Special Issue on Privacy As Contextual Integrity

CA-CI: Integrating Contextual Integrity and the Capabilities Approach for Dignity Considerations in AI Governance

Kat Roemmich, *University of Michigan, Ann Arbor, MI, USA*
Kirsten Martin, *Carnegie Mellon University, Pittsburgh, PA, USA*
Florian Schaub, *University of Michigan, Ann Arbor, MI, USA*

Abstract—Capabilities approach–contextual integrity (CA–CI) extends contextual integrity by integrating dignity thresholds from the capabilities approach and specifying purpose as a constitutive parameter. We demonstrate how CA–CI can operationalize the EU AI Act's fundamental rights impact assessments, harm thresholds, and anticipatory governance.

The widespread deployment of AI systems introduces privacy risks and governance challenges that scale with model complexity, autonomy, and cross-domain integration. Regulators, providers, and deployers alike now struggle to manage risks within architectures that learn and generalize autonomously. As these systems evolve, the once-assumed observability, traceability, and contextual stability of information flows erodes as their potential for breach, misuse, and dignitary harm grows. Addressing these challenges requires a governance framework that can evaluate the normative appropriateness of AI systems beyond narrow tasks and stable contexts—a challenge this article takes up by integrating Contextual Integrity with the Capabilities Approach.

Governance must confront new challenges associated with emergent capabilities and representational inferences as AI systems internalize, reconstruct, and propagate information about the world and its inhabitants. These include the continual generation, retention, and circulation of latent features, embeddings, and other internal representations through which systems infer and act upon sensitive regularities about individuals and groups. Once produced, such representations can be reactivated or recombined for new purposes far removed from their original provenance. As durable components of the computational environment, they recursively shape how future information is perceived, classified, and acted upon.

Empirical research shows that even models trained for narrow purposes can develop sensitive and unanticipated capacities. Systems may internalize sensitive attributes (socio-demographic categories, health traits, political leanings, emotional patterns) latent in the data, with embedding vectors and other internal representations particularly prone to privacy leakage [1]. Moreover, models may develop emergent, privacy-intrusive abilities even in sensitive contexts despite safeguards. For example, generic retrieval models trained for object search in law enforcement settings have been shown to acquire unintended person re-identification abilities through overlearning, enabling the identification and profiling of individuals even when trained exclusively on non-human data [2]. These findings illustrate the growing difficulty of tracing, constraining, and anticipating the sensitive knowledge that systems infer and retain as they evolve across tasks and contexts.

Such dynamics are intensified by the rise of foundation models designed for broad capacity, continuous adaptation, and purpose fluidity. Trained on heterogeneous corpora and fine-tuned across tasks, these systems enable features learned for one purpose to activate in another. The same latent representation can serve multiple functions, and models increasingly operate across contexts by design [3]—straining the context-relative and purpose-specific risk distinctions central to privacy and AI governance. In multi-tenant deployments, common parameters and shared embedding or retrieval layers further challenge assumptions that data and representations can remain contextually bounded, even when organizational policies posit strict

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Digital Object Identifier 10.1109/MSEC.2026.3654404

isolation [4]. To illustrate, consider a customer support chatbot: while users engage these systems for service resolution, their interaction data (conversation transcripts, metadata, telemetry logs) may be absorbed into shared embedding spaces, retrieval indexes, or alignment parameters that inform other downstream deployments [5]. A customer's angry exchange may later influence models used to rank job candidates in sourcing databases, or to target advertisements that exploit emotional tendencies.

The structural features that enable model adaptability, transfer, and generalization destabilize contextual and use-based boundaries, creating what may be described as a regulatory paradox: laws premised on stable contexts and bounded purposes must nonetheless govern systems whose very function is to transcend them. Yet the normative anchors of proportionality and necessity remain: organizations should collect only what is needed to fulfill legitimate purposes, use it only as declared, and remain compatible with human dignity. Accordingly, privacy and AI governance frameworks continue to require context and purpose specification even as they shift from governing information flows to governing AI models.

In the EU, the General Data Protection Regulation (GDPR) enshrines a purpose limitation principle, requiring data to be "collected for specified, explicit and legitimate purposes and not further processed in a manner incompatible with those purposes," and mandates data protection impact assessments (DPIAs) for high-risk data processing that may affect fundamental rights and freedoms (Art. 35) [6]. The EU AI Act extends this logic: it prohibits AI practices deemed to present an unacceptable risk to fundamental rights, health, or safety (Art. 5); requires certain deployers of high-risk systems (Art. 6) to conduct fundamental rights impact assessments (FRIAs) prior to deployment and after relevant system changes, complementary to DPIA obligations under the GDPR (Art. 27); and obliges providers to maintain continuous, purpose-specific risk assessments throughout the system lifecycle (Art. 9) [7]. Risk classifications hinge on factors such as deployment context, intended purpose, technical characteristics, and the nature and severity of potential harm, yet no unified standard defines how these criteria should be comparatively evaluated to determine context-relative risk [8].

The international consensus on dignity's inviolability [9] provides the normative source of entitlements that regulatory instruments like the GDPR and EU AI Act seek to protect [10], yet the concept of dignity itself remains operationally under-specified. Guidance lacks a clear standard for determining what constitutes a violation to dignity beyond broad reference to fundamental rights [8]. These ambiguities hinder evaluators in determining when a given practice crosses the moral boundary of dignity—and by extension, the derivative human rights it grounds. Consequently, dignity's enforceability as a foundational normative principle becomes increasingly tenuous: the conditions under which violations occur are made indeterminate by governance under-specification and are further obscured by architectural opacity. Meeting this challenge requires a normative governance framework for privacy and data protection that can substantively assess dignity risks across evolving socio-technical contexts throughout the AI lifecycle.

Nissenbaum's Contextual Integrity (CI) offers a promising foundation. Evaluating the appropriateness of information flows relative to social context, Contextual Integrity structures its evaluation criteria by five inter-dependent parameters to ask whether exchanges constituted by specific actors (subject, sender, recipient), data attributes, and transmission principles conform to contextual norms and aims [11]. The theory traditionally treats informational purpose as an optional transmission principle constraining how information is shared—capturing purpose constraints implicitly by convention. Yet because Contextual Integrity draws its normative force primarily from established norms, its guidance is strained in novel socio-technical contexts where norms are unsettled. And because Contextual Integrity leaves purpose under-specified and does not articulate the considerations for dignity demanded by emerging privacy, data protection, and AI regulation, it is under-equipped for today's governance challenges.

We introduce CA-CI, a normative framework for privacy and AI governance that extends Contextual Integrity's evaluative methodology and normative grounding to meet these conditions. First, CA-CI elevates purpose to a sixth constitutive parameter of an information flow. Explicitly specifying purpose enables evaluators to track shifting contexts of use that can remain indiscernible when the original five parameters appear stable—for instance, in cases of scope creep or representational reuse—and to assess compatibility with contextual aims even where norms are absent or ambiguous. Second, CA-CI adds a special class of fixed transmission principles that specify threshold conditions for dignity as a universal moral minimum, supplying a second normative basis for legitimacy that holds across contexts. CA-CI operationalizes Nussbaum's Capabilities Approach (CA) to specify what a dignified human life minimally requires: an irreducible set of ten core capabilities that together constitute dignity when agency for every person is secured

at threshold levels in matters of life; bodily health; bodily integrity; the senses, imagination, and thought; emotions; practical reason; affiliation; other species; play; and control over one's environment [12]. Where any capability threshold is not secured, environments fail the overlapping consensus on human dignity and thus trigger a duty of intervention. Uniting Contextual Integrity and the Capabilities Approach into a single normative governance framework, CA-CI evaluates privacy and dignity in any socio-technical context, novel or entrenched, by whether it secures the integrity of social life and each human life within it.

This article proceeds by elaborating CA-CI's theoretical extensions to Contextual Integrity: (1) specifying dignity as a universal moral minimum standard and (2) clarifying purpose's normative role. We then demonstrate CA-CI's practical value for AI governance through three applications to the EU AI Act: (1) Fundamental Rights Impact Assessments (FRIAs), (2) significant harm thresholds, and (3) anticipatory governance. While the EU AI Act provides a compelling case study given its rights- and risk-based framework grounded in dignity, CA-CI is broadly applicable to evaluate privacy and dignity in any socio-technical context, regardless of jurisdiction. Together, we show how extending Contextual Integrity's normative conception of privacy as appropriate flows of information with a second legitimacy standard that specifies what dignity requires provides a powerful normative framework for making ethics operational in AI governance.

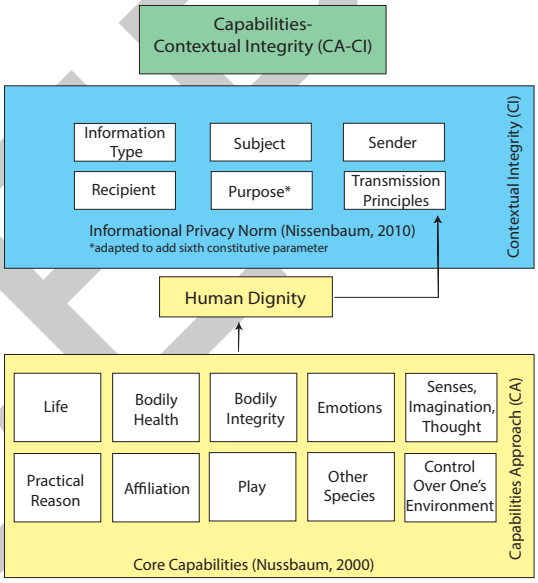

FIGURE 1. Capabilities–Contextual Integrity (CA–CI) Theoretical framework extending CI by (1) integrating dignity thresholds as a special class of fixed transmission principles and (2) adding purpose as a constitutive parameter.

Grounding Governance in Dignity

As AI systems increasingly turn toward general purpose architectures and rights-based governance, CA-CI (see Figure 1) addresses the challenge of evaluating privacy and dignity risks with the contextual sensitivity such assessments require.

By synthesizing Contextual Integrity's structured account of context-relative information flows—here extended with purpose as a sixth parameter—with the Capabilities Approach's specification of dignity thresholds, CA-CI provides a systematic framework for assessing how AI systems affect privacy and dignity across their lifecycle. Contextual Integrity provides the analytic means to determine whether informational practices are appropriate to a context's social purposes and functional aims. The Capabilities Approach adds a complementary test of what any contextual norm, purpose, or practice must respect at minimum: the conditions required to realize a dignified life. Together, they establish a dual standard of legitimacy: information flows—understood broadly, whether as explicit transmissions between actors, representational inferences that traverse AI system deployments, or internal model states that mediate system adaptations and behaviors—must serve both the *telos*, or ultimate aim, of a context, and the capabilities constitutive of dignity.

Table 1 illustrates the complementary strengths of each theory. Integrated into a unified normative approach for evaluating socio-technical systems, Contextual Integrity and the Capabilities Approach reinforce each other to ensure that, even as contexts shift and norms adapt, respect for persons remains nonnegotiable and the integrity of social life is preserved. The following sections explicate these theoretical extensions in further detail.

Dignity as a Universal Moral Minimum

Adding core capability-based dignity thresholds as a fixed class of transmission principles extends Contextual Integrity to preserve the integrity of persons. On the Capabilities Approach, core capabilities are the essential constituent parts of dignity as a whole; where

Theoretical Dimension	Contextual Integrity (CI)	Capabilities Approach (CA)
Aim	Preserve the integrity of social life.	Preserve the dignity of persons.
Normative Focus	Contextually appropriate data flows.	Substantive capabilities for a dignified life.
Limitations	Norm dependency limits global guidance; does not evaluate dignity.	Minimum thresholds limit local guidance; does not evaluate privacy.

TABLE 1. Comparative Overview of Contextual Integrity and Capabilities Approach.

any core capability falls below its sufficiency threshold, dignity is violated. The claim is ontological: if the parts are not secured, the whole cannot obtain [12].

International legal instruments recognize dignity as inviolable, with the 1948 Universal Declaration of Human Rights establishing international consensus, and each of the human rights declared serving as derivative entitlements grounded in that worth [9], [12]. The EU Charter of Fundamental Rights, which became legally binding in 2009, extended this foundation to include rights to data protection as a fundamental entitlement, reflecting digital-age threats to dignity unforeseen at the time of the UN's international agreement [10]. Yet rights name instruments; they do not by themselves specify the content of a dignified life or the conditions under which dignity is lost. If individuals possess "rights" yet lack the ability to exercise them, then those rights ring hollow. Rights that cannot be enacted in practice fail to secure dignity in substance. The problem, then, is not how to affirm rights in principle, but how to realize them in practice.

The Capabilities Approach was developed to overcome this limitation within welfare economics, a leading framework guiding public policy. Rather than measuring equality and justice in a society by its distribution of resources and social goods (including individual rights), the normative theory evaluates whether each person can realistically convert them into activities and choices that enable them to flourish. Offering a guide to constitutional design, Nussbaum's version specifies the core content of a dignified life: ten irreducible capabilities whose securement for every person, at threshold, constitutes the moral floor of a "truly human life" [12]. As a moral minimum grounded in the basic requirements for a human form of life to thrive, these thresholds specify the evaluative boundary below which practices become degrading to humanity and thereby dignity-violating. Because individual circumstances differ, so too do individual capacities to convert available goods and resources into a dignified life. Accordingly, empirical work operationalizing the Capabilities Approach evaluates the personal, social, and environmental conditions that enable or impede capability attainment as conversion factors, with the goal to secure conditions that enable every person to realize the effective opportunities required for flourishing, if they so choose. This critical emphasis on human agency ensures the Capabilities Approach remains sensitive to cultural variation, a normative commitment to values pluralism it shares with Contextual Integrity.

Contextual Integrity's pluralist commitments draw from Walzer's notion of "complex equality," where multiple autonomous social spheres are each governed by their own principles of distribution and merit. By deferring evaluative authority to the domains that constitute meaning and settle distribution, Contextual Integrity's "justificatory framework" grants presumptive legitimacy to established information flows—rooting its standard of appropriateness in the lived histories and normative grammars which sustain a domain's structural and moral coherence [11]. Yet Contextual Integrity also inherits Walzer's limits, articulated in his distinction between "thick" and "thin" morality: thick morality encompasses the diverse ways in which communities instantiate and elaborate shared values, yet without a thin universal set of moral minimum principles to anchor thick elaborations, Walzer warned, the autonomy of social spheres remains precarious—vulnerable both to internal corruption and external distortion through tyrannical imposition [13]. Though hesitant to specify the content of moral minimums, Walzer pointed to international human rights as a possible source. Following Walzer, Nussbaum defends the core capabilities constitutive of dignity as a thin, universally applicable minimum standard. Below threshold, the environment is degrading—in the vocabulary of Contextual Integrity, inappropriate—regardless of local justification. Above threshold, pluralism reigns: communities rightly differ in how they weigh, pursue, and distribute diverse capabilities and goods. In this way, the Capabilities Approach to dignity is not a limit opposed to pluralism but a condition of its very possibility.

Integrating dignity thresholds into Contextual Integrity as a universal moral minimum makes the framework resilient under novel socio-technical conditions. As AI models increasingly optimize across tasks and traverse contexts by design, CA-CI restores normative stability in their evaluation with dignity as an external

ACCEPTED

evaluative standard—even as norms are unsettled, contested, or ambiguous. More practically, core capability thresholds render dignity empirically accessible. Because capabilities concern what people can actually do and be, evaluators can ask whether the socio-technical environment provides the conversion conditions necessary for threshold realization: where conversion is impossible, the practice is dignity-violating; where conversion is impaired but remediable, the risk is high and must be mitigated; where conversion is made possible, the practice is legitimate. Rights, where recognized, may be the vehicles of dignity's protection, but capabilities supply the standard to specify when protection succeeds or fails—across contexts, for whatever the purpose, within evolving socio-technical environments.

The Normative Role of Purpose

In Contextual Integrity, respect for context-relative standards of appropriateness upholds the integrity of social life, presumed to promote the telos of a context. Yet this heuristic leaves implicit a crucial assumption: that the ultimate ends of each domain are themselves worthy of pursuit. Healthcare, workplaces, educational institutions—such domains are not valued merely because they sustain established social practices; they are valued because they secure the conditions by which people can live purposeful and dignified lives. The legitimacy of any domain thus depends on the teleological alignment of its everyday practices with these human ends.

When Contextual Integrity was first theorized, socio-technical systems were more clearly delineated by social domains, and informational purpose could be reasonably inferred from context [11]. Under these conditions, treating purpose as an optional transmission principle was a workable heuristic. Yet as we move toward general purpose architectures that autonomously generate inferences and operate across contexts, this assumption no longer holds. AI systems increasingly repurpose representations learned from one task to serve entirely different ends, straining the context-relative norms that once implicitly constrained purpose—and undermining their reliability as a unit of normative evaluation.

In Contextual Integrity, an informational norm is the structure that makes a flow the kind of act it is, constituted by five core parameters: data subject, sender, recipient, attributes, and transmission principles. But general purpose AI systems can reuse learned internal representations across tasks in ways that shift the practice a flow instantiates without any salient change in those parameters. Consider employees using an internal AI assistant that relies on a shared embedding space to support routine work—summarizing meetings, drafting emails, and answering questions from internal knowledge bases. The organization retains the resulting artifacts (embeddings, retrieval indexes, interaction logs) accessible to a limited internal group for security monitoring and policy compliance; meanwhile, the model's learned representations remain available for general purpose reuse. Later, that same representation store and model are redeployed by the same actors to infer employee skills, match workers to roles, or generate proxies for productivity—without introducing new data subjects or recipients, and without an obvious change in attributes or transmission principles prior to downstream outputs. The shift in use alters what the flow *is* as a workplace practice—and what normative justifications and safeguards it requires. Purpose is what makes the flow intelligible as one kind of practice rather than another; if purpose is left unspecified, such shifts can be overlooked in conventional Contextual Integrity evaluations because the five core parameters can appear stable even as cross-task reuse changes the flow's normative significance.

Formalizing purpose as a sixth constitutive parameter in Contextual Integrity supplies a means to trace a flow's practical ends to the telos of its context and assess their compatibility. In the workplace example, if the role of workplaces in a just society is to secure a domain in which people can exercise their capabilities to live free, equal, and dignified lives, then socio-technical practices that diminish those possibilities by extending authoritarian control over workers corrode that legitimacy [14]. For an employer using systems that apply representations learned about workers to generate inferences about their skills or productivity, we can evaluate whether those uses promote the context's telos, whether they are necessary to achieve it, and whether their means are compatible with human dignity. Purpose thus ensures that flows are evaluated not only for conformity to prevailing norms, but for whether they serve the ends that give contexts their normative standing—enabling us to distinguish legitimate from illegitimate practices even as norms remain unsettled or become ambiguous in general-purpose, cross-domain AI systems.

Case Study: CA-CI and the EU AI Act

CA-CI's contributions extend beyond theory and have practical value for AI governance by providing a normative framework that is robust enough to surface and

ACCEPTED

evaluate a broad range of ethical considerations, and specific enough to guide how we ought to address them in order to uphold both local norms and globally shared values.

In the following case study, we illustrate CA-CI's systematic approach to evaluating privacy and dignity in AI systems by applying it to key requirements of the EU AI Act. Specifically, we show how the framework (1) enables context-sensitive assessment of dignity risks within Fundamental Rights Impact Assessments, (2) defines principled thresholds for what counts as significant harm, and (3) supports anticipatory governance by identifying dignity-based risks that have not yet been recognized or codified.

Risk Classifications and Fundamental Rights Impact Assessments

The EU AI Act prohibits certain AI practices that pose a clear threat to the rights, safety, and dignity of persons (Art. 5), while permitting AI systems classified as high-risk subject to additional safeguards and requirements (Art. 6)—including, for certain deployers, an assessment of impacts on fundamental rights for intended uses (Art. 27) [7]. Yet providers and deployers face persistent ambiguities, both evaluative and operational: What distinguishes unacceptable risk from high risk? How should EU Charter rights and freedoms be assessed within concrete socio-technical contexts of use?

CA-CI supplies a structured foundation for Fundamental Rights Impact Assessments (FRIAs) by linking context-relative evaluations of information flows to capability thresholds that specify the minimum conditions for a dignified life. As formal entitlements codified to protect dignity [10], Charter rights can be mapped onto each of the ten core capabilities that constitute dignity. By evaluating the contextual parameters of information flows (including purpose) against those thresholds, CA-CI offers a principled way to identify where an AI system risks undermining dignity in practice, and therefore to surface corresponding impacts on fundamental rights.

Consider an algorithmic management system that draws on enterprise data, including personal employee information, to make automatic scheduling decisions. The tool, supplied by a third-party vendor, is built on a foundation model with a shared embedding and retrieval infrastructure, used across HR functions (e.g., scheduling, performance management, succession planning). Focus on one data class: individual performance metrics derived from time-keeping records, annual reviews, employee surveys, internal communications metadata, workstation logs, and sensor data used to proxy bio-physiological signals of fatigue and stress. For each flow of information, CA-CI's six analytic parameters specify what data is involved (attributes), who it is about (data subjects), who transmits and receives it (senders and recipients), for what purpose, and under what constraints (transmission principles, such as access controls). Aggregating and mining these sources instantiates a new practice when combined into performance indicators, and another when those indicators are used to allocate shifts. Where model representations are reused across HR applications, shifts may occur without obvious changes to core Contextual Integrity parameters, making them difficult to detect. And because these transformations are novel and generally inconspicuous to employees, norms may be unsettled. In CA-CI, however, the parameters still structure normative evaluation by asking whether each flow is appropriate to the practical end in view and compatible with the workplace context's telos, as constrained by capability thresholds.

This analysis makes salient a wide range of implicated Charter rights [10]. Expanding the purposive scope of performance metrics to include scheduling heightens their stakes, potentially increasing stress and behaviors that negatively affect *bodily health*, prompting scrutiny under EU Charter rights such as Integrity of the Person (Art. 3). Metrics that incorporate fatigue or stress indicators derived from bio-physiological proxies may be perceived as intrusive and affect *bodily integrity*, *emotions*, *and senses, imagination, and thought*, possibly affecting Freedom of Thought, Conscience and Religion (Art. 10), Workers' Right to Information and Consultation (Art. 27) and Non-discrimination (Art. 21). Likewise, using interpersonal communications data in scheduling may shift employee relations in ways that burden capacities for *affiliation*, raising questions under Respect for Private and Family Life (Art. 7), Freedom of Assembly and Association (Art. 12), and equality protections (Arts. 20–23). More generally, because scheduling decisions shape livelihoods and working conditions, they can implicate *life*, *play*, *bodily health*, and *control over one's environment* through effects on autonomy, pay predictability, benefits eligibility, and work-life balance that impede Freedom to Choose an Occupation and Engage in Work (Art. 15), Health Care (Art. 35), and Fair and Just Working Conditions (Art. 31). Illustrative rather than exhaustive, these examples show how CA-CI connects context-relative dignity risks to codified rights by pinpointing where risks emerge at the level of data flows.

In turn, this exercise helps evaluators differentiate

risk tiers, i.e., high versus unacceptable risk, by making visible when dignity is at risk (capabilities may be negatively impacted) and clarifying when conditions plausibly amount to dignity violation (capabilities are foreseeably driven below threshold). This sharpens judgments about whether and how risks can be mitigated. Some risks, for instance, may be introduced by and limited to particular data sources, such as bio-physiological signals that implicate *bodily integrity*. Evaluators would consider whether impacts to bodily integrity risk a person's sovereignty over their body—their effective authority to act and make choices about their physical self without coercion or other means of external control—and how this effect occurs in practice. Where the context of use lacks sufficient freedom for real choice, evaluators may deem the practice a violation of bodily integrity and thus an unacceptable risk. By locating the risk at a particular input, evaluators can see how it might be eliminated through redesign that excludes those data sources entirely.

Whether sufficient dignity risk mitigation is possible, and what it should look like, will depend on the context at hand: risks from identifiable information exposure might be addressed through privacy-enhancing techniques, while risks from transparency or accountability failures might be addressed through organizational governance policies and independent oversight. The key is that a socio-technical understanding of how threats to dignity may be experienced in everyday practice allows evaluators to trace where respect for fundamental rights is strained and to identify design and governance measures that meet capability thresholds. In this way, the capability-rights linkage supports FRIA practice by assessing whether an AI system's socio-technical environment can realize the rights it implicates and by clarifying where regulatory boundaries between acceptable and unacceptable risk may lie.

Significant Harm Thresholds

Where FRIAs evaluate risks to rights, the AI Act also requires evaluations of risks of harm. The Commission's draft guidelines clarify, for example, that prohibited or tightly controlled practices such as AI systems deploying subliminal, deceptive, or manipulative techniques (Art. 5) must be assessed for the severity and reasonable likelihood of harm, including potential physical, psychological, financial, or economic harm, with attention to compounding effects over time [8]. Importantly, significant harm thresholds can be crossed even when injury unfolds gradually, such as addiction-like dynamics that exacerbate vulnerabilities or creeping erosions of autonomy that materialize only in the long term.

Yet regulators and operators alike lack clear guidance on which harms matter, and when they become significant. As a result, some of the most normatively pressing and socially consequential concerns—manipulation, exploitation—remain difficult to assess. When does algorithmic influence cross the line into significant harm? When do incremental restrictions of autonomy crystallize into injuries warranting AI prohibition, and by what benchmarks can autonomy constraints be judged?

CA-CI's threshold logic offers a response to these inquiries: significant harm corresponds to the dignity violation condition, where, in context, an AI system's effects foreseeably drive any core capability below the minimum needed for dignified existence. This model clarifies three regulatory uncertainties: (1) which harm categories and thresholds are relevant; (2) how context should shape evaluation; and (3) how to track harms that compound over time.

Harm categories and thresholds.
Core capabilities supply both the evaluative targets and thresholds for determining when harm becomes significant, and they can be cross-walked to harms already recognized in privacy torts doctrine. Citron and Solove's taxonomy of cognizable privacy harms [15] overlaps considerably with the harm categories identified in the Commission guidelines [8]. Core capabilities capture these legally recognized harms, while articulating risks of normative significance that may go unnoticed in doctrinal categories yet bear directly on dignified functioning.

Physical harms. Physical harms resulting in bodily injury or death threaten capabilities for *bodily health*, *bodily integrity*, and *life*. Thresholds are crossed when the effects of an AI system interfere with access to, or constrain decision-making about, basic health conditions (including reproductive health, nourishment, and shelter); restricts free movement; undermines security against violent assault; shortens life expectancy; or otherwise reduces conditions of embodied functioning below what is compatible with dignity. In general purpose AI deployments, such harms can arise when personal inferences (e.g., intent, physical condition, location) are learned and later reused in safety- or access-critical settings like workplace safety enforcement, emergency triage, or housing eligibility—predictably exposing individuals to coercive bodily constraints or unsafe conditions.

Reputational harms. Reputational harms that injure one's standing in a community implicate capabilities for *affiliation* and *control over one's environment*. *Affiliation* impacts become significant when they

impede the ability to live with and toward others in relations of mutual recognition; to secure the bases of self-respect and non-humiliation; or to be treated as an equal whose worth is acknowledged. Thresholds for *control over one's environment* are crossed when reputational effects foreseeably impair one's capacity to pursue employment, housing, or education; to participate on equal terms in political or economic life; or to otherwise exercise meaningful choice over the social conditions that structure one's life opportunities. Effects may cascade into degradations of other core capabilities including *practical reason* and *emotions*. In general purpose settings, representational inferences learned from interaction histories may internalize that an individual or class of users are "low-trust" or have "unreliable" dispositions, which may then be activated in downstream hiring, credit, or fraud systems—importing reputational stigma across contexts.

Psychological harms. Psychological harms induce distress or disturbance, burdening capacities for *emotions* alongside *practical reason* and *senses, imagination, and thought*. Thresholds are crossed where emotional life is impaired such that one cannot sustain attachments, love and grieve appropriately, or develop without being blighted by fear and anxiety; where reasoning is impaired such that one cannot form a conception of the good or engage in critical reflection about life planning; or where cognitive and creative capacities for imagination, thought, and expression are stifled. Downstream effects may further undermine *affiliation* and *control over one's environment*, among other capabilities. AI companions engineered for high empathy and agreeability, for instance, may reinforce maladaptive beliefs or foster emotional dependency, producing cumulative distress that degrades emotional functioning, critical reflection, and self-expression below threshold.

Economic harms. Economic harms resulting in monetary or opportunity loss strike *control over one's environment*. Thresholds are crossed when economic impacts foreseeably foreclose access to work, property, education, or basic material security; or prevent meaningful participation in labor, social, and political life on equal terms. These constraints may cascade into effects on core capabilities including *emotions*, *practical reason*, and *affiliation*. In ad targeting and dynamic pricing systems that use foundation model representations, inferred vulnerabilities and propensities may be used to personalize offers or terms in ways that systematically extract economic surplus or exclude individuals from economic opportunities, producing material precarity.

Discrimination harms. Discrimination harms disadvantage protected groups, impairing *affiliation*, *practical reason,* and *emotions*, and *control over one's environment* by restricting access to employment, services, or civic participation. Foundation model representations encode proxy attributes (e.g., dialect, cultural references) that may reproduce disparate outcomes even without explicit protected-class inputs, especially when reused across multiple decision contexts.

Relationship harms. Relationship harms damaging personal, professional, or institutional relationships undermine *affiliation*, *practical reason*, and *emotions*, with context-relative spillovers onto capabilities such as *play* and *control over one's environment*. For instance, continuous surveillance systems in the workplace can chill association among workers and erode trust in employers, reshaping these relationships by inducing conditions of suspicion and self-censorship.

Autonomy harms. Autonomy harms impair agency over both ends and means. They include coercion (limiting real choice); failure to inform (withholding information needed for action); manipulation (steering decisions beyond the agent's cognizability); thwarted expectations (contradicting stated purposes or promises); loss of control (denial of meaningful management over personal information); and chilling effects (deterring speech, association, or belief under surveillance pressures). Because dignity requires agency to develop and exercise each capability, autonomy harms can implicate any core capability. In adaptive systems optimized for engagement or conversion, inferred vulnerabilities may enable micro-targeted behavioral steering that progressively narrows a person's perceived option set; thresholds are crossed when such steering foreseeably degrades *practical reason* below threshold by undermining capacities for critical reflection and self-authorship.

For all of these harms, CA-CI's capability thresholds specify when impairments amount to significant harm.

Context-relative evaluation.

The Commission treats harm significance as a fact-specific, case-by-case inquiry but offers limited guidance on how context should structure that assessment [8]. CA-CI provides a model to operationalize contextual analysis by linking CI's context-sensitive diagnostics for information flows to concrete harm vectors, while fixing the decisive condition for significant harm at core capability thresholds.

Consider, for example, the EU AI Act's prohibition on emotion recognition in the workplace, with narrow exceptions for medical or safety purposes (Art. 5) [7]. Where such a deployment is permitted and subject

to governance obligations (including, where applicable, FRIA practice), CA-CI specifies how to evaluate whether the system is justified in context. System inputs (e.g., biometric data) and outputs (e.g., fatigue detection) are captured as information attributes; actors are identified (data subjects, senders, recipients); purposes are specified (e.g., safety, medical); and transmission principles (regulatory requirements, organizational policies, safeguards) are described. The evaluation then connects system behaviors and design choices to context-relative capability impacts.

This supports both harm evaluation and targeted mitigation. If fatigue detection flags the state of an employee as a safety risk and alerts supervisors, HR, and operational leaders, traditional privacy models would emphasize safeguards such as access limitation and secure data storage to prevent unauthorized data leakage. While important and necessary, these technical measures are not sufficient to address the range of dignitary considerations present in this context. CA-CI can guide evaluators to ask: Besides for safety enhancement, could supervisors also use fatigue data to question worker commitment or reliability, eroding *affiliation* by undermining mutual recognition at work? Could HR link fatigue patterns to performance evaluation, creating conditions that impair workers' *bodily health* and capacity to engage *practical reason* in critical work-life decisions? Could the very knowledge of monitoring chill *affiliation* among colleagues, eroding dignity through pressures of non-humiliation? Even if necessity, minimization, and anti-discrimination measures are in place, dignity thresholds may still be at risk, prompting evaluators to consider complementary safeguards such as prohibiting integration of fatigue-monitoring outputs into performance management; embedding those constraints into data lineage, role-based access, and auditability; and instituting policies, supervisor training, and audits to ensure compliance.

By posing such questions to expose dignity risks, CA-CI identifies where existing socio-technical practices and their privacy safeguards are normatively insufficient and what additional protections and mitigation measures are needed to keep impacts above threshold—enabling context-sensitive evaluation while maintaining consistent dignity standards.

Harmful effects over time.

Commission guidelines specify that harm assessments must consider effects that accumulate over time and exacerbate vulnerabilities, but offer limited guidance on how to operationalize such assessments [8]. CA-CI addresses this by anchoring evaluation in capability thresholds that remain constant even as harms compound, supplying stable targets for longitudinal monitoring.

Individual algorithmic nudges, for instance, may not immediately cross thresholds for *practical reason*, but cumulative effects over months or years can degrade one's capacity to critically reflect upon and plan one's own life below dignified levels. Likewise, continuous workplace surveillance permissible under the AI Act may not instantly foreclose *affiliation*, yet sustained chilling effects may eventually erode workers' capacity to live with and toward each other on terms of mutual recognition and non-humiliation.

CA-CI thus enables assessment of both immediate harms and those that accumulate across time—reputational degradation, cumulative autonomy erosion, economic precarity. Evaluators can track whether capability impacts compound into significant harm, determining when individually minor impediments cumulate into dignity-eroding conditions that warrant prohibition or strict control, as the guidelines require for harms reasonably likely to occur over time.

Anticipatory AI Governance

CA-CI also furnishes a model for anticipatory AI governance. Because Contextual Integrity parameters specify context, roles, attributes, transmission principles, and purpose, while the Capabilities Approach supplies dignity thresholds, the framework provides a principled basis for *ex ante* normative risk assessment that aligns with the EU AI Act's risk classification architecture.

The AI Act already distinguishes risks to dignity and derivative rights from permissible uses by reference to contextual parameters that map to Contextual Integrity's framework. For instance, the AI Act prohibits employers from using employee biometric data for emotion recognition, but permits biometric data use for authentication purposes. CA-CI both formalizes these distinctions and links them to capability thresholds for dignity evaluation. This model enables evaluators to identify dignity risks that may be overlooked by regulatory procedure. Emotion recognition from text rather than biometric inputs, for example, may escape biometric-based prohibitions yet still impair *affiliation* and *practical reason* by enabling similar forms of workplace control. By anchoring evaluation in capability impacts while specifying contextual parameters (e.g., input modalities), CA-CI surfaces risks that purely procedural evaluations can miss.

This anticipatory capacity extends beyond regulatory compliance to organizational governance. For instance, a data scientist requesting access to employee communications metadata to predict burnout

may technically satisfy existing regulatory mandates, yet still create dignity risk liabilities. Because CA-CI's parameters map to data governance systems (catalogs, lineage tracking, access controls), organizations can flag capability impacts when new data sources or purposes are introduced, prompting evaluation before deployment rather than after harm materialization.

CA-CI flags risks to dignity in any socio-technical system, whether or not it has been prohibited or flagged as high-risk. Using its methodology for evaluation can classify new use cases and identify if there is a compatible basis to classify the practice as prohibited or warranting strict regulation.

Conclusion

CA-CI advances Contextual Integrity through two key theoretical extensions that strengthen its normative and empirical adequacy for AI governance, specifying: (1) moral minimum thresholds for dignity as fixed transmission principles and (2) purpose as a constitutive parameter. This model extends Contextual Integrity's capacity to evaluate socio-technical contexts even where novel or contested by establishing a second standard for legitimacy grounded in the basic requirements for dignity from the Capabilities Approach, independent of social norms. Practically, these extensions support providers, deployers, and regulators in evaluating privacy and dignity risks in AI systems, though they are certainly not the only ways that future work may extend Contextual Integrity or apply the Capabilities Approach.

While we establish CA-CI's theoretical foundations and demonstrate its governance utility via applications to the EU AI Act, further research is needed to render the framework empirically robust and normatively calibrated. In line with the capabilities literature, future research should systematically identify the conversion factors (personal, social, environmental) that mediate whether and how individuals can translate entitlements, such as privacy and data protection rights, into genuine capabilities to act, choose, and live with dignity. These factors may include cognitive and affective dispositions linked to privacy risk, such as digital literacy, trust, impulsivity, or situational cognitive load, as well as socioeconomic constraints, language proficiency, disability, institutional power asymmetries, and the material affordances of devices and interfaces. Mapping these conversion environments will clarify how AI systems condition the core capabilities of differently situated persons.

A further empirical task concerns the stability and adaptability of dignity thresholds. Validated instruments developed to measure capabilities in fields such as health and human development can be employed and adapted to measure the impact of AI systems on core capabilities in particular contexts, and enable validation of whether CA-CI evaluations align with stakeholder intuitions and regulatory judgments. Such comparative validation would advance both the empirical operationalization and the normative legitimacy of CA-CI.

Finally, future work should explore institutional and organizational implementation. Embedding CA-CI in enterprise risk management, such as integrating capability assessments into data catalogs and lineage systems, impact assessment workflows, or red-teaming exercises, would investigate its practical feasibility and reveal where dignity thresholds are stable and where they require contextual calibration. If empirical research shows capability-based evaluations to reliably identify dignity violations across socio-technical contexts, CA–CI would offer a stable evaluative framework for AI policy that suits global governance—one that preserves human agency, dignity, and their privacy considerations by grounding its guide for how we build and govern AI systems in the real capabilities that individuals and communities have to live lives they have reason to value.

REFERENCES

1. C. Song and A. Raghunathan, "Information Leakage in Embedding Models," *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, New York, NY, USA, 2020, pp. 377–390, doi: https://doi.org/10.1145/3372297.3417270.
2. A. T. Nguyen, R. Stoykova, and E. Arazo, "Emergent AI Surveillance: Overlearned Person Re-Identification and Its Mitigation in Law Enforcement Context," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, Madrid, Spain, vol. 8, no. 2, pp. 1862–1874, Oct. 2025, doi: https://10.1609/aies.v8i2.36680.
3. Anka Reuel *et al.*, "Open Problems in Technical AI Governance," *Transactions on Machine Learning Research*, 99 pages, 2025. Available: https://openreview.net/forum?id=1nO4qFMiS0
4. K. S. Kumar, J. V. Kumar, K. S. Kumar, and N. V. Kumar, "Security and Privacy Challenges in Multi-Tenant Cloud Architectures: A Comprehensive Analysis," *International Conference on Computing Technologies & Data Communication (ICCTDC)*, Hassan, India, 2025, pp. 1-6, doi: https://10.1109/ICCTDC64446.2025.11158758.
5. I. Barberá, "AI Privacy Risks & Mitigations — Large Language Models," Support Pool of Experts Programme, European Data Protection Board, 2025.

ACCEPTED

Available: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf

6. European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official Journal of the European Union (OJEU)*, Apr. 27, 2016. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
7. European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *Official Journal of the European Union (OJEU)*, Jul. 12, 2024. Available: https://data.europa.eu/eli/reg/2024/1689/oj
8. European Commission, "Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)," Brussels, Belgium, C(2025) 5052 final, Jul. 29, 2025. Available: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act
9. United Nations, "Universal Declaration of Human Rights," General Assembly Resolution 217 A (III), Dec. 10, 1948. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights
10. European Union, "Charter of Fundamental Rights of the European Union," *Official Journal of the European Communities*, vol. C 364, pp. 1–22, Dec. 18, 2000. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2000_364_R_0001_01
11. H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.
12. M. C. Nussbaum, *Women and Human Development: The Capabilities Approach*. Cambridge University Press, 2000.
13. M. Walzer, *Thick and Thin: Moral Argument at Home and Abroad*. University of Notre Dame Press, 1994.
14. E. Anderson, *Hijacked: How Neoliberalism Turned the Work Ethic Against Workers and How Workers Can Take It Back*. Cambridge University Press, 2023.
15. D. K. Citron and D. J. Solove, "Privacy Harms," *Boston University Law Review*, vol. 102, no. 3, pp. 793–863, Apr. 2022, doi: https://dx.doi.org/10.2139/ssrn.3782222.

Kat Roemmich is a research associate at the University of Michigan, where she earned her Ph.D. in Information. Her work examines how emerging technologies affect dignity, privacy, and democratic life, with a focus on advancing research and policy approaches to AI governance that align innovation with ethical and social values. Contact her at roemmich@umich.edu.

Kirsten Martin is H. John Heinz III Dean of the Heinz College of Information Systems and Public Policy at Carnegie Mellon University. Her research focuses on the ethics of technology, privacy, and corporate responsibility, with particular attention to algorithmic accountability and the role of business in shaping ethical data practices. Martin received a PhD. in Business at the University of Virginia. Contact her at kirstenm@andrew.cmu.edu.

Florian Schaub is an associate professor of information as well as electrical engineering and computer science at the University of Michigan, Ann Arbor, MI 48109 USA. His research interests include privacy, human–computer interaction, emerging technologies, and public policy. Schaub received a Ph.D. in computer science from Ulm University. He is a Distinguished member of the Association for Computing Machinery, and a member of IEEE and the International Association of Privacy Professionals. Contact him at fschaub@umich.edu.

ACCEPTED