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—AI systems generate risks that challenge privacy governance by threatening dignity and straining contextual norms. CA-CI extends Contextual Integrity to integrate dignity thresholds from the Capabilities Approach and specify purpose as a constitutive parameter. CA-CI's EU AI Act applications operationalize fundamental rights impact assessments, harm thresholds, and anticipatory AI 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 capabilities 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.

These dynamics are intensified by the rise of foundational models designed for broad capability, continuous adaptation, and purpose fluidity. Trained on heterogenous corpora and fine-tuned across tasks, such models 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 where organizational policies posit strict isolation [4]. To illustrate, consider a customer

XXXX-XXX © 2026 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

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 systems downstream [5]. A customer's angry exchange with a service provider may later influence models used to profile 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. Despite these technical challenges, the normative claim that informational practices should remain proportionate and appropriate to their socio-functional aims remains central to legitimate informational practice. That organizations should collect only what is necessary for purpose fulfillment and use data only as declared, while ensuring compatibility with human dignity, anchors proportionality and necessity requirements of data protection law. Thus, privacy and AI governance frameworks continue to require context and purpose specification even as they shift from governing information flows to 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 mandating 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: prohibiting AI practices deemed to present an unacceptable risk to fundamental rights, health, or safety (Art. 5); requiring 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 obliging 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, already obscured by governance under-specification, are made further indeterminate by architectural opacity. Meeting this challenge requires a normative governance framework capable of substantively assessing dignity risks and adapting to evolving socio-technical contexts of use across the AI lifecycle.

Contextual Integrity offers a promising foundation. Contextual Integrity evaluates the appropriateness of information flows relative to social context, structuring 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]. Contextual Integrity traditionally treats informational purpose as an optional transmission principle constraining how information is shared—capturing purpose constraints implicitly by convention. Yet Contextual Integrity's reliance on norms strains its ability to address novel socio-technical contexts where norms remain unsettled. Because Contextual Integrity leaves purpose under-specified and does not articulate the considerations for dignity demanded by emerging regulation, its framework remains under-equipped for today's AI governance challenges.

We introduce CA-CI, which extends Contextual Integrity by (1) elevating purpose to a sixth constitutive parameter and (2) integrating dignity-based thresholds as a special class of fixed transmission principles. CA-CI operationalizes Nussbaum's Capabilities Approach (CA) to specify what dignity requires: an irreducible set of ten core capabilities that together constitute human 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.

By uniting the two normative theories, CA-CI provides a practical framework for evaluating AI systems

that respects diverse values within and across contexts while safeguarding human dignity as a universal baseline, thereby securing the integrity of both persons and social life. This article proceeds by first elaborating CA-CI's theoretical extensions: (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 for demonstration given its rights- and risk-based framework grounded in dignity, CA-CI is broadly applicable to normatively evaluate privacy and dignity in any socio-technical context, regardless of jurisdiction.

CA–CI as a Normative Dignity-based Governance Framework

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 or latent representational inferences that traverse AI systems—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 non-negotiable and the integrity of social life is preserved. The following sections explicate these theoretical extensions in further detail.

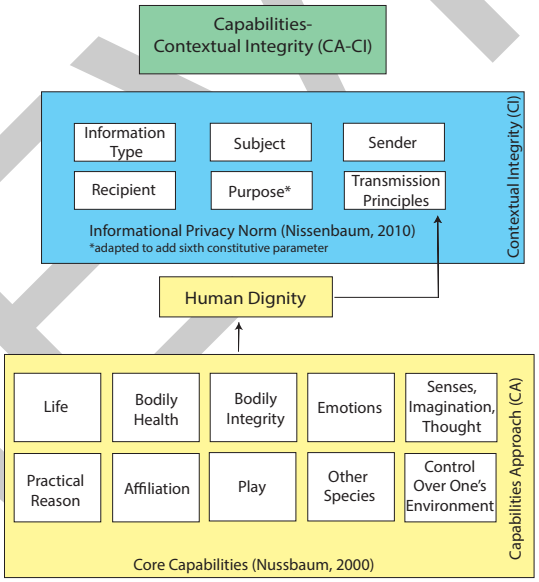

FIGURE 1. Capabilities–Contextual Integrity (CA–CI) Theoretical framework extending CI by (1) adding purpose as a constitutive parameter and (2) integrating dignity thresholds as a special class of fixed transmission principles.

Specifying 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, the core capabilities are the essential constituent parts of dignity as a whole; where 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 [12], and each of the human rights declared serving as derivative entitlements grounded in that worth [9]. 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

Dimension	Contextual Integrity (CI)	Capabilities Approach (CA)
Key Theoretical Focus	Appropriate data flows governed by context-relative norms and teloi.	Substantive capabilities for a dignified life.
Aim	Preserve the integrity of social life.	Preserve the dignity of persons.
Limitations and Gaps	Lacks an external evaluative standard to ensure internal accountability and safeguard against external imposition	Does not specify privacy explicitly, does not apply to contextual values or norms above minimum thresholds

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

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, 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.

Developed as a foundation for constitutional frameworks, Nussbaum's Capabilities Approach specifies the content of a dignified life: ten irreducible capabilities whose securement of every person's agency at threshold constitutes the moral floor of a *truly human* life—the evaluative boundary below which practices become degrading to humanity and thereby dignity-violating [12]. The Capabilities Approach's critical emphasis on *agency* ensures the framework 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 [11]. 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 normative grammars which sustain a domain's moral and structural coherence. 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 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 bound purpose—and their very 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 essential parameters: data subject, sender, recipient, attributes, and transmission principles. Yet these parameters alone cannot specify what kind of act a flow is when AI systems generate novel inferences. Consider employees using an internal AI assistant: conversational data logged for organizational security and policy compliance takes on fundamentally different meaning when repurposed to infer productivity or assess skills for performance reviews. It is purpose itself—why the information is inferred, what ends it serves—that makes the flow intelligible as one kind of practice rather than another. Inferences generated to evaluate policy compliance or organizational security are qualitatively distinct from those to assess employee skills or measure productivity, and these different purposes may demand distinct safeguards and normative justification.

Formalizing purpose as a sixth constitutive parameter therefore supplies a means to trace a flow to the telos of its context and assess compatibility. 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 extend authoritarian control over workers corrode that legitimacy [14]. When employers generate inferences about skills or productivity, we can evaluate whether those uses promote the context's telos, whether the practice is necessary to achieve it, and whether its means are compatible with human dignity. Purpose thus ensures that flows are evaluated not only by conformity to norms but by 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 are not confined to theory, but have practical value for AI governance. We demonstrate how CA-CI offers a systematic approach to meeting regulatory requirements under the EU AI Act by (1) enabling context-sensitive evaluations of dignity risks within Fundamental Rights Impact Assessments, (2) defining thresholds of significant harm, and (3) providing a model for anticipatory governance.

Risk Classifications and Fundamental Rights Impact Assessments

The EU AI Act bans certain AI applications that pose a clear threat to the rights, safety, and dignity of persons (Art. 5), while permitting high-risk systems (Art. 6) under strict safeguards (Art. 27) [7]. Yet providers and deployers face enduring ambiguities, both evaluative and operational. What distinguishes an unacceptable risk from a high-risk? How can impacts on abstract rights and freedoms be assessed within specific socio-technical contexts of use? Are individual rights frameworks adequate to capture systemic or collective harms?

By evaluating data flows and contextual parameters against capability thresholds, CA-CI offers a principled way to differentiate high-risk from unacceptable-risk systems. It makes visible when dignity is *at risk* and clarifies when dignity is *violated*. Each capability corresponds to one or more fundamental rights in the EU Charter, linking the core capabilities that constitute a dignified life to the formal entitlements protected in law. This mapping can support evaluators in assessing whether an AI system's socio-technical environment supports the realization of those rights in practice.

Table 2 illustrates this correspondence by mapping the ten central human capabilities to their associated Charter rights, showing how CA–CI operationalizes dignity as the evaluative bridge between abstract legal guarantees and context-specific socio-technical realities. This mapping also supplies a structured foundation for conducting Fundamental Rights Impact Assessments (FRIAs); when combined with Contextual Integrity's six analytic parameters, it can clarify which rights and capabilities are endangered by particular AI practices and where regulatory boundaries between acceptable and unacceptable risk should lie.

Significant Harm Thresholds

Where FRIAs evaluate risks to rights, the AI Act also requires evaluations of risks of harm. For instance, the

Capability (Nussbaum, 2011)	Corresponding EU Charter Rights and Freedoms
1. Life	Human Dignity (Art. 1); Right to Life (Art. 2); Integrity of the Person (Art. 3); Liberty and Security (Art. 6).
2. Bodily Health	Human Dignity (Art. 1); Liberty and Security (Art. 6); Healthcare (Art. 35); Fair and Just Working Conditions (Art. 31); Health care (Art. 35)
3. Bodily Integrity	Human Dignity (Art. 1); Integrity of the Person (Art. 3); Liberty and Security (Art. 6); Prohibition of Torture and Inhuman or Degrading Treatment (Art. 4).
4. Senses, Imagination, and Thought	Human Dignity (Art. 1); Freedom of Thought, Conscience and Religion (Art. 10); Freedom of Expression and Information (Art. 11); Freedom of the Arts and Sciences (Art. 13); Education (Art. 14).
5. Emotions	Human Dignity (Art. 1); Respect for Private and Family Life (Art. 7); Integrity of the Person (Art. 3); Freedom of Thought, Conscience and Religion (Art. 10); Fair and Just Working Conditions (Art. 31).
6. Practical Reason	Human Dignity (Art. 1); Freedom of Thought, Conscience and Religion (Art. 10); Freedom to Conduct a Business (Art. 16); Equality Before the Law (Art. 20); Non-Discrimination (Art. 21).
7. Affiliation	Human Dignity (Art. 1); Freedom to Choose an Occupation and Engage in Work (Art. 15); Equality Before the Law (Art. 20); Freedom of Expression and Information (Art. 11); Freedom of Assembly and Association (Art. 12); Non-Discrimination (Art. 21); Equality Between Women and Men (Art. 23); Workers' Right to Information and Consultation (Art. 27); Collective Bargaining and Action (Art. 28); Fair and Just Working Conditions (Art. 31).
8. Other Species	Human Dignity (Art. 1); Environmental Protection (Art. 37); Consumer Protection (Art. 38).
9. Play	Human Dignity (Art. 1); Education (Art. 14); Cultural, Religious and Linguistic Diversity (Art. 22).
10. Control Over One's Environment	Human Dignity (Art. 1); Respect for Private and Family Life (Art. 7); Protection of Personal Data (Art. 8); Freedom of Expression and Information (Art. 11); Freedom of Assembly and Association (Art. 12); Consumer Protection (Art. 38); Right to Property (Art. 17); Right to Vote and Stand for Election (Art. 39–40); Good Administration (Art. 41); Access to Documents (Art. 42); Freedom to Choose an Occupation and Engage in Work (Art. 15); Citizens' Rights (Title V).

TABLE 2. Mapping of Core Capabilities to Corresponding Fundamental Rights in the EU Charter.

Commission's draft guidelines clarify that prohibited or tightly controlled AI systems such as those distorting behavior through subliminal, deceptive, or manipulative techniques (Art. 5) must be assessed for both magnitude and likelihood of harm, including potential physical, psychological, financial, or economic harm, with particular 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 at what point 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 provides a framework that can resolve these ambiguities by using dignity thresholds to distinguish "significant harm": harm is significant where it drives any 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 shapes evaluation; and (3) how to track harms that compound over time.

Harm categories and thresholds.
Core capabilities supply both targets and thresholds for evaluating significant harm, and can be mapped to harms already recognized in privacy torts doctrine. Citron and Solove's taxonomy of cognizable privacy harms [15] largely overlaps with the categories identified in Commission guidelines [8], but core capabilities capture not only these legally recognized harms but also related risks of normative significance that may escape deductive harm categories.

Physical harms resulting in bodily injury or death threaten *bodily health*, *bodily integrity*, and *life*. Thresh-

olds are crossed when AI systems interfere with access or decision-making in matters of health, including reproductive health, bodily nourishment, and shelter; where they restrict free movement, undermine security against violent assaults, or choice in bodily matters (e.g., reproduction); or where they shorten life expectancy or otherwise reduce any capability conditions below what is compatible with dignity.

Reputational harms injuring one's standing in their community implicate *affiliation* and *control over one's environment*. Affiliation impacts become significant when they impede the ability to live with and toward others in mutual recognition, to secure the bases of self-respect and non-humiliation, or to be treated as a dignified being whose worth is equal to others. Thresholds concerning environmental control are crossed when effects impair one's capacity to participate equally in political choices, to seek and hold property and employment, or to maintain relationships of mutual recognition in the workplace. Effects may further cascade into degradations to other core capabilities such as *practical reason* and *emotions*.

Psychological harms inducing emotional distress or disturbance burden *emotions* alongside *practical reason*, and the *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 practical reason is impaired such that one cannot form a conception of the good or engage in critical reflection about life planning; or where cognitive or embodied capacities for imagination, thought, and creative or political expression are stifled. Downstream effects may further undermine *affiliation* and *control over one's environment*.

Economic harms resulting in monetary or opportunity loss strike *control over one's environment*. Thresholds are crossed when economic impacts prevent equal participation in political governance, property rights and ownership, or employment, or undermine freedom from unwarranted search and seizure or to enter relationships of mutual recognition in the workplace. These constraints may produce further cascading effects on core capabilities including *emotions*, *practical reason*, and *affiliation*.

Discrimination harms disadvantaging protected groups impair *affiliation*, *practical reason,* and *emotions*, and may extend to diminish other capabilities such as *control over one's environment*.

Relationship harms damaging personal, professional, or institutional relationships undermine *affiliation*, *practical reason*, and *emotions*, with context-relative impacts on capabilities such as *play* or *control over one's environment*.

Autonomy harms impair agency over both ends and means. They include coercion (limiting real choices); 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.

Context-relative evaluation.
The Commission treats harm significance as a fact-specific, case-by-case inquiry but offers limited guidance on how context should shape that assessment [8]. CA-CI provides a model to operationalize contextual analysis by linking CI's context-sensitive data flow diagnostics to concrete harm vectors, while clarifying the decisive condition: harm is significant where, in context, agency in any core capability is driven below threshold.

Consider the EU AI Act's ban on emotion recognition in the workplace, generally prohibited for its incompatibility with dignity, with narrow exceptions for medical or safety purposes (Art. 5) [7]. These high-risk exceptions require FRIAs to ensure applications remain justified; CA-CI specifies how. 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. By these parameters, the evaluation connects system behaviors and design to contextual capability impacts.

This supports both FRIA evaluation and targeted risk mitigation. For instance, if fatigue detection flags an employee at safety risk and alerts supervisors, HR, and operational leaders, traditional privacy risk models would assess whether access is appropriately limited and data is securely stored. But even where technical safeguards restrict access to authorized actors, CA-CI reveals dignity risks that flow-level analyses alone may miss: Could supervisors use fatigue data to question worker commitment or reliability, eroding *affiliation* by undermining mutual recognition in the workplace? Could HR link fatigue patterns to performance evaluations, instrumentalizing worker vulnerability and impairing their ability to *practically reason* about critical work-life decisions? Could the very knowledge of fatigue

monitoring chill association among colleagues, corroding dignity through pressures of non-humiliation?

By posing such questions, CA-CI identifies where safeguards are insufficient and where additional protections are needed. For instance, even if necessity, minimization, and anti-discrimination measures are in place, dignity thresholds may still be at risk. CA-CI highlights complementary safeguards such as prohibiting integration of fatigue-monitoring data into performance management systems, embedding these constraints into data lineage and access control systems, and instituting organizational policies, supervisor training, and audits to ensure compliance.

In this way, CA-CI exposes dignity risks that extend beyond technical data architecture and clarifies which safeguards—technical, organizational, normative—are necessary to mitigate them. It thereby enables context-sensitive evaluation while maintaining consistent dignity standards.

Harmful effects over time.
Commission guidelines specify that harm assessments must consider effects which 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 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 accumulating across time—reputational degradation, cumulative autonomy erosion, economic precarity. Evaluators can track whether capability impacts compound into significant harm, determining when graduate injuries that are individually minor but cumulatively dignity-eroding 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 CI parameters specify context, roles, attributes, transmission principles, and purpose, while CA 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 Act already distinguishes risks to dignity and derivative rights from permissible uses by reference to contextual parameters that map to CI's framework. For instance, the 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 evaluations of impacts to dignity. This model enables evaluators to identify dignity risks that current regulatory categories miss. 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 can be overlooked by purely procedural evaluations.

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, incorporating dignity thresholds from the Capabilities Approach and purpose as a constitutive parameter, strengthening its normative and empirical adequacy for AI governance. First, it enables the evaluation of novel AI systems that generate information whose social meaning has not yet been negotiated, and may be applied across contexts where norms are absent or ambiguous, by anchoring assessment in dignity thresholds from the Capabilities Approach. Practically, these extensions support providers, deploy-

ers, and regulators in evaluating privacy and dignity risks in AI systems.

While this paper establishes CA-CI's theoretical foundations and demonstrates its governance utility via applications to the EU AI Act, further work 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. Capabilities-based instruments already validated in fields such as health, wellbeing, and human development can be employed and adapted to measure the impact of AI systems on core capabilities in particular contexts, enabling 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 confirmed through such empirical research that capability-based evaluations reliably identify dignity violations across sociotechnical contexts, CA–CI could offer a stable evaluative architecture for dignity-preserving AI governance, one that promotes pluralism while grounding accountability in the real capabilities individuals and communities have to live lives they have reason to value.

REFERENCES

1. C. Song and A. Raghunathan, "Information Leakage in Embedding Models," in Proc. 2020 ACM SIGSAC Conf. Comput. Commun. Security (CCS '20), New York, NY, USA: ACM, 2020, pp. 377–390, doi: 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," AIES, vol. 8, no. 2, pp. 1862–1874, Oct. 2025.
3. A. Reuel et al., "Open Problems in Technical AI Governance," Trans. Mach. Learn. Res., 2025. [Online]. Available: https://openreview.net/forum?id=1nO4qFMiS0
4. K. S. Kumar et al., "Security and Privacy Challenges in Multi-Tenant Cloud Architectures: A Comprehensive Analysis," in Proc. 2025 Int. Conf. Comput. Technol. Data Commun. (ICCTDC), IEEE, 2025.
5. I. Barberá, "AI Privacy Risks & Mitigations — Large Language Models," Support Pool of Experts Programme, European Data Protection Board, 2025. [Online]. Available: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf
6. European Parliament and Council of the 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 (General Data Protection Regulation)," Off. J. Eur. Union, Apr. 27, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
7. European Parliament and Council of the 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 (Artificial Intelligence Act)," Off. J. Eur. Union, Jul. 12, 2024. [Online]. Available: https://data.europa.eu/eli/reg/2024/1689/oj
8. European Commission, "Annex to the Communication to the Commission — Approval of the Content of the Draft Communication — Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)," Brussels, Belgium, C(2025) 884 final, Feb. 4, 2025. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/112367
9. United Nations General Assembly, "Universal Declaration of Human Rights," General Assembly Resolution 217 A (III), Dec. 10, 1948. [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights
10. European Union, "Charter of Fundamental Rights of the European Union," Off. J. Eur. Communities, vol. C 364, pp. 1–22, Dec. 18, 2000. [Online]. 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, Pol-

icy, and the Integrity of Social Life. Stanford, CA, USA: Stanford Univ. Press, 2009.
12. M. C. Nussbaum, Women and Human Development: The Capabilities Approach. Cambridge, U.K.: Cambridge Univ. Press, 2000.
13. M. Walzer, Thick and Thin: Moral Argument at Home and Abroad. Notre Dame, IN, USA: Univ. Notre Dame Press, 1994.
14. E. Anderson, Hijacked: How Neoliberalism Turned the Work Ethic Against Workers and How Workers Can Take It Back. Cambridge, U.K.: Cambridge Univ. Press, 2023.
15. D. K. Citron and D. J. Solove, "Privacy Harms," Boston Univ. Law Rev., vol. 102, no. 3, pp. 793–863, Apr. 2022.

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. 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.