

AI and digital ethics in the age of generative systems-principles, standards and accountability across cultures

Ognen Firfov

University American College Skopje, North Macedonia; e-mail: ognen.firfov@uacs.edu.mk

Abstract: *Generative AI intensifies ethical risks around opacity, bias, responsibility gaps, and cross-cultural legitimacy. This paper synthesizes contemporary literature and proposes a layered governance model with three layers that translates universal ethical principles into culturally adaptive and sector-specific controls. The contribution is a practical pathway from principles to standards, enabling transparency, accountability, and contestability across the AI lifecycle.*

KEY WORDS: GENERATIVE AI, DIGITAL ETHICS, TRANSPARENCY, ACCOUNTABILITY, STANDARDS, GOVERNANCE

1. Introduction

Generative artificial intelligence (AI) systems are more and more included in important areas with possible high impact such as public administration, telecommunications, healthcare, education, and finance. Their outputs which are probability based, scale of coverage, and limited interpretability introduce ethical challenges that existing governance regimes struggle to manage [1,2]. A central problem is in fact operational: how to move from widely recommended ethical principles to implementable operational standards that remain legitimate across different cultures and various contexts.

2. Prerequisites and means for solving the problem

Ethical AI governance draws on wide known recurring principles — transparency, accountability, fairness, privacy, and human autonomy — yet generative AI complicates implementation because harms can be diffuse, emergent, and value-laden [4,3]. Further in this section, the state of the art is reviewed and gaps that motivate the proposed governance model are identified.

2.1 Literature review: principles, practice, and socio-technical governance

Early AI ethics work emphasized that ethical quality is not only a property of algorithms but of socio-technical systems — institutions, incentives, and accountability processes surrounding them [4]. Floridi et al. propose a principled foundation [beneficence, non-maleficence, autonomy, justice, explicability] that has influenced subsequent debates, while also highlighting explicability as a distinct requirement for AI-mediated societies [1].

However, principle proliferation has raised concerns about ‘ethics washing’ and non-actionability, things would be done on paper and formally but then no effect could be seen later. Mittelstadt argues that principles are too indeterminate without enforcement, metrics, and institutional mechanisms [2]. This critique is reinforced by the observation that different stakeholders require different forms of transparency and justification, making single, universal transparency mandates inadequate [5].

Opacity has been conceptualized and demonstrated in multiple ways. Burrell distinguishes intentional secrecy, technical complexity, and inherent model inscrutability, noting that modern machine learning can be opaque even absent malintent [6]. Interpretable machine learning research frames explainability as a scientific challenge and calls for evaluation criteria to avoid misleading explanations [7]. In governance contexts, the emphasis

shifts from full interpretability to auditability and procedural transparency that can support oversight and contestability [5].

Accountability research similarly stresses socio-technical embedding. Binns connects algorithmic fairness to political philosophy and argues that normative commitments (e.g., equality) require explicit choices and trade-offs rather than purely technical fixes [8]. Selbst et al. warn that abstraction can sever fairness methods from social context, producing ‘false precision’ and governance blind spots [9].

Operational accountability mechanisms are more and more intensively discussed in the context of audits and documentation related to AI related ethics. Raji et al. propose structured auditing processes to close the ‘accountability gap’ between stated principles and deployed systems, emphasizing organizational controls and independent review [10]. Yeung’s concept of algorithmic regulation highlights the institutional design challenges of delegating governance to computational systems and the need for oversight that preserves democratic accountability [11].

Cultural diversity complicates any move toward universal standards which can be applied across the board, also due to the reason that this diversity is multilayered and different in different societies or groups. Jobin et al. find convergence at high-level values across guidelines but significant divergence in interpretation, scope, and enforcement proposals [3]. Ess argues that digital media ethics must take plural moral traditions seriously, not merely as ‘local noise’ but as sources of legitimacy and societal stability [12]. Hagerty further cautions that global AI ethics can reproduce power asymmetries unless standard-setting includes diverse voices and socio-economic realities [13].

Together, it is quite obvious from the mentioned above that this literature indicates a persistent gap: [i] principles are widely accepted, [ii] implementation mechanisms remain uneven, and [iii] universalization efforts risk cultural illegitimacy. The next section responds by proposing a layered governance model that connects ethical baselines to context-sensitive standards and sector controls.

2.2 Gap statement and design requirements

From the reviewed literature, the governance solution should satisfy four requirements: [R1] enable accountability across the AI lifecycle (design–deployment–monitoring), [R2] provide procedural transparency sufficient for audit and contestability, [R3] maintain interoperability through a minimal universal baseline, and [R4] allow cultural and sectoral adaptation without ethical relativism [2,10,12].

3. Solution of the examined problem

To meet these requirements, we propose a layered governance model for generative AI that translates universal principles into implementable standards and controls. The model is designed to bridge the ‘principles-to-practice’ gap identified in the literature [2].

The model consists of three layers: [L1] a universal ethical baseline (e.g., human dignity, fairness, accountability, safety); [L2] cultural and regional interpretation [legal traditions, ethical norms, risk tolerances, and institutional capacity]; and [L3] sector-specific operationalization (policies, technical controls, documentation, audits, and redress). Figure 1 presents the already mentioned model.

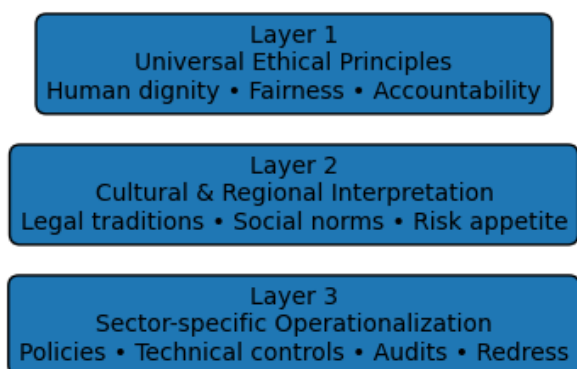


fig. 1. Layered governance model for ethical generative AI systems

In practice, L1 functions as a minimum normative floor (the basis) that prevents harmful ‘race-to-the-bottom’ dynamics, L2 provides legitimacy and interpretive alignment across jurisdictions, and L3 delivers measurable controls aligned with concrete risks (e.g., discrimination, privacy violations, misinformation). This structure supports procedural transparency and auditable accountability without assuming full model explainability [6,5].

4. Results and discussion

Because this study is a conceptual governance analysis, the ‘results’ are presented as analytical findings derived from synthesis of the literature and design requirements. The core proposed outcome is that a layered approach presented on Figure 1, can operationalize ethics while still keeping alive cultural legitimacy and interoperability.

4.1 Interpretation: transparency as procedural, not total disclosure

A key implication from opacity research is that transparency should be specified by purpose and audience. Rather than total disclosure of model internals in all cases, governance should require procedural transparency: documentation of intended use, training data provenance at an appropriate granularity, known limitations, and evaluation protocols. This aligns with critiques of the ‘transparency ideal’ and supports contestability and oversight [5,7].

Within the layered model, transparency obligations are distributed: L1 defines the transparency minimum (e.g., explainability proportional to risk), L2 maps this to local regulatory and cultural expectations, and L3 prescribes specific artifacts (e.g., model cards, audit logs, incident reporting). The result is a feasible pathway to oversight even for complex models where interpretability remains limited [6].

4.2 Interpretation: accountability as lifecycle and institutional control

The accountability gap is appearing when responsibility is diffused across creators, developers, deployers, and downstream integrators. Prior work emphasizes organizational governance mechanisms such as independent audits and structured risk assessments [10]. In the layered model, L3 enforces lifecycle controls, including pre-deployment impact assessments, ongoing monitoring, and corrective actions when harms occur. This institutional framing aligns with socio-technical perspectives on machine behavior and governance [14,11].

A practical implication is that accountability should be clearly documented, visible and assignable to a person, position or unit: who(or which unit) approves deployment, who monitors performance drift, and who is responsible for user-facing remedies. These controls support traceability and provide the basis for redress (remedy to set right), an often under-specified element in principle-only frameworks [2].

4.3 Cultural diversity: avoiding both relativism and ethical imperialism

The literature shows that universal principles can be agreed at a high level, yet their operational meaning differs across cultures [3]. L2 explicitly accommodates pluralism by allowing interpretive translation (e.g., how autonomy, consent, and harm are socially understood) while preserving the normative floor defined by L1. This approach aims to avoid ethical imperialism while still preventing relativistic ‘anything goes’ deployment [12,13].

4.4 Implications and future research directions

Future research should test and refine the layered model via empirical studies: [i] comparative case analyses of generative AI governance across jurisdictions, [ii] evaluation of auditing artifacts and their effectiveness in detecting harm, and [iii] development of metrics for ‘procedural transparency’ and ‘contestability’ that can be standardized without forcing uniform moral assumptions. Work is also needed on governance for model supply chains (foundation models, fine-tuning, and downstream deployment) and on culturally legitimate stakeholder participation in standard-setting [10,11].

5. Conclusion

This paper argues that ethical governance of generative AI requires a shift from abstract principles to implementable standards embedded in institutional processes. By integrating universal ethical baselines with culturally adaptive interpretation and sector-specific controls as shown on Figure 1, the layered governance model offers a feasible and desirable direction toward transparency, accountability, and legitimacy.

5.1 Tangible hypotheses

H1: Generative AI deployments that implement lifecycle auditing and documentation [L3] will show lower rates of recurring ethical incidents than deployments relying only on principle statements.

H2: Governance regimes that explicitly include cultural/regional interpretation [L2] will have higher perceived legitimacy among affected stakeholders than regimes that apply uniform standards without adaptation.

H3: A minimal universal baseline [L1] combined with sector controls [L3] will improve cross-border interoperability without increasing ethical risks, compared to purely localized governance.

H4: Procedural transparency artifacts (documentation + audit logs) will enable more effective redress and contestation than technical 'black-box' explanations alone.

These hypotheses can be tested through mixed-method research combining incident datasets, operational audits, stakeholder surveys, and comparative regulatory analysis. If supported, they would substantiate the model's practical impact and provide evidence-based foundations for universal standards that remain culturally legitimate.

6. Synthesis and practical applications

Reviewing the study's insights and view, the layered governance model can be applied as a design template for organizations and regulators. At L1, governmental institutions and commercial entities should define a non-negotiable baseline: commitments to non-discrimination, safety, accountability, and meaningful transparency proportional to risk which is present. At L2, these commitments should be converted into locally legitimate requirements (e.g., consent expectations, data protection norms, acceptable uses). At L3, organizations should implement operational controls: documentation of the model, assessment impact before deployment, monitoring for drift and typical misuse, incident response plans, and user-facing remedy directions.

For high-impact areas (e.g., critical infrastructure, high hazard applications and telecommunications), practical controls include access controls for model outputs, robust logging and hazard logging and retention policies, third-party auditing, and governance committees with cross-functional representation. When it comes to public generative AI systems, practical protection points should contain public disclosure of AI-generated content, abuse monitoring, and typical escalation procedures. The broader impact is a governance architecture that can quite easily scale up and grow with innovation while maintaining ethical accountability and societal trust.

7. References

1. Floridi L. AI4People – An ethical framework for a good AI society – *Minds and Machines*, 28, 2018, 689–707.
2. Mittelstadt B. Principles alone cannot guarantee ethical AI – *Nature Machine Intelligence*, 1, 2019, 501–507.
3. Jobin A. The global landscape of AI ethics guidelines – *Nature Machine Intelligence*, 1, 2019, 389–399.
4. Mittelstadt B. The ethics of algorithms: Mapping the debate – *Big Data & Society*, 3, 2016.
5. Ananny M. Seeing without knowing: Limitations of the transparency ideal – *New Media & Society*, 20, 2018, 973–989.
6. Burrell J. How the machine thinks: Understanding opacity in machine learning algorithms – *Big Data & Society*, 3, 2016.
7. Doshi-Velez F. Towards a rigorous science of interpretable machine learning – arXiv, 2017.
8. Binns R. Fairness in machine learning: Lessons from political philosophy – *Proceedings of FAT* (FAccT)*, 2018, 149–159.
9. Selbst A. Fairness and abstraction in sociotechnical systems, in: *Proceedings of FAT* (FAccT)*, 2019, 59–68.
10. Raji I. Closing the AI accountability gap, in: *Proceedings of FAT* (FAccT)*, 2020.
11. Yeung K. Algorithmic regulation: A critical interrogation – *Regulation & Governance*, 12, 2018, 505–523.
12. Ess C. *Digital Media Ethics*, Cambridge, Polity Press, 2020.
13. Hagerty A. Global AI ethics: A review of the social impacts of artificial intelligence – *AI & Society*, 34, 2019, 497–512.
14. Rahwan I. Machine behaviour – *Nature*, 568, 2019, 477–486.