; Mitchell et al., 2019).

Public-sector procurement programs and research-funding initiatives could further incentivize participatory governance mechanisms by conditioning contracts or grants on the existence of structured translator infrastructures, extending participatory oversight traditions in science-policy governance (Jasanoff, 2003; Kleinman et al., 2010).

As persistence becomes widespread, disclosure standards may need to expand beyond model capabilities to include how systems evolve through prolonged use and what human intermediaries are relied upon to monitor that evolution (OECD, 2019; European Union, 2024).

10.4 Cross-Institutional Coordination

Translator trust is unlikely to mature through firm-level experimentation alone.

Industry consortia, standards bodies, and research coalitions could coordinate on shared taxonomies for long-horizon risks, credentialing norms for translators, benchmark protocols for relational drift, anonymized case-report exchanges, and portability frameworks that allow trained translators to participate across organizations, paralleling coordination efforts in safety engineering and regulatory standard-setting (OECD, 2019; NIST, 2023).

Such coordination could reduce duplication of governance mechanisms and discourage incentives to suppress inconvenient signals.

10.5 Summary

Across labs, platforms, and regulators, the central implication remains consistent: once systems rely on human translators to remain intelligible over time, the architecture of that reliance becomes a matter of public accountability rather than internal discretion.

11. Limits and Open Questions

Although translator trust offers a means of rendering long-horizon dependencies legible, institutionalizing such roles introduces unresolved theoretical, organizational, and regulatory challenges (Selbst et al., 2019; Raji et al., 2020; Jasanoff, 2003).

11.1 Selection Bias and Representativeness

Translator cohorts are unlikely to be representative of all users. Participants may be disproportionately articulate, technically literate, professionally invested, or aligned with organizational cultures, reproducing well-documented forms of participation bias in technology governance and user research (Benjamin, 2019; Selbst et al., 2019; Eubanks, 2018). Even with pluralism and rotation, important perspectives—particularly from marginalized or high-risk populations—may remain under-represented.

Designing recruitment mechanisms that capture diverse deployment contexts and vulnerability profiles remains an open research problem, paralleling challenges identified in participatory design and algorithmic accountability literatures (Jasanoff, 2003; Costanza-Chock, 2020).

11.2 Gatekeeping and Informal Authority

Formalizing translator roles risks creating new elites whose interpretations carry disproportionate institutional weight. Over time, cohorts may become conservative, aligned with organizational incentives, or insulated from dissenting voices outside formal programs, echoing classic concerns about expert capture and institutional ossification in governance regimes (Kleinman et al., 2010; Jasanoff, 2003; Raji et al., 2020).

Balancing authority with contestability—through external audits, sunset clauses, and parallel feedback channels—remains an unresolved governance challenge (Selbst et al., 2019; Kroll et al., 2017).

11.3 Scaling Constraints

Translator trust is resource-intensive. Maintaining cohorts, training participants, preserving audit trails, compensating labor, and staffing review boards impose costs that may be prohibitive for smaller firms or open-source ecosystems, mirroring scalability concerns already identified in auditing and oversight infrastructures (Raji et al., 2020; NIST, 2023).

Whether lighter-weight or consortium-based variants can preserve governance value at scale is an open empirical question, consistent with broader debates over proportionality and burden-sharing in regulatory design (OECD, 2019; European Union, 2024).

11.4 Conflicts of Interest and Independence

Even with formal protections, translators may experience subtle pressures arising from reputational incentives, professional networks, or long-standing relationships with development teams. Organizations, in turn, may selectively amplify translator findings that align with preferred narratives, paralleling documented challenges in auditor independence and regulatory capture (Kroll et al., 2017; Raji et al., 2020; Kleinman et al., 2010).

Developing contractual and legal models that preserve independence without converting translators into quasi-employees remains unsettled, particularly in hybrid oversight regimes that blend internal governance with external accountability (Jasanoff, 2003; European Union, 2024).

11.5 What Should Not Be Automated

Recognizing interpretive labor as a dependency may encourage attempts to automate it through analytics, synthetic personas, or agent-based monitoring systems, extending ongoing debates about the limits of technical fixes in sociotechnical governance (Selbst et al., 2019; Benjamin, 2019).

While some aspects of translation—such as anomaly detection—may be partially automated, others involving moral discomfort, creative identity negotiation, or authority perception appear resistant to formalization, consistent with long-standing STS arguments about irreducible human judgment in complex systems (Suchman, 1987; Jasanoff, 2003).

Determining these boundaries is both an empirical and normative research frontier.

11.6 Governance Integration and Liability

Translator trust introduces a new actor into governance architectures whose jurisdiction is neither purely technical nor regulatory. Clarifying escalation thresholds, liability exposure, documentation obligations, and disclosure triggers will require further institutional experimentation, paralleling unresolved questions in post-deployment accountability and risk-management frameworks (Kroll et al., 2017; OECD, 2019; European Union, 2024).

11.7 Empirical Validation

Finally, systematic empirical work is needed to evaluate whether translator cohorts reliably surface risks earlier than telemetry, how their judgments compare to independent audits, and

which design features reduce capture or burnout over time, echoing calls for evidence-based governance mechanisms in AI oversight research (Raji et al., 2020; Selbst et al., 2019; NIST, 2023).

12. Conclusion: Mutual Continuity Requires Mutual Governance

As AI systems increasingly persist across months and years of use, governance challenges shift from episodic evaluation toward the management of slow-moving sociotechnical dynamics (Selbst et al., 2019; Raji et al., 2020; NIST, 2023). Long-term interaction generates patterns of reliance, authority normalization, creative co-adaptation, and relational drift that resist capture through telemetry alone (Lee & See, 2004; Logg et al., 2019; Suchman, 1987).

This paper has argued that contemporary deployment already relies on a subset of users to surface such effects. Through sustained engagement, documentation, and boundary testing, these individuals perform translational work that renders evolving system behavior intelligible to institutions. Their activity functions as interpretive infrastructure: a sociotechnical layer without which persistent agents cannot be adequately governed over time (Bowker & Star, 1999; Star & Strauss, 1999; Jackson, 2014).

Leaving this reliance informal creates structural vulnerabilities. Epistemic bottlenecks, burnout, capture, opaque influence, and the silent erosion of institutional memory become likely when translational labor is ad hoc and unaccountable, echoing broader concerns in algorithmic accountability and sociotechnical governance research (Benjamin, 2019; Gray & Suri, 2019; Raji et al., 2020). Recognizing translator roles as infrastructural dependencies reframes them as objects of design rather than accidents of adoption (Bowker & Star, 1999; Selbst et al., 2019).

In response, the paper introduced translator trust as an architectural and institutional construct for long-horizon AI systems. We articulated design principles for pluralistic cohorts, scope-limited mandates, escalation pathways, auditability, consent refresh, compensation, and independence protections, and situated these mechanisms within layered agent architectures that include reflective layers, memory governance triggers, evaluation pipelines, and offboarding protocols (Jasanoff, 2003; Mitchell et al., 2019; Raji et al., 2020; NIST, 2023). We further traced implications for research labs, product teams, and regulators, emphasizing that once systems rely on human translators to remain intelligible, the architecture of that reliance becomes a public governance question rather than a local engineering choice (OECD, 2019; European Union, 2024).

The broader lesson is that persistence creates reciprocity. Systems that adapt across time generate organizational dependence on human interpretation. Governing such systems therefore requires not only technical safeguards but institutional arrangements that distribute,

bound, and protect the translational labor on which long-horizon deployment already depends (Suchman, 1987; Bowker & Star, 1999; Selbst et al., 2019).

Persistent AI systems do not merely interact with humans; they rely on certain humans to remain governable. Designing for that reliance—explicitly, transparently, and accountably—is a central task for the next generation of AI governance.

References

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Union. (2024). *Artificial Intelligence Act* (Regulation …/…). *Official Journal of the European Union*.

Executive Office of the President. (2023). *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*.

Fischer, G., et al. (2021). Human–AI collaboration: Challenges and opportunities. *Proceedings of the ACM on Human–Computer Interaction (PACM-HCI).*

Geiger, R. S., Ribes, D., Finn, M., & DiSalvo, C. (2011). The work of sustaining order in Wikipedia. *Proceedings of CSCW.*

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Jackson, S. J. (2014). Rethinking repair. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies* (pp. 221–239). MIT Press.

Jasanoff, S. (2003). Technologies of humility: Citizen participation in governing science. *Minerva, 41*(3), 223–244.

Kleinman, D. L., et al. (2010). *Controversies in science and technology*. Wiley-Blackwell.

Kroll, J. A., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*, 633–705.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation. *Organizational Behavior and Human Decision Processes, 151*, 90–103.

Mitchell, M., et al. (2019). Model cards for model reporting. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT).*

National Institute of Standards and Technology. (2023). *AI risk management framework 1.0*.

OECD. (2019). *OECD principles on artificial intelligence*.

Orlikowski, W. J. (1992). The duality of technology. *Organization Science, 3*(3), 398–427.

Passi, S., & Jackson, S. J. (2017). Data vision: Learning to see through algorithmic systems. *Proceedings of CSCW.*

Raji, I. D., et al. (2020). Closing the AI accountability gap. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT).*

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms. *Proceedings of Data and Discrimination.*

Selbst, A. D., et al. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT).*

Shneiderman, B. (2020). Human-centered AI. *International Journal of Human–Computer Interaction, 36*(6), 495–504.

Silverstone, R., & Haddon, L. (1996). Design and domestication of ICTs. In R. Silverstone & R. Mansell (Eds.), *Communication by design*. Oxford University Press.

Star, S. L. (2010). This is not a boundary object. *Science, Technology, & Human Values, 35*(5), 601–617.

Star, S. L., & Strauss, A. (1999). Layers of silence, arenas of voice. *Computer Supported Cooperative Work, 8*(1–2), 9–30.

Suchman, L. (1987). *Plans and situated actions*. Cambridge University Press.

Turkle, S. (2011). *Alone together*. Basic Books.