

The Reversal of Visibility: Transparency, AI Governance, and the Ethics of Hiddenness

JONAS MIKLAVČIČ

Faculty of Theology, University of Ljubljana

jonas.miklavcic@teof.uni-lj.si

ORCID: 0000-0003-2393-348X

Abstract. Transparency is repeatedly invoked as a foundational requirement for ‘trustworthy’ artificial intelligence (AI), yet the term is often treated as self-evident. This article argues that transparency is conceptually plural, ethically ambivalent, and—under contemporary digital conditions—capable of producing a reversal of visibility: persons become increasingly exposed and predictable, while the infrastructures that classify and steer them remain partially opaque. Using conceptual analysis in dialogue with empirical research in critical transparency studies and AI ethics, the paper (1) distinguishes key meanings of transparency (notice, explainability, auditability, and accountability), (2) shows how datafied environments generate epistemic asymmetries and diffuse responsibility, and (3) interprets these dynamics through a science-and-faith lens by contrasting ‘omniscience without grace’—knowledge detached from relational responsibility—with theological accounts of hiddenness, dignity, and moral agency. Against both naïve demands for total openness and cynical acceptance of black-box governance, the paper proposes a norm of proportional transparency: transparency where it protects persons and enables contestation, and justified opacity where it safeguards intimacy, conscience, and human flourishing.

Keywords: AI ethics; transparency; hiddenness; omniscience; surveillance; human dignity; theology of technology.

Contribution. The article advances AI-governance debates by identifying and theorizing an asymmetry of visibility that standard transparency frameworks fail to address. It develops a normative proposal (“proportional transparency”) that specifies when and for whom transparency is ethically warranted, strengthening accountability while limiting harmful exposure of individuals. The author designed the study, performed the analysis and synthesis of the literature, and drafted and revised the manuscript.

Use of AI. There was no use of generative AI in the process of writing this paper.

Funding. This research was funded by research program: *The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theologica Insights* (University of Ljubljana).

Introduction

Calls for “transparency” have become a near-unquestioned moral ideal in contemporary AI governance. Major policy frameworks—from the European Union’s Artificial Intelligence Act to standards work such as the NIST AI Risk Management Framework—treat transparency as a central safeguard for trustworthy AI (European Parliament and Council of the European Union 2024; NIST 2023). Likewise, global ethics instruments emphasize disclosure and explainability as prerequisites for responsible deployment (OECD 2019; UNESCO 2021). At first glance, the normative direction seems obvious: if AI systems influence hiring, credit, welfare allocation, policing, or information ecosystems, then affected persons deserve to know when and how such systems operate.

Yet transparency is not a single practice. Even in technical and legal discussions it functions as an umbrella term covering different, sometimes conflicting, expectations: (a) *notice* (being informed that an AI system is used), (b) *traceability and documentation*, (c) *explainability* of the decision-making process, (d) *auditability* for independent oversight, and (e) *accountability* mechanisms that enable contestation and redress (Ananny and Crawford 2018; Burrell 2016; European Commission High-Level Expert Group on AI 2019). Moreover, the social demand for transparency unfolds within a digital environment that is itself asymmetrical:

individuals are asked to become “transparent” through data disclosure, while the infrastructures that observe, infer, and monetize that data often remain opaque and hidden (Zuboff 2019; Pasquale 2015).

To avoid equivocation, I use “transparency” in two primary senses throughout: (1) institutional or system transparency—practices that make AI systems and their deployers answerable through notice, explanation, documentation, audit, and redress; and (2) personal transparency—the growing exposure and predictability of persons produced by pervasive data capture and inference. The paradox explored here concerns their interaction: calls for the former can coexist with, and sometimes legitimate, the latter.

This tension motivates the paper’s central claim: contemporary transparency discourse risks masking a paradoxical reversal of visibility. While transparency is invoked to illuminate AI systems, datafied governance can produce increasing legibility of persons—through profiling, prediction, and behavioral steering—combined with persistent opacity of institutional decision-making. The result is an ethical shift from transparency as a condition of public reason to transparency as a tool of control.

The argument is developed not only as a philosophical critique but also as a contribution to science-and-faith dialogue. The paper asks how theological accounts of hiddenness, dignity, and moral agency can clarify what is ethically at stake when “transparency” becomes a totalizing ideal (Turner 1995; Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education 2025). The guiding research questions are: (1) Which meanings of transparency operate in AI ethics and regulation? (2) How does the digital environment transform transparency into an asymmetrical regime of visibility? (3) What normative criteria—also informed by theological anthropology—can distinguish liberating transparency from reductive exposure?

Methodologically, the paper combines (i) conceptual analysis of transparency, visibility, and opacity, (ii) engagement with interdisciplinary research in AI ethics and critical transparency studies, and (iii) a theological-ethical interpretation of hiddenness and omniscience. The goal is not to research a technical side of a model explanation, but to clarify

the moral grammar that structures current debates on transparency and to propose normative criteria that could in turn influence AI governance.

The paper proceeds in four steps. Section 1 situates the contemporary transparency imperative and clarifies the plural meanings of transparency in AI governance. Section 2 looks at transparency as an epistemic and political ideal and shows how digital infrastructures transform it into a pervasive demand for exposure. Section 3 analyzes the resulting “reversal of visibility” and illustrates it with empirical and regulatory examples. Sections 4–6 then draw out the governance and anthropological implications, culminating in a science-and-faith interpretation (Section 5) and the proposal of proportional transparency (Section 6) as an ethic for human-centric AI.

1. The Transparency Imperative in AI Governance

In the last decade, “transparency” has become a policy keyword across AI governance. In the EU AI Act, transparency is treated both as a general aspiration for trustworthy AI and as a set of specific obligations for certain systems. Article 50 requires, among other things, that people be informed when they are interacting with an AI system, that synthetic content be machine-readably marked, and that deployers disclose the use of emotion recognition or biometric categorisation systems (European Parliament and Council of the European Union 2024, art. 50).

Earlier European guidance framed transparency as part of a broader package of ‘trustworthy AI’ requirements. The Ethics Guidelines for Trustworthy AI (HLEG) explicitly connect transparency to human agency, explainability, and traceability, and they recommend documentation and auditability as governance practices (European Commission High-Level Expert Group on AI 2019). Similarly, the OECD recommendation urges transparency and responsible disclosure, and UNESCO’s recommendation highlights transparency as a condition for accountability and human rights protection (OECD 2019; UNESCO 2021). Outside Europe, NIST’s AI Risk Management Framework treats transparency and interpretability as

characteristics that support risk identification, accountability, and public trust (NIST 2023).

These documents show that transparency is not merely a moral slogan; it is operationalized through concrete governance instruments: notices, documentation, labeling, impact assessment, and audit. At the same time, the documents also reveal internal tensions. First, transparency is frequently invoked as a means of *trust*, but trust is not produced by exposure alone; it requires competence, integrity, and effective accountability. Second, transparency is often framed as a remedy for “black boxes,” yet some systems remain technically hard to explain, while others are legally or commercially protected as intellectual property (Burrell 2016; Rudin 2019). Third, transparency can conflict with other goods and values—privacy, security, and the protection of vulnerable persons. As a result, practical governance requires criteria for *how much* transparency, *for whom*, and *for what purpose*.

To clarify the concept of transparency, we can distinguish five dimensions that are often collapsed in policy debates:

- Notice and intelligibility: timely, accessible information that AI is being used and what role it plays in a decision (European Parliament and Council of the European Union 2024, art. 50).
- Explainability: reasons or understandable accounts of how outputs were produced; this may range from local explanations of a single decision to global model descriptions (Doshi-Velez and Kim 2017).
- Traceability and documentation: records that allow reconstruction of design choices, training data provenance, and deployment context (European Commission High-Level Expert Group on AI 2019).
- Auditability: the possibility of independent evaluation of performance, bias, and compliance through testing, access, or mandated disclosures (Ananny and Crawford 2018; NIST 2023).
- Accountability and contestability: procedures for responsibility assignment, appeals, and redress—so that transparency can actually change outcomes rather than merely inform (OECD 2019; UNESCO 2021).

The core thesis of this paper is that these dimensions do not automatically align. A system can provide ‘notice’ without enabling

contestation; it can be explainable to engineers but unintelligible to affected persons; it can be auditable by regulators while remaining experientially opaque to ordinary users. This mismatch opens the door to a cultural reversal: the *appearance* of transparency coexists with growing informational power asymmetries that we will discuss later in the paper.

2. The Concept of Transparency: From Illumination to Digital Exposure

Although transparency can refer to a variety of practices and sometimes its meaning in a given context is not at all obvious, as we saw above, it is difficult to dispute transparency as an ethical norm in almost all situations. To understand why transparency is treated as self-evidently good, it helps to note its historical layering. Modernity often associates truth with illumination: what is rational is what can be publicly articulated, inspected, and criticized. In political contexts, transparency functions as a safeguard against arbitrary power—an extension of ideals of publicity and accountable governance. (Alloa 2018) But late modernity also turns transparency into a cultural demand for exposure, where *visibility* itself becomes a value.

Contemporary diagnoses of a ‘transparent society’ capture this shift. Dennett and Roy argue that digital infrastructures have produced a new environment in which secrecy becomes difficult and institutions must adapt to radical visibility (Dennett and Roy 2015). Byung-Chul Han, by contrast, warns that the ‘transparency society’ transforms trust into control and turns subjects into self-exposing data producers (Han 2015). Critical transparency studies emphasize that transparency is never simply the absence of concealment; it is a selective practice that redistributes power by deciding what becomes visible, to whom, and under which interpretive frames (Hansen, Christensen, and Flyverbom 2015; Alloa 2018).

This matters because it clarifies an ethical ambiguity: transparency can serve emancipation when it exposes domination, but it can also become domination when it demands total visibility. In other words,

transparency is not only an epistemic tool but also a political technology and as such cannot necessarily be self-evidently good. The question is therefore not whether transparency is ‘good’ in the abstract, but whether specific transparency practices enlarge or diminish human agency. The point is not abstract. In the digital sphere, transparency is operationalized through infrastructures that continuously *produce* visibility, often without deliberate human choice. This is where the ethical question sharpens: whose visibility is increased, by what mechanisms, and to whose advantage?

For example, digital systems generate pervasive traces—so called digital footprints. Search queries, location histories, social media interactions, and device sensor data are routinely collected and combined into profiles that can be used for prediction and influence. The central point is not merely that information is recorded, but that recording is automated, cheap, and scalable. As Zuboff argues, data extraction has become an economic logic in which personal experience is translated into behavioral data for prediction and control (Zuboff 2019). In such environments, transparency is frequently demanded from individuals—through consent interfaces, data sharing, and identity verification—while institutional operations remain shielded by complexity and proprietary claims. This is personal transparency: individuals become legible by default, while institutional decision infrastructures remain partially opaque.

Public debates about state surveillance made this asymmetry visible. The 2013 Edward Snowden disclosures about NSA surveillance of communications—including surveillance that reached European leaders—and subsequent analyses highlighted the intertwining of large-scale data infrastructures, security rationales, and platform intermediaries (Lyon 2015). Even when surveillance is justified in the language of safety or efficiency, it forms a background condition in which ‘visibility’ becomes default and ‘hiddenness’ appears suspicious. Furthermore, empirical research shows that datafied systems can produce discriminatory or manipulative outcomes even when designers intend neutrality. For example, a field study on online ad delivery found that a gender-neutral STEM job advertisement was shown disproportionately

to men, illustrating how platform optimization can create disparate exposure without explicit discriminatory intent (Lambrecht and Tucker 2019). Such findings matter for transparency: the more decisions are shaped by complex intermediaries, the harder it becomes to identify who is responsible, what ‘the decision’ even is, and which explanation would be meaningful to affected persons.

Herein lies the transparency paradox. Alloa describes transparency as a ‘keyword’ whose promise of clarity often hides the selective and strategic character of making-visible (Alloa 2018). Ida Koivisto develops this same point by arguing that transparency demands can simultaneously illuminate and obscure, because they may overwhelm users with information while leaving the decisive mechanisms untouched (Koivisto 2022; Koivisto 2016). In simple terms, to make something really transparent and visible, something else often has to be concealed. The paradox is therefore not simply that transparency is difficult to achieve; it is that transparency can function as a ritual of legitimacy that conceals deeper asymmetries. To explain and develop this insight further, the next section therefore looks at how AI systems participate in this reversal of visibility.

3. The Reversal of Visibility: When Transparency Becomes Asymmetry

There seems to be a shift in what is *exposed* and what is *concealed* in modern technics. In many earlier technologies, the artefact was publicly visible and intelligible (cars, airplanes, telephones, even much of the early computers and web), while the user largely remained socially and informationally “hidden” behind it (“hidden behind keyboard”). With contemporary AI-driven digital infrastructures, the situation increasingly flips: the system’s operative logic becomes opaque (complex, proprietary, distributed), while persons become continuously legible through their data traces. The result is an asymmetry of visibility in which individuals are transparent by default and institutions are not – this is what I call the reversal of visibility.

It is a question of control as well. The modern ideal of transparency was historically directed upward: citizens demanded that rulers and institutions justify decisions, publish reasons, and expose corruption. In digital environments, however, transparency often flows downward: people and their behaviors are made legible to institutions and platforms, while the decision infrastructures remain partially opaque. Put differently: in this reversal personal transparency expands, while institutional transparency remains partial at best. A classical metaphor for asymmetrical visibility is the panopticon. Foucault's analysis of this Bentham's idea on panoptic power describes a regime in which the many are potentially visible to the few, internalizing surveillance and disciplining themselves (Foucault 1995). Contemporary datafied systems share the core panopticon structure: individuals are rendered observable, classifiable, and predictable through continuous data capture (Zuboff 2019; Lyon 2015).

AI intensifies this asymmetry because it does not merely record behavior; it infers traits and probabilities. Profiling systems reduce persons to clusters, risk scores, or predicted preferences—operational representations that can be acted upon even when they are incomplete or erroneous (Eubanks 2018; Barocas and Selbst 2016). The ethical stakes are concrete: when a person is treated as a risk category or as a marketing segment, their autonomy is constrained by decisions that are both consequential and difficult to contest.

The other important aspect of combining AI and reduced transparency is that AI itself is non-transparent or, as we sometimes say, inexplicable. Transparency initiatives often assume that if we can “explain” an AI model, accountability will follow. But AI opacity, that makes us not understand the reasons for an AI decision, has multiple sources. Burrell distinguishes opacity arising from intentional secrecy (e.g., proprietary systems), technical illiteracy (users lacking the background to interpret outputs), and the intrinsic complexity of some machine-learning models (Burrell 2016). Ananny and Crawford add that ‘opening the black box’ is rarely a single act; it requires institutional arrangements, interpretive labor, and the politics of who gets access and for what purpose (Ananny and Crawford 2018). This means that transparency-as-explanation is

performative. Platforms may provide interface disclosures, or publish general principles, without revealing the specific data, model features, or optimization objectives that shape outcomes. Similarly, organizations may rely on the sheer complexity of socio-technical systems to diffuse responsibility: “the algorithm decided,” “the model recommended,” “the system flagged.” Such language displaces accountability from persons and institutions to an apparently neutral technical artifact (Pasquale 2015).

But the problem is not only epistemic but moral as well. If transparency is treated as merely informational—something that can be satisfied by disclosing a fact—then it may fail to secure the deeper goals that motivate transparency demands: fair treatment, the possibility of contestation, and the protection of dignity.

Empirical research and public policy debates provide concrete illustrations of the reversal of visibility. First, automated decision systems in public services can render vulnerable populations intensely visible to bureaucratic scoring, while making institutional decision criteria harder to contest. Eubanks documents cases in which welfare and social service systems use data-driven tools to profile and ration access, often deepening existing inequalities (Eubanks 2018) without being explicitly visible themselves.

Second, targeted advertising and recommender systems can shape public discourse while remaining opaque in their steering logic. Even when ad targeting is framed as personalization, it depends on classification and prediction about persons; it can also amplify structural bias, as in the disproportionate delivery of STEM ads to men (Lambrecht and Tucker 2019). Third, risk scoring in criminal justice has raised questions about whether explanations are sufficient when the underlying data encode historical injustices. Barocas and Selbst show how ‘big data’ can reproduce disparate impact through proxies and feedback loops even without explicit protected-class features (Barocas and Selbst 2016).

These examples show why the transparency debate cannot be reduced to a technical problem of interpretable models. What is at stake is a social order in which persons are made transparent as data objects, while institutions become opaque as complex, distributed, and proprietary

systems. Every time we find ourselves in an online environment (and today we practically live in it), every move we make makes us more visible to algorithm designers, without us knowing that they are following us – without AI revealing itself, in other words. Invisible technology makes us visible, and therein lies the reversal.

4. Invisible AI: Explainability, Audit, and Accountability

If Section 3 described the reversal of visibility, this section asks what it takes to respond: why focusing only on model explanations is insufficient, and how documentation, audit, and contestability can restore institutional transparency. The notion of ‘invisible AI’ captures two related phenomena. First, AI is increasingly embedded in everyday infrastructures—search, news feeds, navigation, credit and insurance, hiring platforms—so that decision mediation becomes ordinary and unremarkable. Second, many AI systems operate in ways that are not directly observable by those affected: model inference happens at scale, often in the background, and is protected by proprietary or security constraints. Invisibility is therefore both experiential (users do not perceive the system’s operation) and structural (access is limited).

From a governance perspective, three distinct challenges emerge: (1) *technical opacity* (how the model works), (2) *data opacity* (what data and proxies drive outcomes), and (3) *institutional opacity* (who is responsible, and what remedies exist). AI ethics discussions often focus on (1), but the reversal of visibility is mainly driven by (2) and (3) (Ananny and Crawford 2018; Pasquale 2015).

Explainable AI research helps but does not solve these problems. Doshi-Velez and Kim argue for a more rigorous science of interpretability, including clarity about the intended audience and the purpose of explanations (Doshi-Velez and Kim 2017). Rudin goes even further by warning against the routine use of *post-hoc* explanations for black-box models in high-stakes domains, arguing that interpretable models can often achieve comparable performance and are preferable when decisions must be justified (Rudin 2019).

Regulation increasingly recognizes these distinctions. The EU AI Act mixes transparency obligations (Article 50) with broader requirements for high-risk systems—risk management, documentation, and human oversight—thereby treating transparency as part of an accountability ecosystem rather than as an isolated disclosure (European Parliament and Council of the European Union 2024). Similarly, NIST frames transparency as a characteristic that supports accountability and governance functions, not merely user information (NIST 2023).

Still, a normative question remains: transparency *for what*? If transparency is pursued primarily to optimize system acceptance or to manage reputational risk, it may reinforce the reversal of visibility. If it is pursued to protect persons and enable contestation, it must include institutional mechanisms—audits, rights to explanation and appeal, and meaningful public oversight.

A useful way to sharpen the concept is to treat transparency not as a property of an object but as a *relational norm*. Something is transparent *to someone, for a purpose, within an interpretive framework*. This aligns with critiques that ‘transparency’ can become a fetishized solution that ignores the social conditions required to make information meaningful (Ananny and Crawford 2018; Alloa 2018). Relational transparency clarifies why the reversal of visibility is ethically troubling: the relationship is asymmetrical. Individuals are transparent to institutions (through data capture and inference), while institutions are not equally transparent to individuals (through complexity, secrecy, or legal shields). This asymmetry can erode autonomy, because agency requires both the capacity to understand how one is being governed and the possibility to contest or opt out. Relational transparency also helps specify the normative question more precisely: which party in the relationship bears the burden of being transparent, and to what extent? In contexts of unequal power, transparency should track responsibility—those who exercise authority through classification, prediction, and decision-making owe stronger disclosure and justifiability, whereas those who are subject to these decisions require protected forms of opacity (privacy, discretion, and the ability not to be exhaustively rendered into data) if agency is to remain possible. This is where the discussion begins to open onto theological ethics, because it frames

“visibility” not as a neutral technical condition but as a moral relation: to know someone rightly is not merely to extract information about them, but to stand in a posture of accountability to the person known.

5. Omniscience without Grace: A Theological-Ethical Interpretation

This dynamic of invisible control of highly visible ordinary people, often referred to as “the Big Brother,” can also be recognized in the religious context of the omnipresent God, whom we cannot see. And it is precisely this dynamic that will lead us to the answer of how to overcome the problem of inverted visibility. The reversal of visibility is thus not only a policy problem; it touches deeper anthropological and theological questions about what it means to know a person.

Digital systems increasingly treat persons as knowable through data patterns—predictable and governable from the outside. A science-and-faith dialogue can illuminate why this is reductive: knowledge detached from relational responsibility can become a form of domination. Appeals to “hiddenness” in what follows should not be read as an apology for institutional secrecy. On the contrary: where institutions exercise power over persons, transparency and accountability are the priority; the value of the hidden concerns the protection of interiority, conscience, and relational disclosure in human life.

Christian theological traditions on the other hand insist that true knowledge of persons is inseparable from love, freedom, and the irreducibility of the human subject. Apophatic and mystical strands stress that God’s hiddenness is not mere absence but a protection against objectifying mastery—an invitation to humility before the mystery of the other (Pseudo-Dionysius 1987; Turner 1995). In this light, the problem with *digital datafied ‘omniscience’* is not that it knows too much, but that it knows in the wrong mode: as extraction and prediction rather than as relational recognition.

The phrase *omniscience without grace* names this mode. It describes an informational gaze that aims at total legibility, but without the moral

commitments that would make such knowledge protective and just. It can therefore produce a distorted anthropology: the person appears as a dataset, the conscience as a variable, the future as a probability distribution. Recent Catholic reflection on AI ethics resonates with this concern. The Rome Call for AI Ethics emphasizes that AI must serve human dignity and the common good, not reduce persons to mere means (Pontifical Academy for Life 2020). The Vatican note *Antiqua et Nova* similarly warns against confusing intelligence with mere calculation and urges that technological systems respect the relational and spiritual dimensions of the human person (Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education 2025). These interventions do not replace legal standards, but they supply an anthropological criterion: governance is inadequate if it treats transparency as total exposure rather than as a tool for protecting human dignity.

In both ordinary moral experience and theological anthropology, certain forms of hiddenness are constitutive of moral agency. Conscience, intention, and the inner life are not fully transparent objects; they are disclosed in trust and relationship, not extracted as raw data. When systems demand total visibility—continuous tracking, emotion inference, biometric categorization—they risk treating what should be relationally disclosed as something to be mined.

This has a direct connection to contemporary regulation. The EU AI Act recognizes that some uses, such as emotion recognition and biometric categorization, require explicit transparency duties and are bounded by other legal constraints (European Parliament and Council of the European Union 2024, art. 50). From a theological-ethical angle, the point is broader: even when such systems are lawful, they raise questions about whether inferring inner states from outward signals respects the dignity of the person or converts interiority into an object of control.

Hiddenness therefore has a positive ethical meaning: it protects the space in which persons can deliberate, repent, grow, and relate freely. An ethic of transparency that ignores this will drift into a culture where the good is equated with the visible and the hidden is equated with the suspicious (Han 2015).

6. Recovering the Value of the Hidden: Toward Proportional Transparency

If transparency can both protect and harm, the ethical task is to articulate criteria for *proportional transparency*: transparency as a means of safeguarding persons, not as an unconditional demand for exposure. Proportional transparency has two sides: (1) *maximal transparency about institutions* when they exercise power over persons, and (2) *protected opacity of persons* where intimacy, conscience, and contextual integrity are at stake.

First, institutional transparency should be strongest where decisions are high-stakes and rights-affecting. In such contexts, *post-hoc* public relations disclosures are insufficient. What is required is a governance stack: documentation, independent audit, clear responsibility assignments, and enforceable rights to contestation (NIST 2023; OECD 2019; UNESCO 2021). This corresponds to the intuition behind the EU AI Act's approach of combining transparency duties with broader compliance requirements for high-risk systems (European Parliament and Council of the European Union 2024).

Second, proportional transparency insists on protecting persons from being made fully transparent as data objects. Privacy theory describes this not as secrecy but as *contextual integrity*: information flows should be appropriate to context and governed by norms of legitimacy (Nissenbaum 2010). When AI governance demands more data disclosure than a context warrants, it turns transparency into coerced exposure.

Third, proportional transparency reframes explainability. The point of explanation is not to satisfy curiosity but to enable agency and accountability: to allow affected persons to understand a decision enough to challenge it, and to allow oversight bodies to test and correct system behavior (Doshi-Velez and Kim 2017; Rudin 2019). Where full model transparency is not feasible, auditability and contestability can still be achieved through documentation, standardized reporting, and secured access for regulators or independent auditors (Ananny and Crawford 2018; European Commission High-Level Expert Group on AI 2019).

Practically, proportional transparency can be implemented through layered disclosures: (a) clear notices to users; (b) meaningful explanations and reasons for affected persons in high-stakes decisions; (c) robust documentation for regulators; and (d) audit access for qualified independent evaluators. Such layering addresses the problem of ‘information overload’ that can make transparency meaningless at the user level (Koivisto 2022).

A science-and-faith dialogue strengthens this proposal by adding an anthropological criterion. If persons are more than profiles, then governance must resist reductionism: it must recognize that not everything that can be inferred should be inferred, and not everything that can be made visible should be made visible. This does not reject transparency; it orders transparency toward dignity. In this sense, theological reflection complements technical and legal work by articulating why certain forms of opacity—protected spaces of interiority, confession, and trust—are essential for moral life (Turner 1995; Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education 2025).

Finally, proportional transparency reframes the transparency paradox. The paradox is resolved not by total openness but by a disciplined distribution of visibility: institutions must be made accountable where they govern, while persons must be protected from being governed *as if* they were fully transparent objects.

Conclusion

The paper argued that contemporary AI ethics discourse risks a reversal of visibility: persons become increasingly transparent through data capture and inference, while AI infrastructures remain opaque. By distinguishing different dimensions of transparency and situating them in the political economy of datafied life, the paper showed why transparency cannot be treated as a simple moral remedy.

A science-and-faith perspective clarifies the deeper stakes. When knowledge is detached from relational responsibility, it resembles ‘omniscience without grace’: an informational gaze that can reduce

persons to profiles. Theological accounts of hiddenness and dignity therefore support a norm of proportional transparency: transparency where it protects persons and enables contestation, and justified opacity where it safeguards moral agency and the integrity of the inner life.

AI governance is, and should be, a scientific, technical, legal, and ethical enterprise; but it also presupposes an anthropology. Clarifying the ethical meaning of transparency is therefore part of clarifying what kind of human-centered technological future we are willing to build.

References

- Alloa, Emmanuel. 2018. "Transparency: A Magic Concept of Modernity." In *Transparency, Society and Subjectivity: Critical Perspectives*, edited by Emmanuel Alloa and Dieter Thomä, 21–55. Cham: Palgrave Macmillan.
- Alloa, Emmanuel. 2022. "Seeing Through a Glass, Darkly: The Transparency Paradox." In *This Obscure Thing Called Transparency: Politics and Aesthetics of a Keyword*, edited by Emmanuel Alloa, 9–28. Leuven: Leuven University Press.
- Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20 (3): 973–89. <https://doi.org/10.1177/1461444816676645>.
- Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732.
- Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 1–12. <https://doi.org/10.1177/2053951715622512>.
- Dennett, Daniel C., and Deb Roy. 2015. "Our Transparent Future." *Scientific American* 312 (3): 64–69. <https://doi.org/10.1038/scientificamerican0315-64>.
- Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education. 2025. *Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence* (28 January 2025). Vatican City: The Holy See.
- Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv:1702.08608.
- Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

- European Commission High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.
- European Parliament and Council of the European Union. 2024. *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)*. Official Journal of the European Union L, 2024/1689 (12 July 2024).
- Foucault, Michel. 1995. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan. New York: Vintage Books.
- Han, Byung-Chul. 2015. *The Transparency Society*. Translated by Erik Butler. Stanford, CA: Stanford University Press.
- Hansen, Hans Krause, Lars Thøger Christensen, and Mikkel Flyverbom. 2015. "Introduction: Logics of Transparency in Late Modernity: Paradoxes, Mediation and Governance." *European Journal of Social Theory* 18 (2): 117–31.
- Koivisto, Ida Ilmatar. 2016. *The Anatomy of Transparency: The Concept and Its Multifarious Implications*. EUI Working Paper MWP 2016/09. Florence: European University Institute.
- Koivisto, Ida Ilmatar. 2022. *The Transparency Paradox: Questioning an Ideal*. Oxford: Oxford University Press.
- Lambrecht, Anja, and Catherine Tucker. 2019. "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science* 65 (7): 2966–81. <https://doi.org/10.1287/mnsc.2018.3093>.
- Lyon, David. 2015. *Surveillance after Snowden*. Cambridge: Polity Press.
- National Institute of Standards and Technology. 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST. <https://doi.org/10.6028/NIST.AI.100-1>.
- Nissenbaum, Helen. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press.
- Organisation for Economic Co-operation and Development. 2019. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Paris: OECD.
- Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Pontifical Academy for Life. 2020. *Rome Call for AI Ethics* (28 February 2020). Vatican City: Pontifical Academy for Life.
- Pseudo-Dionysius. 1987. *Pseudo-Dionysius: The Complete Works*. Translated by Colm Luibheid. New York: Paulist Press.

- Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1: 206–15. <https://doi.org/10.1038/s42256-019-0048-x>.
- United Nations Educational, Scientific and Cultural Organization. 2021. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO.
- Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.