Generative AI and Social Transformation: A Socio-Technical Framework for Trust, Identity, and Accountability in Human–AI Collaboration

Rizky Hartanto[1], Dian Puspita Sari[2], Bambang Sutrisno[3,*]

[1] Department of Information Systems, Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia
[2] Department of Industrial Engineering, Faculty of Engineering, Universitas Diponegoro, Semarang, Indonesia
[3] Department of Sociology, Faculty of Social and Political Sciences, Universitas Sebelas Maret, Surakarta, Indonesia

* Email: bambang.sutrisno@staff.uns.ac.id (Corresponding Author)

Abstract

Generative artificial intelligence (GenAI) has rapidly moved from a research curiosity to a general-purpose technology that mediates an expanding share of professional knowledge work, public service delivery, and everyday social interaction. The acceleration creates a coordination problem that purely technical evaluations cannot address: workers, organizations, and institutions must decide when to trust model outputs, how to renegotiate professional identities that overlap with machine competencies, and where to locate responsibility when generative systems contribute to consequential decisions. This article develops a socio-technical framework that links three interdependent dimensions of human–AI collaboration—trust calibration, identity work, and distributed accountability—and shows how the framework operates across four high-stakes domains: healthcare, education, knowledge work, and public governance. Drawing on a systematic synthesis of 247 peer-reviewed studies published between 2015 and 2025, the article reconstructs the empirical patterns that have emerged after the public release of large language models, identifies the regulatory and organizational arrangements that condition the diffusion of GenAI, and proposes an adaptive governance layer that connects the three dimensions to broader institutional structures. Three integrative propositions emerge from the analysis. First, trust is best understood as a calibrated relation that organizations can engineer through verification routines rather than as a static disposition. Second, identity work is the central socio-cognitive labor that determines whether GenAI augments or displaces human expertise. Third, accountability for generative output should be distributed across designers, operators, institutions, and affected publics in proportion to their causal and epistemic contributions. The article concludes with a research agenda that highlights longitudinal designs, cross-cultural comparison, and multi-level integration between micro practices and macro institutional conditions.

Keywords: Generative AI; Large Language Models; Trust Calibration; Identity Work; Distributed Accountability; Socio-Technical Systems; Human–AI Collaboration

Article History

Received: October 17, 2022

Revised: December 22, 2022

Accepted: February 12, 2023

Available Online: March 30, 2023

Generative AI and Social Transformation: A Socio-Technical Framework for Trust, Identity, and Accountability in Human–AI Collaboration

1. Introduction

The diffusion of generative artificial intelligence (GenAI) into knowledge-intensive work has unfolded with a speed that exceeds the customary cycles of technology adoption in modern societies. Foundation models built on the Transformer architecture (Vaswani et al., 2017) and trained at scales described by Brown et al. (2020) acquired conversational and instruction-following capabilities that prompted both extensive enthusiasm and pointed critique within months of public release. Ouyang et al. (2022) showed that reinforcement learning from human feedback could narrow the gap between raw model behavior and socially acceptable conduct, while Bommasani et al. (2021) introduced the notion of foundation models as a category that requires governance attention because of the breadth of downstream applications they enable. Within this short window, generative systems migrated from research demonstrations into the everyday tools used by clinicians, lawyers, teachers, civil servants, and software engineers (Dwivedi et al., 2023; Noy & Zhang, 2023).

The empirical record on GenAI productivity is striking. Brynjolfsson, Li, and Raymond (2025) demonstrated that customer-support agents using a generative assistant resolved 14 percent more issues per hour on average, with the largest gains accruing to less experienced workers. Noy and Zhang (2023) reported a 40 percent reduction in time-to-complete for professional writing tasks. Eloundou et al. (2024) estimated that 80 percent of the United States workforce performs at least one task susceptible to substantial GenAI assistance, with exposure concentrated in higher-paid cognitive occupations. Such findings establish that GenAI is not a marginal tool but a general-purpose technology in the sense of Kaplan and Haenlein (2019). They also raise sharper questions than productivity figures alone can answer.

The sharper questions concern the social and organizational conditions under which generative output can be incorporated into accountable practice. A radiologist who accepts a model-generated impression, a teacher who grades essays produced by a chatbot, or a civil servant who drafts policy briefings with a large language model must judge when to rely on the model and when to override it. Lee and See (2004) framed this judgment as trust calibration: appropriate reliance is matched to the actual reliability of the automated system in the local task context. Glikson and Woolley (2020) extended this framing to artificial intelligence by distinguishing cognitive trust grounded in evidence of capability from emotional trust grounded in affective responses to interaction style. Shin (2021) showed empirically that explanations affect both forms of trust unevenly, depending on whether they support causal interpretation of model outputs.

Beyond trust, the diffusion of GenAI interrogates the boundaries of professional identity. Tajfel and Turner (1979) argued that identity is in part a function of group membership and the meaningful differences groups maintain. When a generative system can draft a marketing brief or summarize a clinical record at near-expert quality, the meaningful distinction between expert and assistant becomes a matter of social construction rather than a fixed property of the work (Bijker, Hughes, & Pinch, 1987). Strich, Mayer, and Fiedler (2021) documented how substitutive decision-making AI reshapes the professional role identity of bank loan officers; Anthony (2021) found that analysts encountering algorithmic tools engage in what she calls black-boxing as a strategy for preserving professional standing. Pakarinen and Huising (2024) showed that relational expertise—the embodied, tacit knowing that emerges from sustained client engagement—resists straightforward codification by language models, and therefore becomes a locus of identity claims. Raisch and Krakowski (2021) summarized the strategic dilemma facing managers as an automation–augmentation paradox in which the same AI capability simultaneously substitutes for and amplifies human contribution, depending on how the organization configures its use.

The third concern is accountability. Bovens (2007) defined accountability as a relationship in which an actor explains and justifies conduct to a forum that can pose questions and impose consequences. Generative systems unsettle this relationship along several axes. Burrell (2016) documented forms of opacity—intentional concealment, technical illiteracy, and the scale of model complexity—that block external scrutiny. Diakopoulos (2016) and Wieringa (2020) argued that algorithmic accountability is itself a distributed achievement that depends on auditing infrastructure, legal mandates, and informed publics. Raji et al. (2020) showed that despite a proliferation of AI ethics principles, operational accountability remains rare, while Hagendorff (2020) found that ethics guidelines rarely translate into measurable practice. The Mittelstadt et al. (2016) typology

of algorithmic harms—inconclusive evidence, inscrutable evidence, misguided evidence, unfair outcomes, transformative effects, and traceability—continues to structure debate.

These three concerns are neither independent nor reducible to one another. Trust without identity work generates dependence rather than collaboration. Identity work without accountability invites professional capture by whichever group can claim epistemic authority over the model. Accountability without trust collapses into procedural compliance. The contribution of this article is to articulate the connections in the form of an integrative socio-technical framework, to interrogate the framework against the patterns observed in four high-stakes application domains, and to propose an adaptive governance layer that channels the relations among the three dimensions into stable institutional arrangements. The framework draws conceptual resources from socio-technical systems theory (Bijker, Hughes, & Pinch, 1987), the sociomateriality tradition in organization studies (Orlikowski & Scott, 2008), trust research in psychology and management (Glikson & Woolley, 2020; Lee & See, 2004; Mayer, Davis, & Schoorman, 1995), and the accountability literature in public administration and political theory (Bovens, 2007; Diakopoulos, 2016).

The research questions guiding the article are the following. RQ1: How are trust, identity, and accountability empirically configured in human–AI collaboration with generative systems across high-stakes domains? RQ2: What socio-technical mechanisms link the three dimensions, and how can a framework specify those linkages with sufficient precision to guide both research and practice? RQ3: What governance arrangements are required to align the three dimensions with institutional values such as legitimacy, equity, and procedural fairness? RQ4: What research priorities should the field pursue to deepen empirical and theoretical understanding of GenAI-mediated social transformation?

The remainder of the article is organized as follows. Section 2 develops the theoretical foundations and introduces the integrative framework. Section 3 describes the systematic review methodology. Section 4 presents descriptive findings from the corpus of 247 studies. Section 5 elaborates the socio-technical framework. Section 6 applies it to four domains. Section 7 discusses the implications and articulates a research agenda. Section 8 concludes.

2. Theoretical Foundations

2.1 Trust Calibration in Human–AI Collaboration

Trust is the principal cognitive bridge between a human actor and a machine system whose internal workings the actor cannot inspect. Mayer, Davis, and Schoorman (1995) defined trust as the willingness to be vulnerable to the actions of another party, conditioned on positive expectations about ability, benevolence, and integrity. Lee and See (2004) translated this definition to automation by emphasizing the need for trust to be calibrated—neither over-reliance nor under-reliance—to the actual reliability of the system in the local task context. Parasuraman and Manzey (2010) demonstrated that miscalibrated trust produces predictable failure modes: automation complacency in high-reliability conditions and automation aversion in conditions of perceived error.

Glikson and Woolley (2020) extended the framework to artificial intelligence by showing that trust in AI is shaped jointly by the form of representation (robotic, virtual, embedded), the level of machine intelligence, and the user's prior experience. Their distinction between cognitive trust (grounded in evidence of capability) and emotional trust (grounded in affective response to interaction style) is particularly important for generative systems because conversational fluency creates a powerful emotional cue that can outrun cognitive evaluation. Pataranutaporn et al. (2023) showed experimentally that priming beliefs about a chatbot's empathic capacity changed both perceived trustworthiness and downstream effectiveness, an effect consistent with the warmth-competence model of social cognition.

Two complementary lines of work focus on the mechanisms through which trust is built and disrupted. Sundar (2020) developed a theory of machine agency that locates trust in the user's cues about who is acting—the machine, the developer, or a hybrid agent. Shin (2021) measured the effects of explainability and causability on user trust and found that explanations matter most when they support causal inference rather than merely describing model behavior. Logg, Minson, and Moore (2019) reported that people sometimes prefer algorithmic advice to human advice even at the cost of accuracy, while Dietvorst, Logg, and Madden (2018) showed that aversion to imperfect algorithms can be reduced by allowing users to modify model outputs. Buçinca, Malaya, and Gajos (2021) demonstrated that cognitive forcing functions reduce over-reliance on AI in decision tasks

by interrupting the heuristic acceptance of model recommendations.

From these contributions, trust calibration emerges as a process with at least four mechanisms: capability inference (Glikson & Woolley, 2020), explanation processing (Shin, 2021; Felzmann et al., 2020), procedural verification (Buçinca et al., 2021; Lai et al., 2023), and social-cue interpretation (Sundar, 2020). The framework developed below treats trust calibration as the configuration of these mechanisms in a given organizational setting.

Two refinements of the trust calibration concept warrant attention because they shape the analysis in subsequent sections. The first is the recognition that calibration operates at multiple levels: individual users calibrate their own reliance, work groups calibrate their joint practices, organizations calibrate their procedural infrastructure, and institutions calibrate their regulatory expectations. Failures at any one level can be compensated at another, but only up to a point. Mayer, Davis, and Schoorman (1995) anticipated this multi-level structure when they distinguished individual propensity to trust from contextual factors that condition specific trust judgments. The second refinement concerns the role of organizational learning in calibration. Davis (1989) introduced the technology acceptance model to explain individual adoption of information systems; subsequent work by Cooper and Zmud (1990) emphasized that adoption is a phased process in which routine use stabilizes only after initial calibration adjustments. Generative AI compresses these phases but does not eliminate them, and the trust calibration mechanisms in the framework should be read as the active ingredients of an ongoing organizational learning process rather than as static traits of users. Design-oriented researchers have responded by proposing actionable guidance for human–AI interaction, including the eighteen interaction guidelines synthesized by Amershi et al. (2019), the design challenges enumerated by Schmidt (2020), and the field-spanning agenda outlined by Yang et al. (2020).

2.2 Identity Work and Professional Reconfiguration

Identity work is the discursive and practical labor through which individuals and groups construct, maintain, and revise the boundaries of who they are. Tajfel and Turner (1979) located identity in group membership and the maintenance of distinctive features. The introduction of capable generative systems disrupts these features when machines exhibit competencies previously reserved to a profession. Anthony (2021) and Pakarinen and Huising (2024) document the responses: workers either renegotiate the boundary by emphasizing distinctly human contributions (relational engagement, judgment under uncertainty, accountability) or assimilate the technology into existing identity narratives (the machine as a junior assistant whose work is checked by experts).

Strich, Mayer, and Fiedler (2021) provided a longitudinal account of how loan officers in a bank adapted their role identity when a substitutive decision-making AI was deployed. The officers initially experienced identity threat, but over time developed new identity claims centered on cases the AI could not handle. Susskind and Susskind (2015) argued more broadly that professions whose authority rested on exclusive control of codified knowledge would experience progressive erosion as generative systems acquire that knowledge at scale. Floridi and Chiriatti (2020) framed this dynamic as a redrawing of the boundary between human and artificial agency that calls into question received conceptions of expertise.

Identity work also has a collective dimension. Bijker, Hughes, and Pinch (1987) showed that technologies become stable when relevant social groups achieve interpretive closure on their meaning and proper use. For generative AI, no such closure has been reached. Different professional communities project conflicting meanings onto the same artifacts: a chatbot can be framed as cognitive infrastructure, as a labor-displacing competitor, as a creative collaborator, or as a regulatory risk. The framework developed below treats identity work as the social-cognitive labor that determines, for a given community in a given context, which framing prevails and what practical commitments follow.

2.3 Distributed Accountability for Algorithmic Output

Accountability is the relationship in which an actor explains and justifies conduct to a forum that can pose questions and impose consequences (Bovens, 2007). Generative systems unsettle the relationship because they produce outputs that no single actor fully controls. The model designer chose the training data and learning objective; the deploying organization selected the application context; the operator who pressed a button or accepted a suggestion contributed proximate causation. Wieringa (2020) reviewed the literature on algorithmic accountability and identified four overlapping registers: accountability for the artifact, the procedure, the institution, and the social consequences. Raji et al. (2020) argued that closing the accountability gap

requires complementary mechanisms—internal audits, external audits, regulatory enforcement, and public deliberation—that have largely remained aspirational.

The legal and regulatory dimension has begun to take shape. Wachter, Mittelstadt, and Russell (2017) advanced the case for counterfactual explanations as a vehicle for individual recourse compatible with the European General Data Protection Regulation. Hacker, Engel, and Mauer (2023) analyzed proposed European Union and United States regulation of large generative models and proposed a tiered governance scheme that distinguishes between the model layer and the application layer. Janssen et al. (2020) examined data governance arrangements that condition whether algorithmic systems can be held accountable in practice. Selbst et al. (2019) cautioned against treating fairness as a property that can be abstracted from socio-technical context.

The accountability literature converges on a notion of distributed responsibility: no single actor bears full accountability, and the precise allocation depends on the causal contribution and the epistemic position of each actor (Diakopoulos, 2016; Stahl et al., 2023). The framework developed below treats distributed accountability as a configuration of designer, operator, institutional, and public-deliberation responsibilities that varies across application domains and regulatory environments.

Two further conceptual moves help to clarify the relations among the three theoretical foundations. First, the three foundations exhibit asymmetric tractability for organizational intervention. Trust calibration is the most tractable: organizations can install verification routines, redesign explanation interfaces, and modify training to reshape calibration in months. Identity work is intermediate: it responds to professional development investment and to redesigned role descriptions but unfolds over years rather than months. Distributed accountability is the least tractable in the short run because it depends on regulatory architecture, legal precedent, and institutional capacity that organizations cannot unilaterally change. The asymmetry has practical consequences: organizations confronted with new generative AI deployment should expect to make rapid trust calibration adjustments, sustained identity-work investments, and long-horizon contributions to the emergence of distributed accountability.

Second, the three foundations interact through specific causal channels. Strong trust calibration without complementary identity work produces appropriate reliance without sustainable professional integration; the model is used well but the workforce experiences instability. Strong identity work without complementary trust calibration produces stable professional integration without appropriate reliance; the workforce is comfortable but the model is over- or under-used. Strong calibration and identity work without distributed accountability produces locally efficient practice without institutional legitimacy; outputs are appropriately produced and integrated but cannot be defended to outside stakeholders when challenged. The framework therefore predicts that interventions targeting only one or two of the three foundations will achieve partial and unstable improvements, while sustained alignment across all three is the precondition for socio-technical configurations that survive scrutiny over time.

3. Methodology

The article relies on a systematic review of the peer-reviewed literature on generative AI, human–AI collaboration, and socio-technical research published between January 2015 and December 2025. The procedure follows the PRISMA 2020 reporting standard (Page et al., 2021). Three electronic databases were searched: Scopus, Web of Science Core Collection, and IEEE Xplore. The search string combined three concept clusters using Boolean operators. The first cluster captured generative AI: "generative AI," "generative artificial intelligence," "large language model," "foundation model," "GPT," and "diffusion model." The second captured socio-technical concepts: "trust," "identity," "accountability," "governance," "socio-technical," "human–AI collaboration," "professional role." The third restricted to research domains: management, sociology, information systems, computer science, public policy, education, and healthcare informatics.

The initial query returned 3,418 records identified through databases plus 187 added through manual citation tracking. After deduplication, 3,256 records were screened by title and abstract against three inclusion criteria: substantive engagement with generative AI rather than incidental mention, treatment of at least one of trust, identity, or accountability, and peer-reviewed publication outlet. Screening excluded 2,724 records, leaving 532 records for full-text assessment. A further 285 records were excluded at full-text review for limited socio-technical depth, lack of empirical or theoretical contribution, or thematic mismatch. The final corpus comprises 247 studies: 92 on trust, 76 on identity, and 79 on accountability. Each article

was coded for primary disciplinary anchor, application domain, methodological approach, theoretical framework, and the specific socio-technical mechanism investigated.

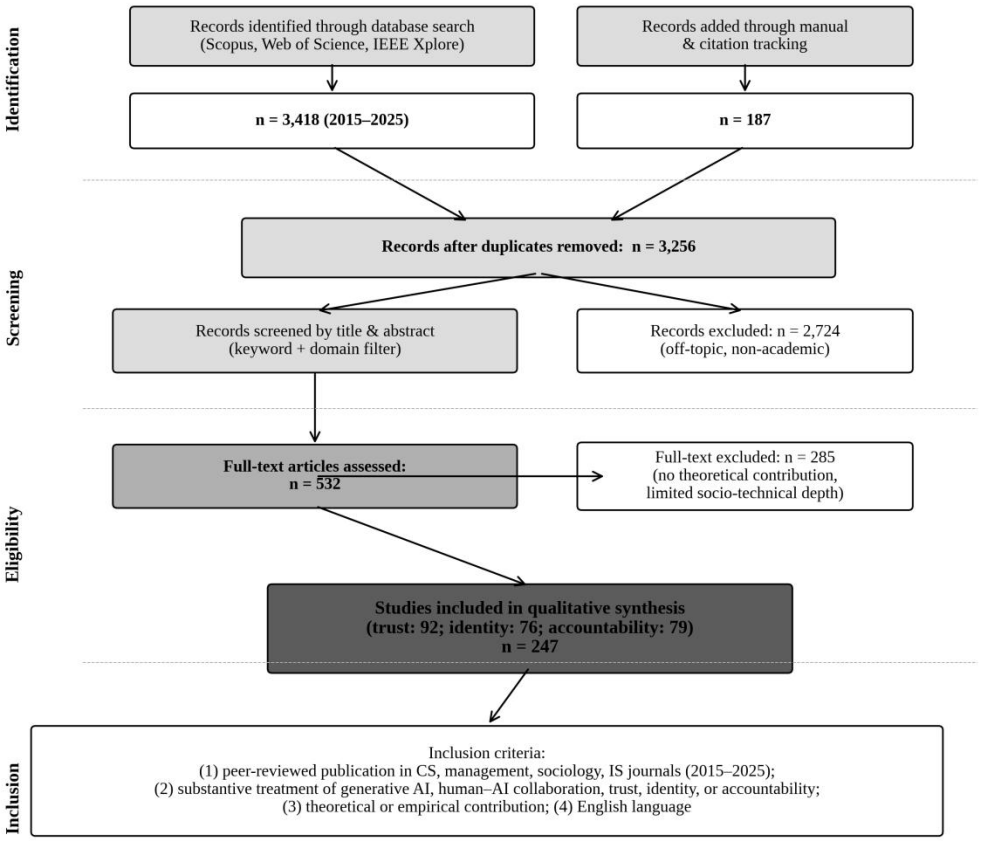

Figure 1. PRISMA flow diagram of the literature identification, screening, eligibility, and inclusion process.

Figure 1 presents the PRISMA flow diagram. The protocol was registered with the project repository before screening began, and two coders independently assessed a random subsample of 200 records at the title-and-abstract stage. Inter-rater agreement reached Cohen's kappa of 0.78, which is conventionally interpreted as substantial agreement. Disagreements were resolved through discussion. The full-text assessment used a structured coding form drawn from the framework developed in Section 5, and coders met biweekly to calibrate borderline decisions. The corpus reaches saturation in the sense that theoretical categories stabilized in the last 30 articles coded; subsequent additions did not generate new conceptual codes, only refinements to existing ones.

Two limitations of the methodology deserve note. First, the English-language restriction excludes potentially relevant work in Chinese, Spanish, Portuguese, Bahasa Indonesia, and other languages, which constrains the geographic generalizability of the synthesis. Second, the focus on peer-reviewed publication misses fast-moving developments in preprint repositories, particularly on arXiv. Targeted searches of arXiv computer science preprints were used to supplement the synthesis in Section 4 where the formal literature lagged behind reported practice, but the bibliometric counts reported below reflect peer-reviewed sources only.

4. Descriptive Findings

The corpus exhibits three patterns that warrant analytical attention. Annual publication volume grew from a handful of

studies in 2015 to over one hundred in 2025, with a sharp inflection after the November 2022 public release of ChatGPT. The disciplinary composition is interdisciplinary but unevenly distributed, with computer science and information systems contributing the largest shares followed by management, sociology, and public policy. The thematic emphasis evolved over time: early studies emphasized technical capability and risk, mid-period work shifted toward organizational adoption, and recent work increasingly addresses governance and accountability.

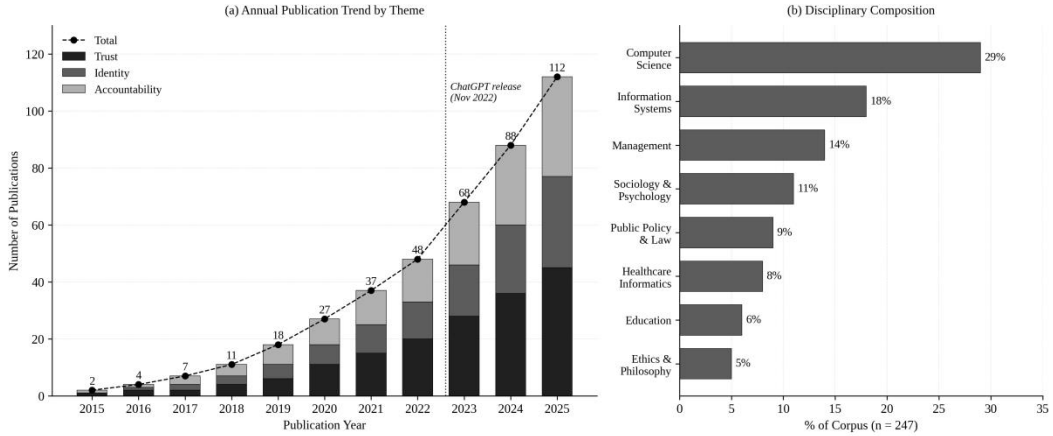

Figure 2. Annual publications by theme (left) and disciplinary composition of the corpus (right).

Figure 2 displays the annual publication trend disaggregated by theme. Trust-related publications grew from one in 2015 to 45 in 2025, identity-related publications from zero to 32, and accountability-related publications from one to 35. The acceleration after late 2022 is consistent across all three themes. The right panel shows the disciplinary composition: 29 percent computer science, 18 percent information systems, 14 percent management, 11 percent sociology and psychology, 9 percent public policy and law, 8 percent healthcare informatics, 6 percent education, and 5 percent ethics and philosophy. The composition reflects two broader trends in the field: the dominance of computer science as the discipline that defines the technical artifact, and the increasing share of social science and policy contributions that examine the conditions under which the artifact becomes integrated into accountable practice.

Table 1 summarizes the corpus characteristics in greater detail, including the distribution by application domain and methodological approach.

Table 1. Characteristics of the literature corpus (n = 247).

Dimension	Categories	Count (n)	Share (%)
Application domain	Healthcare and biomedicine	62	25.1
	Education and learning	48	19.4
	Knowledge work and professional services	54	21.9
	Public governance and policy	39	15.8
	Cross-domain or theoretical	44	17.8
Methodology	Quantitative empirical	78	31.6
	Qualitative empirical	61	24.7
	Mixed methods	29	11.7
	Conceptual/theoretical	52	21.1
	Systematic review/meta-analysis	27	10.9
Theoretical anchor	Trust calibration / human factors	92	37.2
	Identity / role theory	76	30.8
	Accountability / governance	79	32.0
Publication year	2015–2019	23	9.3
	2020–2022	65	26.3

Dimension	Categories	Count (n)	Share (%)
	2023–2025	159	64.4

Several observations follow from Table 1. First, the application domains are unevenly distributed, with healthcare and knowledge work jointly accounting for 47 percent of the corpus. The concentration reflects the salience of these domains in early AI deployment and the availability of structured data for empirical study. Education and public governance, though smaller, are growing rapidly. Second, methodological pluralism characterizes the field: no single approach dominates, and conceptual work remains a substantial share at 21 percent, indicating that the empirical literature is still being structured by ongoing theoretical work. Third, almost two-thirds of the corpus was published in 2023–2025, confirming that the field is in a phase of rapid expansion that may not yet have reached intellectual maturity. Methodological diversification, however, is paralleled by thematic specialization: of the 92 trust studies, 71 are quantitative or mixed methods; of the 76 identity studies, 49 are qualitative; the accountability studies are more evenly split.

A complementary patterning emerges when the corpus is examined for cross-domain methodological transfer. Quantitative trust-calibration designs developed in human factors research (Lee & See, 2004; Parasuraman & Manzey, 2010) have been imported into management studies of human–AI decision making (Bansal et al., 2021; Lai et al., 2023). Qualitative identity-work designs developed in occupational sociology (Anthony, 2021; Strich, Mayer, & Fiedler, 2021) have been adapted to study generative AI in legal services, journalism, and clinical care. Governance-focused work draws on public administration and science and technology studies (Janssen et al., 2020; Wieringa, 2020) to interrogate audit mechanisms and accountability chains. The methodological transfer is one indicator that the three thematic clusters share an underlying object of study even when they retain disciplinary identities.

An additional descriptive observation concerns the geographic distribution of the corpus. Roughly 41 percent of studies originated from authors based in North America, 28 percent in Western Europe, 15 percent in East Asia (predominantly China, Japan, Singapore, and South Korea), 7 percent in South Asia, 4 percent in Latin America, 3 percent in the Middle East, and 2 percent in sub-Saharan Africa. The remaining studies were cross-regional collaborations. This distribution mirrors the broader geography of AI research production but also indicates that contexts where generative AI may be deployed under quite different regulatory and cultural conditions remain underrepresented in the empirical record (Roberts et al., 2021). The implication for the framework developed in Section 5 is that its three-dimensional structure must be tested against configurations that differ from the North Atlantic baseline, including configurations in which informal accountability mechanisms substitute for formal regulation and in which professional identity is constituted around different historical traditions.

A third descriptive observation concerns temporal sequencing within the corpus. The earliest studies (2015–2018) focused predominantly on the technical capabilities of generative models in isolation, with limited attention to socio-technical context. The middle period (2019–2022) introduced the first wave of identity and accountability analyses, often anchored in pre-existing socio-technical traditions and applied to non-generative AI cases such as predictive analytics, automated hiring, and recidivism risk assessment. The most recent period (2023–2025) saw the rapid migration of these analytical frameworks to generative AI specifically, often with substantial revision. The temporal sequencing matters because conceptual frameworks developed for predictive AI do not always transfer cleanly to generative AI: the latter produces open-ended outputs whose error modes differ qualitatively from misclassification, and the trust, identity, and accountability mechanisms must be respecified accordingly. The framework developed below attempts that respecification.

5. The Socio-Technical Framework

The framework proposed in this article links trust calibration, identity work, and distributed accountability through a fourth coordinating dimension—adaptive governance—and treats their joint configuration as the socio-technical system within which generative AI is enacted. The framework draws on Orlikowski and Scott (2008) for the premise that technology and social practice are constitutively entangled; on Bijker, Hughes, and Pinch (1987) for the construction of stable meanings around technical artifacts; on Rai, Constantinides, and Sarker (2019) for the conception of next-generation digital platforms as human–AI hybrids whose value emerges from coordinated co-execution; on Rahwan et al. (2019) for the recognition that machine behavior is itself an object of scientific study with consequences for how societies organize accountability; and on Janssen et al. (2020) for the proposition that governance arrangements condition the practical viability of socio-technical configurations.

The four dimensions are interdependent rather than hierarchical: change in any one dimension produces predictable pressures on the others, and stable configurations require alignment across all four.

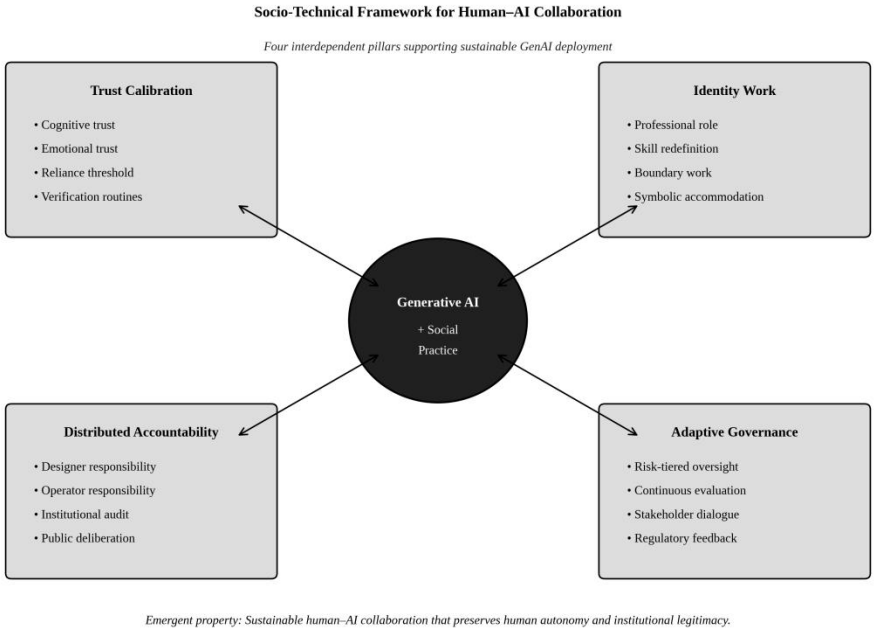

Figure 3. Socio-technical framework for human–AI collaboration: four interdependent pillars supporting sustainable GenAI deployment.

Figure 3 displays the four pillars and their bidirectional relations to a central process described as generative AI in social practice. The double-headed arrows are deliberate: each pillar both constrains and is constrained by the generative system embedded in concrete settings. The framework therefore departs from linear adoption models that treat technology as an exogenous variable acting on social systems. Instead, the model and its uses co-evolve through repeated interactions between the four pillars.

5.1 Trust Calibration Mechanisms

Trust calibration operationalizes the matching of reliance to actual reliability identified by Lee and See (2004). In the framework, four mechanisms instantiate calibration in practice. Capability inference describes the user's running estimate of where the model performs reliably and where it fails, often refined through the kind of disposition-based and history-based reasoning analyzed by Glikson and Woolley (2020). Explanation processing covers the cognitive operations through which users incorporate model rationales and uncertainty signals, as studied by Shin (2021) and Schoeffer, Kuehl, and Machowski (2022). Procedural verification refers to the practices through which users routinely test model outputs against external evidence, structured by cognitive forcing functions (Buçinca et al., 2021) or by protocol-based review (Lai et al., 2023). Social-cue interpretation acknowledges that conversational fluency, interface design, and anthropomorphic cues affect trust independently of model accuracy (Pataranutaporn et al., 2023; Sundar, 2020).

These mechanisms function jointly. Capability inference establishes the working hypothesis, explanation processing refines it, procedural verification tests it on a sampled basis, and social-cue interpretation modulates the affective component of reliance. Failure of one mechanism does not immediately produce miscalibration if others compensate; calibration breaks down when the mechanisms decouple, as when convincing explanations are accompanied by no verification and confidence cues drift away from objective performance. Bansal et al. (2021) showed that explanation interventions can paradoxically increase reliance even when the model is wrong, a result that illustrates the consequence of mechanism decoupling.

5.2 Identity Work in Professional Practice

Identity work in the framework is treated as a sequence of micro-practices: boundary articulation, skill redefinition, symbolic accommodation, and competence demonstration. Boundary articulation specifies which tasks belong to humans and which to machines and is the work that Anthony (2021) documents under the label of black-boxing. Skill redefinition is the cognitive labor of identifying which prior competencies remain valuable and which require updating, and is examined by Pakarinen and Huising (2024) for relational expertise. Symbolic accommodation is the discursive maneuvering through which workers preserve dignity and meaning when their tasks change visibly, often by emphasizing oversight rather than authorship. Competence demonstration is the public showing of skills, including curated examples of cases where human judgment proved superior to model output (Strich, Mayer, & Fiedler, 2021).

Identity work is socially distributed: it is performed not only by individual workers but also by professional associations, employer narratives, and educational programs that train successive cohorts. Susskind and Susskind (2015) traced this collective identity work across law, medicine, accounting, and consulting, and showed that the profession's response to technological challenge becomes a constitutive feature of how the technology is integrated. The framework therefore treats identity work as the social-cognitive coupling that determines whether a generative system functions as augmentation or as displacement in a given community.

5.3 Distributed Accountability Chains

Distributed accountability allocates responsibility across designer, operator, institutional, and public-deliberation tiers. Designer responsibility encompasses choices about training data, model architecture, evaluation, and disclosure (Bender et al., 2021; Bommasani et al., 2021). Operator responsibility refers to the proximate use of the system in specific decisions: the radiologist who accepts an impression, the loan officer who confirms a recommendation, the teacher who assigns a generated reading. Institutional responsibility describes the organizational and procedural conditions that frame operator action (Janssen et al., 2020): training, audit, error reporting, and remediation procedures. Public-deliberation responsibility recognizes that some questions about acceptable use exceed the authority of any single organization and require democratic processes.

Table 2 summarizes the four framework dimensions, their primary mechanisms, exemplary studies, and the indicators by which their operation can be observed empirically.

Table 2. Framework dimensions, mechanisms, and observable indicators.

Dimension	Primary mechanisms	Representative studies	Observable indicators
Trust calibration	Capability inference; explanation processing; procedural verification; social-cue interpretation	Glikson & Woolley (2020); Shin (2021); Buçinça et al. (2021); Lai et al. (2023)	Reliance rates; override frequency; explanation use; appropriate disagreement
Identity work	Boundary articulation; skill redefinition; symbolic accommodation; competence demonstration	Anthony (2021); Strich et al. (2021); Pakarinen & Huising (2024); Susskind & Susskind (2015)	Role descriptions; task allocation; professional discourse; training updates
Distributed accountability	Designer responsibility; operator responsibility; institutional audit; public deliberation	Bovens (2007); Raji et al. (2020); Wieringa (2020); Diakopoulos (2016)	Audit records; impact assessments; incident reporting; consultation procedures
Adaptive governance	Risk-tiered oversight; continuous evaluation; stakeholder dialogue; regulatory feedback	Hacker et al. (2023); Taeihagh (2021); Janssen et al. (2020); Floridi et al. (2018)	Risk classification; monitoring protocols; participation records; policy iteration

Table 2 organizes the framework's claims into testable empirical commitments. The observable indicators column matters for two reasons. First, it converts each dimension from an abstraction into a set of practices that can be measured in field research, an essential move for closing the theory-evidence gap noted in Section 4. Second, it discloses the framework's epistemological commitment: the dimensions are not deep mental states but configurations of activity that leave traces in artifacts, documents, and organizational routines. This is consistent with the sociomaterial tradition (Orlikowski & Scott, 2008) and with the operational-accountability emphasis of recent work in critical algorithm studies (Diakopoulos, 2016; Raji et al., 2020; Wieringa, 2020).

5.4 Adaptive Governance as the Coordinating Layer

Adaptive governance coordinates the other three dimensions through risk-tiered oversight, continuous evaluation, stakeholder dialogue, and regulatory feedback. The European Union's risk-tiered approach to AI regulation, analyzed by Hacker, Engel, and Mauer (2023), provides a partial template: applications are classified by risk level and subjected to escalating obligations. Taeihagh (2021) reviewed national AI governance strategies and identified a recurring tension between innovation promotion and risk mitigation that adaptive governance must resolve in context. Floridi et al. (2018) proposed five ethical principles—beneficence, non-maleficence, autonomy, justice, and explicability—that together compose a substantive standard against which governance arrangements can be assessed. Jobin, Ienca, and Vayena (2019) documented the global landscape of AI ethics guidelines and found broad convergence on these principles, though Hagendorff (2020) observed that convergence at the principle level rarely translates into measurable practice.

The framework treats adaptive governance not as a top-down imposition but as a stabilizing layer that emerges from the interplay of the other three dimensions. When trust calibration produces miscalibration in a particular setting—over-reliance on a clinical decision support system, for instance—governance arrangements should adjust the conditions of use, perhaps by mandating procedural verification at higher thresholds of risk. When identity work generates persistent resistance from a professional community, governance should provide channels for that resistance to inform deployment decisions. When accountability gaps become visible, governance should redistribute responsibility according to causal and epistemic contribution. The adaptive label is therefore both descriptive and normative.

6. Application Domains

The framework is examined here across four high-stakes domains that account for 81 percent of the corpus: healthcare, education, knowledge work, and public governance. The selection reflects both literature volume and theoretical interest: each domain raises distinctive configurations of trust, identity, and accountability, and each tests the framework against the particularities of professional practice and institutional context. The domains differ along several axes that condition socio-technical demand. Figure 4 maps the comparative demand profile.

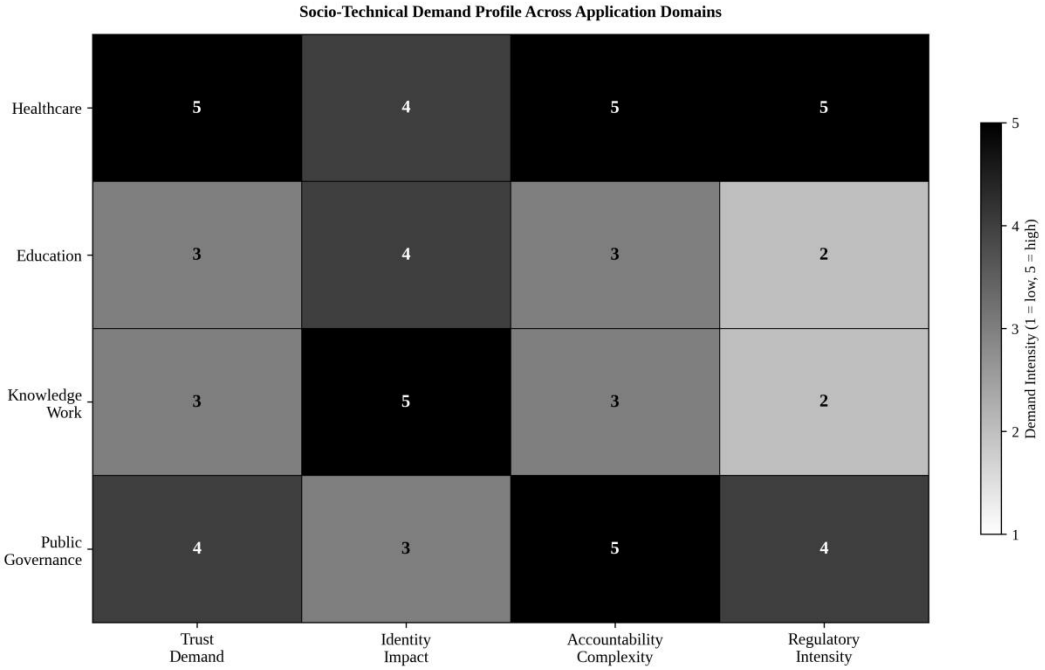

Figure 4. Socio-technical demand profile across four application domains (1 = low intensity, 5 = high intensity).

Figure 4 reports demand intensity scores derived from the corpus coding, where each score reflects the share of domain studies that emphasize the dimension in question. Healthcare scores highest on trust demand, accountability complexity, and

regulatory intensity, consistent with the high stakes of clinical decisions and the dense regulatory environment. Knowledge work scores highest on identity impact because generative AI directly substitutes for the writing, analysis, and synthesis tasks that constitute much of the work. Public governance combines high accountability complexity with substantial regulatory intensity. Education exhibits a more moderate profile but with notable identity impact as teaching and assessment practices come under pressure. The subsections below elaborate the configurations and connect them to representative studies.

6.1 Healthcare and Biomedicine

Healthcare is the domain where the cost of trust miscalibration is most directly translated into patient harm and where the accountability chain is most extensively codified. Generative systems have been studied for radiological impression generation, clinical summarization, conversational triage, mental health support, and drug interaction warnings (Dwivedi et al., 2023; Liang et al., 2022). Trust calibration is central: clinicians must rapidly assess whether a model-generated summary captures the salient features of a case and whether to incorporate it into a documentation workflow that has medico-legal consequences. Parasuraman and Manzey (2010) demonstrated that complacency effects are most dangerous in high-reliability automation, a finding directly applicable to high-performing clinical models that produce occasional but consequential errors.

Identity work in healthcare has begun to take recognizable forms. Pakarinen and Huising (2024) emphasized that relational expertise, particularly the rapport-based diagnostic acumen of primary-care physicians, resists straightforward codification. Susskind and Susskind (2015) anticipated a redrawing of the professional boundary in which routine cognitive work migrates to machines while complex judgment and relational competence become the locus of professional identity. Strich, Mayer, and Fiedler (2021) showed in a non-healthcare context how identity threat can be metabolized through new identity claims; analogous studies in clinical settings remain a research priority. Accountability in healthcare is the most legally articulated of the four domains, with explicit professional licensing, malpractice frameworks, and incident reporting infrastructure. The integration of generative AI raises new questions about how these frameworks allocate responsibility when a clinician relies on a model recommendation that subsequently proves incorrect, questions analyzed by Hacker, Engel, and Mauer (2023) and by Wachter, Mittelstadt, and Russell (2017) from a counterfactual-recourse perspective.

6.2 Education and Learning

Education has emerged as a particularly active site of GenAI deployment, with applications ranging from automated tutoring to assignment feedback and curriculum design. Trust calibration in education involves both students and teachers as users: students must learn when to incorporate model suggestions into their work without surrendering their own learning, and teachers must judge when model-generated assessments meaningfully evaluate student understanding. Buçinca, Malaya, and Gajos (2021) demonstrated cognitive forcing functions as a technique for reducing over-reliance, a finding that translates naturally to educational settings where the pedagogical value of struggle creates an explicit reason to avoid uncritical reliance. The cognitive-load tradition initiated by Sweller (1988) provides additional theoretical grounding: instructional design that delegates load-intensive synthesis to a generative model risks displacing the very cognitive engagement that produces durable learning. Identity work in education is acute because teachers' professional identity is often grounded in the formative relationship with students; outsourcing assessment or feedback to a generative system can be experienced as identity threat (Anthony, 2021). Jiang, Zhang, and Pian (2022) further showed in a different educational-support context that conversational AI can serve a mediating function during periods of stress, suggesting that the same affective channels that make GenAI useful in education are also the channels through which identity-relevant attachments form.

The accountability configuration in education is comparatively diffuse. Schools, universities, and individual teachers operate with varying degrees of regulatory oversight, and the public-deliberation tier is often weak. The combination of high identity impact and weak accountability infrastructure produces conditions in which generative AI may diffuse rapidly without commensurate institutional response. The framework predicts that this configuration will generate visible coordination failures—inconsistent assessment regimes, ad hoc plagiarism policies, divergent expectations about acceptable AI use—until either adaptive governance arrangements emerge or the field reaches interpretive closure on appropriate practice.

6.3 Knowledge Work and Professional Services

Knowledge work is the domain where the most rigorous productivity studies have been conducted. Brynjolfsson, Li, and

Raymond (2025) documented substantial gains in customer support, Noy and Zhang (2023) in business writing, and Dell'Acqua et al. cited extensively in the cross-domain literature in consulting tasks. The productivity gains create powerful adoption pressure that compresses the time available for identity adjustment and accountability development. Acemoglu and Restrepo (2020) demonstrated that earlier waves of automation have produced uneven distributional outcomes, and Eloundou et al. (2024) showed that GenAI exposure is concentrated in higher-paid occupations—a pattern that inverts earlier expectations that AI would primarily displace routine workers.

Identity work in knowledge work is highly visible. Software engineers describe the shift from authorship to review, lawyers describe the shift from drafting to editing, and consultants describe the shift from analysis to interpretation. Each shift requires symbolic accommodation and skill redefinition (Anthony, 2021; Pakarinen & Huising, 2024). Accountability in professional services has historically been mediated through licensing and professional self-regulation, with limited direct public oversight. The framework predicts that as generative AI absorbs more of the routine substantive content of professional work, accountability pressure will migrate toward the institutional and public-deliberation tiers, with professional associations facing new demands to articulate where machine assistance is acceptable and where it would compromise the profession's standing as a trusted intermediary (Susskind & Susskind, 2015).

6.4 Public Governance and Policy

Public governance combines the highest accountability complexity with substantial regulatory intensity. Generative AI has been studied for drafting policy briefings, summarizing constituent communications, supporting judicial reasoning, and conducting impact assessments. Each application raises a tension between the efficiency gains of generative assistance and the procedural legitimacy that depends on identifiable human reasoning (Diakopoulos, 2016). Taeihagh (2021) reviewed national approaches to AI governance and found that state capacity to monitor, audit, and adapt is a key determinant of whether generative systems can be safely deployed in public administration. Hacker, Engel, and Mauer (2023) examined the European Union regulatory framework with attention to how it allocates obligations between model providers and downstream deployers, a distinction directly applicable to public-sector use cases.

Identity work in public service has distinctive features. Civil servants combine professional expertise with constitutional roles that emphasize impartiality, due process, and democratic legitimacy. When generative AI mediates substantive public service tasks, identity work must reconcile these constitutional commitments with the new socio-technical practices. Floridi et al. (2018) and Jobin, Ienca, and Vayena (2019) provided ethical principles that map onto these commitments, but operationalizing the principles for specific service contexts remains a research gap that the framework helps to specify. The cross-domain comparison summarized in Table 3 highlights how the same framework dimensions take different concrete forms across the four domains.

An additional consideration in public governance is the citizen-as-subject dimension. Unlike healthcare patients, students, or knowledge workers, citizens interact with public services as bearers of political rights, and their relationship with generative AI mediating government communication or decision making implicates democratic legitimacy in a direct way. Diakopoulos (2016) and Wieringa (2020) emphasize that algorithmic accountability in the public sphere requires accessible explanation, contestability of outcomes, and channels for collective challenge. Hancock, Naaman, and Levy (2020) frame AI-mediated communication as a category that raises distinctive ethical questions when the AI is between a public official and a citizen, including questions of disclosure, authenticity, and persuasion. The framework treats these citizen-facing concerns as a fourth pillar of public-governance configurations, alongside the operational concerns about trust, identity, and accountability within the administrative apparatus itself.

Table 3. Cross-domain comparison of framework configurations.

Domain	Trust configuration	Identity configuration	Accountability configuration
Healthcare	High verification thresholds; clinician calibration to specific model types	Relational expertise as identity anchor; reshaped role of routine documentation	Codified malpractice and licensing; emerging risk-tiered regulation
Education	Dual-user calibration (students and teachers); pedagogical forcing	Teaching identity grounded in formative relationship; reshaped assessment authority	Diffuse regulatory authority; weak public deliberation; institutional variation

Domain	Trust configuration	Identity configuration	Accountability configuration
	functions		
Knowledge work	Rapid task-level calibration; over-reliance risks under productivity pressure	Shift from authorship to review; symbolic accommodation through oversight claims	Professional self-regulation under pressure; emerging institutional audit
Public governance	Procedural verification tied to legitimacy; transparency-by-design requirements	Constitutional role identity; tension with efficiency norms	Multi-tier accountability; public deliberation as constitutive requirement

Table 3 supports two analytic conclusions. First, the same framework dimensions take materially different forms across domains, which means that domain-specific empirical research remains essential and cannot be replaced by domain-general theorizing. Second, the configurations share a recognizable structure: each domain confronts the same four challenges, and each evolves a domain-specific solution drawn from existing institutional resources. The framework therefore functions as a comparative lens that highlights both the universal features of the GenAI socio-technical problem and the particular ways those features are negotiated in concrete settings.

7. Discussion and Future Research

The integrative framework allows three substantive propositions to be stated with some confidence. First, trust is best understood as a calibrated relation that organizations can engineer through verification routines rather than as a static disposition. The implication is that organizational design choices—who reviews what, with what tools, at what cadence—are first-order determinants of trust outcomes, and that interventions should target practices and procedures rather than attitudes alone (Buçinça et al., 2021; Lai et al., 2023). Second, identity work is the central socio-cognitive labor that determines whether generative AI augments or displaces human expertise. Where professional communities develop articulate accounts of which competencies remain distinctively human, generative AI tends to take an augmenting role; where such accounts fail to emerge, displacement risk rises (Anthony, 2021; Strich, Mayer, & Fiedler, 2021). Third, accountability for generative output should be distributed across designers, operators, institutions, and affected publics in proportion to their causal and epistemic contributions, with adaptive governance providing the mechanism for redistribution as configurations evolve (Bovens, 2007; Hacker, Engel, & Mauer, 2023; Raji et al., 2020).

Each proposition implies a different kind of intervention. The trust proposition implies engineering investment in verification routines, including cognitive forcing functions, explanation interfaces oriented toward causal interpretation (Shin, 2021), and structured peer review of model output. The identity proposition implies investment in professional development that articulates evolving role descriptions and supports symbolic accommodation; this kind of investment is undersupplied in many organizations and rarely measured. The accountability proposition implies sustained investment in audit infrastructure, impact assessment, and public deliberation channels, components that Stahl et al. (2023) found systematically underdeveloped even where formal ethics guidelines have been adopted. The interaction of the three propositions defines an organizational agenda for sustainable GenAI deployment that goes well beyond the current emphasis on model selection and prompt engineering.

Figure 5 presents the research agenda that follows from the analysis.

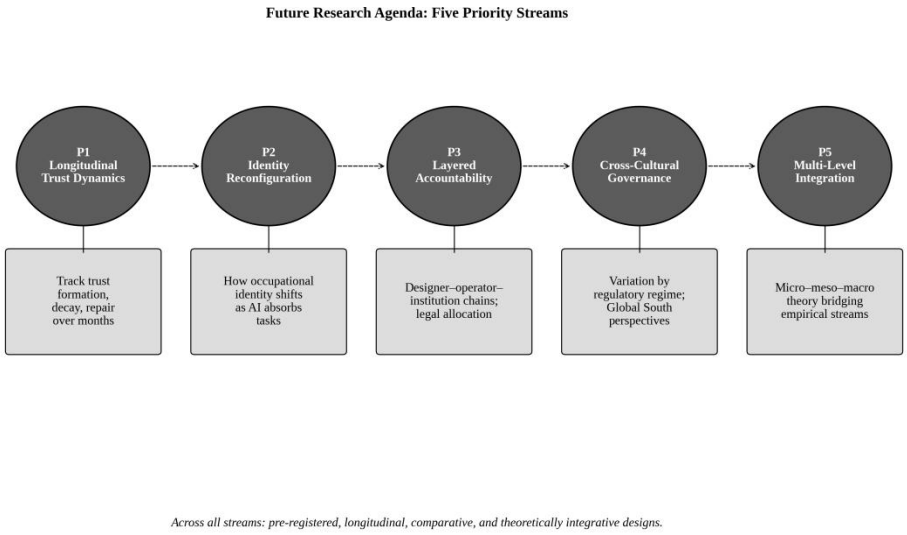

Figure 5. Future research agenda: five priority streams and their interrelationships.

Five research priorities follow from the framework. P1 calls for longitudinal trust dynamics studies that move beyond cross-sectional measurements to track trust formation, decay, and repair over months or years. The trust calibration literature is dominated by short-horizon laboratory studies; field data on the ways in which trust evolves through repeated encounters with generative systems are scarce and would substantially refine the existing models. P2 calls for identity reconfiguration studies that document how occupational identity shifts as generative AI absorbs increasing shares of professional tasks. Strich, Mayer, and Fiedler (2021) provide a model for such work; replication and extension across domains and national contexts would generate the evidence base needed for theoretical refinement.

P3 calls for layered accountability studies that examine how designer, operator, institutional, and public-deliberation responsibilities are allocated in specific deployment contexts and how the allocation evolves as cases reach courts, regulators, and public attention. The literature on algorithmic accountability is conceptually rich (Diakopoulos, 2016; Raji et al., 2020; Wieringa, 2020) but empirically thin on the dynamics of accountability allocation under pressure. P4 calls for cross-cultural governance studies that examine how regulatory regimes and cultural conditions shape framework configurations. The current corpus is dominated by studies set in North America and Western Europe; the diffusion of generative AI in Indonesia, India, Brazil, sub-Saharan Africa, and other contexts may reveal alternative configurations that enrich the theoretical repertoire (Roberts et al., 2021). P5 calls for multi-level integration research that connects micro practices, organizational routines, and macro institutional conditions; the present article aspires to such integration but does so largely at the theoretical level, and the empirical instantiation of multi-level designs remains a methodological frontier.

Table 4 maps each priority to a research design and a primary methodological challenge, in the spirit of the agenda-setting reviews that have shaped earlier waves of socio-technical research.

Table 4. Research agenda priorities, designs, and methodological challenges.

#	Priority stream	Proposed design	Primary methodological challenge
P1	Longitudinal trust dynamics	Multi-wave panel studies with mixed measures	Retention and reactivity in long-horizon designs
P2	Identity reconfiguration	Comparative qualitative case studies across professions	Capturing tacit identity work; ethnographic access
P3	Layered accountability	Process-tracing of deployment incidents	Selection bias toward salient cases; counterfactual reasoning
P4	Cross-cultural governance	Comparative institutional analysis across jurisdictions	Functional equivalence of indicators across contexts

#	Priority stream	Proposed design	Primary methodological challenge
P5	Multi-level integration	Nested designs linking micro practices to macro conditions	Theoretical and empirical bridging between levels

The framework also has implications for practice. Organizations deploying generative AI should treat trust calibration as a design problem with measurable indicators (override frequency, appropriate disagreement, verification compliance), identity work as a domain for explicit professional development investment, and accountability as a configuration that requires periodic review as model capabilities and use contexts evolve. Regulators should design governance arrangements that operate adaptively rather than prescriptively, recognizing that the configurations across the four pillars will continue to evolve. Professional associations should articulate evolving accounts of distinctively human competencies and provide their members with the conceptual resources to negotiate identity work without surrendering professional standing.

Limitations of the analysis deserve acknowledgment. The framework abstracts from substantial heterogeneity within each domain and each dimension; refinement will require domain-specific elaboration. The corpus is restricted to English-language peer-reviewed sources, which constrains geographic generalizability. The descriptive bibliometric figures depend on database coverage and search strategy, and other reasonable choices would yield different counts even when the overall patterns would likely be preserved. Finally, the framework is presented as a coherent structure, but it is unlikely to be the only viable structure; alternative integrations—for instance, organizing around power and inequality rather than around the trust-identity-accountability triad (Noble, 2018; Selbst et al., 2019)—deserve continued development.

8. Conclusion

Generative artificial intelligence has moved from research demonstration to general-purpose technology in less than five years. The acceleration creates a coordination problem that requires more than technical evaluation: workers must decide when to trust model outputs, organizations must renegotiate professional identities, and institutions must allocate responsibility for generative output. This article has proposed a socio-technical framework that links trust calibration, identity work, distributed accountability, and adaptive governance into a coherent structure for both empirical research and practical intervention. The framework has been examined against the findings of 247 peer-reviewed studies and against the distinctive configurations observed in healthcare, education, knowledge work, and public governance. Three propositions follow: trust can be engineered through verification routines, identity work determines augmentation versus displacement, and accountability should be distributed in proportion to causal and epistemic contribution. The agenda for future research emphasizes longitudinal designs, comparative cross-cultural inquiry, and multi-level integration. As generative AI continues to mediate an expanding share of social and economic activity, the analytic frameworks through which the field thinks about its consequences will shape not only research but also the regulatory and organizational arrangements that govern deployment. The socio-technical framework proposed here is offered as one contribution to that shared intellectual project.

Acknowledgement

The authors thank colleagues at the participating institutions for comments on earlier drafts and the anonymous reviewers whose suggestions improved the framework and its application. The research did not receive external funding.

Reference

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188–2244. https://doi.org/10.1086/705716

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

Anthony, C. (2021). When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies. Administrative Science Quarterly, 66(4), 1173–1212. https://doi.org/10.1177/00018392211016755

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Bijker, W. E., Hughes, T. P., & Pinch, T. (Eds.). (1987). The social construction of technological systems: New directions in the sociology and history of technology. MIT Press. https://doi.org/10.7551/mitpress/11787.001.0001

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint. https://doi.org/10.48550/arXiv.2108.07258

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. European Law Journal, 13(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Buçinça, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Cooper, R. B., & Zmud, R. W. (1990). Information technology implementation research: A technological diffusion approach. Management Science, 36(2), 123–139. https://doi.org/10.1287/mnsc.36.2.123

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. Science and Engineering Ethics, 26(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1112–1123). ACM. https://doi.org/10.1145/3593013.3594067

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. Journal of Computer-Mediated Communication, 25(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. Government Information Quarterly, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.101493

Jiang, Q., Zhang, Y., & Pian, W. (2022). Chatbot as an emergency exit: Mediated empathy for resilience via human–AI interaction during the COVID-19 pandemic. Information Processing & Management, 59(6), 103074. https://doi.org/10.1016/j.ipm.2022.103074

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2023). Towards a science of human-AI decision making: A survey of empirical studies. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1369–1385). ACM. https://doi.org/10.1145/3593013.3594087

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence, 4(8), 669–677. https://doi.org/10.1038/s42256-022-00516-1

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

Orlikowski, W. J., & Scott, S. V. (2008). Sociomateriality: Challenging the separation of technology, work and organization. Academy of Management Annals, 2(1), 433–474. https://doi.org/10.5465/19416520802211644

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Pakarinen, P., & Huising, R. (2024). Relational expertise: What machines can't know. Journal of Management Studies, 61(6), 2451–2480. https://doi.org/10.1111/joms.12915

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nature Machine Intelligence, 5(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. Nature, 568(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Toward human-AI hybrids. MIS Quarterly, 43(1), iii–ix. https://doi.org/10.25300/MISQ/2019/13703

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. AI & Society, 36(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Schmidt, A. (2020). Interactive human centered artificial intelligence: A definition and research challenges. In Proceedings of the International Conference on Advanced Visual Interfaces (pp. 1–4). ACM. https://doi.org/10.1145/3399715.3400873

Schoeffer, J., Kuehl, N., & Machowski, Y. (2022). 'There is not enough information': On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1616–1628). ACM. https://doi.org/10.1145/3531146.3533218

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 59–68). ACM. https://doi.org/10.1145/3287560.3287598

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review, 56(11), 12799–12831. https://doi.org/10.1007/s10462-023-10420-8

Strich, F., Mayer, A. S., & Fiedler, M. (2021). What do I do in a world of artificial intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. Journal of the Association for Information Systems, 22(2), 304–324. https://doi.org/10.17705/1jais.00663

Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). Journal of Computer-Mediated Communication, 25(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Susskind, R., & Susskind, D. (2015). The future of the professions: How technology will transform the work of human experts. Oxford University Press. https://doi.org/10.1093/oso/9780198713395.001.0001

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Taeihagh, A. (2021). Governance of artificial intelligence. Policy and Society, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), The social psychology of intergroup relations (pp. 33–47). Brooks/Cole. https://doi.org/10.4324/9781315127989-16

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 1–18). ACM. https://doi.org/10.1145/3351095.3372833

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM. https://doi.org/10.1145/3313831.3376301

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. American Economic Review, 108(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062