

ORIGINAL RESEARCH

Open access

Understanding Non-Transparency in Digital Platforms: Ethical, Governance, and Accountability Perspectives

Nokuthula Dlamini^{1*}, Siphso Zulu¹, Thabo Nkosi², Lerato Molefe³

Abstract

The integration of digital architectures powered by artificial intelligence (AI) into daily routines is expanding at an unprecedented rate. Within this paradigm, the twin dilemmas of accountability and openness endure as critical issues that have yet to be satisfactorily resolved. This study intends to delineate the specific domains of opacity within digital ecosystems and formulate actionable frameworks for mitigating these visibility challenges. Initially, by synthesizing existing scholarship, interrogating the dynamics of platformization, establishing a cross-disciplinary conceptualization of AI, and applying ethical AI frameworks, we map the dimensions of opacity inherent in digital environments. Subsequently, having mapped these obscure domains, we move beyond a standard literature synthesis to offer an independent perspective on mitigating these systemic vulnerabilities. Our evaluation of the literature indicates the existence of three primary spheres of opacity within digital networks: an inadequate understanding of platform content originators; a deficiency of visibility regarding the background labor force, their explicit operational outputs, and their occupational environments; and a lack of openness concerning the engineering and oversight of algorithmic models. Given the abundance of abstract, high-level ethical tenets in current scholarship that resist pragmatic implementation, this investigation represents an effort to reconcile theoretical concepts with operational execution.

Keywords AI ethics, Accountability, Governance, Transparency, Platformization

*Correspondence:

Nokuthula Dlamini
dlaminin@ukzn.ac.za

¹ Department of Digital Platform Governance and Transparency, Faculty of Engineering, University of KwaZulu-Natal, Durban, South Africa

² Department of Non-Transparency and Ethical AI, Faculty of Engineering, University of Cape Town, Cape Town, South Africa

³ Department of Accountability and Platform Ethics, Faculty of Engineering, University of the Witwatersrand, Johannesburg, South Africa

Introduction

This inquiry explores the dynamics of openness within digital networks that integrate artificial intelligence (AI) systems. Owing to their seamless accessibility and user-centric designs, digital infrastructures have become deeply embedded in contemporary societal routines. These architectures function as structural intermediaries, organizing and modulating how individuals engage with various categories of web-based information. For example, these systems facilitate interpersonal dynamics on networking sites, content distribution and streaming, knowledge gathering, and the provision of localized utilities spanning from ride-hailing to the distribution of sustenance and consumer goods. Furthermore, rapid innovations in computation have empowered AI models to refine and automate operational workflows, transforming this technology into an intrinsic element of these networks. The socio-economic weight of these systems as internet intermediaries has catalyzed the formulation of "platformization" as an analytical framework, which investigates the societal repercussions of digital architectures through a triad of dimensions: economic models, infrastructural control, and regulatory governance

[1]. Nevertheless, despite their manifest utilities, pressing moral dilemmas regarding these networks persist and demand resolution. Chief among these issues—and the core focus of this investigation—is the challenge of systemic openness.

Since the inception of digital networks, researchers have rigorously analyzed both their constructive and detrimental societal consequences. In terms of positive outcomes, extensive scholarly attention has focused on how these infrastructures catalyze social and civic mobilization. The Occupy movements [2, 3] and the Arab Spring [4, 5] serve as prominent illustrations of digitally coordinated civic actions. Conversely, severe controversies have emerged that stem directly from the opacity and deficient oversight of these systems. Illustrative historical instances include the 2016 United States presidential election [6], Brexit [7], the humanitarian crisis in Myanmar [8], and the 2019 mosque shootings in Christchurch, New Zealand [9]. Consequently, researchers like Gorwa and Ash [7] alongside Leone de Castris [10] highlight the vulnerabilities surrounding platform visibility and regulatory models, emphasizing the necessity of engineering robust disclosure policies.

From an engineering perspective, these networks and the deep integration of artificial intelligence models within them exhibit extreme structural complexity. This elevated sophistication, compounded by the proprietary nature of corporate assets, significantly drives these visibility deficits. Furthermore, because AI algorithms operate as corporate trade secrets with concealed operational logic, detecting and correcting algorithmic and data biases presents an exceptional challenge [11, 12]. Similarly, because these networks are predominantly controlled by profit-driven enterprises, corporate strategic directives frequently exacerbate opacity concerns. As established in contemporary scholarship, "...despite technical and organizational efforts to improve explainability, transparency and accountability, massive zones of non-transparency remain, caused by both the sheer complexity of technological systems and by strategic organizational decisions" [13] (p. 110). Beyond these engineering and corporate challenges, the deficit of enforceable state mandates coupled with a reliance on corporate self-regulation further complicates these ethical dilemmas. Currently, public policy measures are ill-equipped to adequately manage platform operations, allowing major technology conglomerates to self-govern using independently authored ethical frameworks and behavioral guidelines.

Mitigating these visibility deficits across all three dimensions (technological, organizational, and regulatory) requires foundational investigative steps: interrogating the mechanics of platformization, identifying a unified conceptual definition of AI applicable across these networks, and determining how ethical AI frameworks can be leveraged to expose hidden spheres of opacity. An effective operational definition of AI within this context must explicitly map the objectives, boundaries, targets, and constraints of computational systems within an ethical scope, thereby bridging the divide between theory and practice. As observed by Crawford *et al.* [14] and Mittelstadt [15], the absence of a unified, clear definitions of AI hinders attempts to translate abstract principles into operational practices. Along similar lines, Chouldechova and Roth [11] note that while research regarding algorithmic fairness is substantial, AI ethics still lacks unambiguous definitions, leaving numerous unresolved questions that demand attention.

While AI ethics broadly encompasses reflection on the societal and moral dimensions of artificial intelligence, this paper utilizes an approach where AI ethics addresses "...ways of deviation or distancing oneself from problematic routines of action, with uncovering blind spots in knowledge, and of gaining individual self-responsibility" [13] (p. 115). Our subsequent platform analysis is guided by this specific conceptual framework. We argue that from this ethical standpoint, the pervasive deployment and creation of platforms constitute the very 'problematic routines of action' from which society must distance itself. We apply the concept of "the platformization of cultural production" advanced by David B. Nieborg and Thomas Poell [1] to expose gaps in current knowledge, aiming to illuminate specific zones of non-transparency in platform engineering and consumption. We emphasize an urgent requirement to elevate awareness regarding the opacity risks tied to the use of dominant digital platforms, specifically Google, Apple, Facebook, Amazon, and Microsoft (GAFAM). Consequently, we investigate three distinct zones of non-transparency within these platforms:

- Lack of transparency with regard to who contributes to platforms.
- Lack of transparency with regard to who is working behind platforms, the contributions of those workers, and the working conditions of digital workers.

- Lack of transparency with regard to how algorithms are developed and governed.

To thoroughly address the issue of transparency in AI-enabled digital platforms, the subsequent section first refines our specific operational definitions of digital platforms, AI, AI ethics, and transparency.

Reflecting on the Concept

Digital platforms

The deployment of platforms within commercial frameworks is not a recent innovation. Historically, platforms have acted as vital mechanisms for linking providers with end-users in an efficient manner. For example, Van Alstyne, Parker, and Choudary [16] point to print media and retail malls as traditional vehicles for pairing audiences with marketers, and distinct brands with consumers, respectively. Similarly, Nishikawa and Orsato [17] note that ancestral local marketplaces operated under identical structural foundations. Nevertheless, the transition of these configurations into digital platforms over recent decades represents a profound turning point in their evolution. Rapid developments within information and communication technology (ICT) ecosystems have lowered expenses and simplified the process of scaling and extending the operational boundaries of these architectures [16].

Chen, Richter, and Patel [18] describe digital platforms as “digital systems that facilitate communications, interactions, and innovations to support economic transactions and social activities” [18] (p. 1306). Although this conceptualization prioritizes the relational dynamics between the various actors engaging with the system, Asadullah, Faik, and Kankanhalli [19] highlight the technical parameters as primary defining elements. From this structural viewpoint, emphasis is placed on the platform's capacity to operate as a foundational layer that supports an assortment of distinct digital services [19].

Owing to their pervasive implementation across diverse domains, digital platforms exert a powerful influence over cultural structures, prompting the formulation of the “platformization” paradigm [1]. Nieborg and Poell [1] articulate platformization “as the penetration of economic, governmental, and infrastructural extensions of digital platforms into the web and app ecosystems, fundamentally affecting the operations of the cultural industries” (p. 4276). Examining this phenomenon through economic, structural, and regulatory perspectives illuminates the multifaceted challenges surrounding ethical AI and openness in platform ecosystems.

From an economic perspective, although users ostensibly access these networks without monetary cost, there is a distinct lack of clarity and transparency regarding how individual data assets and behavioral footprints are extracted for data-mining and targeted marketing. Furthermore, individual users have shifted from passive media consumers to active content generators. This convergence of creation and consumption within user activities has led researchers to adopt the term ‘prosumption’ [20]. It remains highly ambiguous how individuals are remunerated for this voluntary labor. Consequently, the dimension of transparency requires a more rigorous investigation within this financial dynamic.

The processes through which platform architectures are engineered, sustained, and refactored introduce significant opacity challenges. Both individual consumers and complementors—which include corporate entities like content developers and advertisers—are profoundly dependent on these networks. Nieborg and Poell [1] characterize a platform's infrastructure as “a sociotechnical system that is widely shared and increasingly perceived as essential. Infrastructural access to Application Programming Interfaces (APIs) and Software Development Kits (SDKs) is among the primary ways in which platforms control complementors” (p. 4281). These foundational frameworks, manifest as APIs and SDKs, dictate the parameters of user engagement by defining the permissible formats, production mechanisms, and distribution channels for information. Moreover, due to the highly asymmetrical distribution of power between platform engineers and the public, users have little recourse but to conform to these structural rules [1, 21].

Regarding the regulatory lens, platform proprietors possess the authority to unilaterally institute policies and baselines governing how their environments are utilized. These regulations materialize as end-user license agreements and terms of service documents, which frequently obscure operational details and fail to offer substantive safeguards against

"targeting, harassment, and marginalization" [21] (p. 3). From an alternative standpoint, this regulatory perspective provides an avenue to evaluate how successfully institutional policymakers are responding to the broader pressures of platformization. Because digital platforms have a relatively brief history and evolve at an accelerated pace, institutional legal frameworks have struggled to keep pace with the legislative demands of these networks. For instance, Duguay [21] notes that "a lot of tech companies make their home between the moment some new way to make money is discovered and the moment some government entity gets around to deciding if it's actually legal" (p. 8).

Digital platforms comprise a highly heterogeneous spectrum, varying according to their user base, ecosystem parameters, and ultimate commercial objectives. As a consequence, platform governance—meaning the regulatory principles and guidelines dictating platform operations—is highly context-dependent. Gillespie [22] delineates two interconnected facets of this regulatory dynamic: the governance of platforms and governance by platforms. The governance of platforms encompasses external legislative mandates that dictate a platform's legal accountability for the behavior and content of its user base, while governance by platforms describes the internal mechanisms through which networks moderate information and regulate user conduct.

Governance of platforms

State legislative systems have generally failed to adapt agilely to the structural demands of the digital frontier, and digital platforms are deeply entangled in this regulatory lag. Specific statutes designed to confront the nuances of cyberspace remain insufficient, and the few legal frameworks currently in place exhibit fundamental deficiencies. Section 230 of the Communications Decency Act of 1996 in the United States serves as an instructive example of such legislation.

Section 230 of the Communications Decency Act decrees that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" [23]. Advocates of this statute argue that it paved the way for the internet's rapid commercial growth; conversely, detractors critique the law on the grounds that it fails to adequately curtail illicit conduct across online spaces [24]. Daub [25] (p. 39) underscores this legislative constraint by clarifying the functional divergence between platforms and traditional publishers, emphasizing that publishers bear an "editorial responsibility" that digital platforms consistently refuse to assume.

To mitigate some of the enforcement loopholes associated with Section 230 and compel platforms to bear greater accountability for hosted content, the United States Congress passed the Allow Victims and States to Fight Online Sex Trafficking Act of 2017 (FOSTA) alongside the Senate's Stop Enabling Sex Trafficking Act of 2017 (SESTA). Nonetheless, these statutory revisions suffer from systemic flaws, including the deployment of ambiguous and non-transparent legal terminology. This conceptual vagueness permits divergent interpretations that can ultimately lead to either excessively narrow or overly broad liabilities, culminating in either under-policing or over-policing online environments [24].

Ultimately, as Daub [25] observes, "regulation is supposed to be slow-moving, deliberate, a little bit after-the-fact" (p. 8). Operating within that paradigm, it is unrealistic to anticipate that external legal frameworks will possess the agility required to foresee and neutralize emerging challenges prospectively, or even to rectify them swiftly once they have manifested.

The legal immunity granted to digital platforms via Section 230 prompts an obvious inquiry: if these networks possess sophisticated capabilities for tracking user trajectories and mining personal datasets for behavioral profiling and target advertising, why do they claim that monitoring illicit content and executing preventative safety measures is an insurmountable technical hurdle when asked to police unethical exploitation? Aside from the challenges surrounding the external regulation of platforms, how these enterprises internally govern their own operational domains constitutes an equally pressing concern.

Governance by platforms

While the governance of platforms focuses on the legislative structures devised by independent external entities, governance by platforms concerns the internal capacity of these companies to moderate user content and arbitrate

interactions among diverse stakeholders. In an analysis of the literature surrounding crowdwork platform regulation, Gol, Stein, and Avital [26] summarize two primary vectors of internal governance: control and coordination. They outline the distinct characteristics of these two dimensions and illustrate how they intersect in practice. Within this regulatory framework, control signifies the capacity to monitor all interactions and operational processes occurring inside the platform environment. Conversely, coordination denotes the organizational capability to manage the "dependencies among tasks and resources that exist in the process" [26] (p. 9). It should be emphasized that their scholarly synthesis applies not only to crowdwork environments but extends to other platform archetypes as well.

Centralized and decentralized architectures of platform governance rely on fundamentally different control and coordination mechanisms. Each organizational methodology presents unique structural tradeoffs that, for the sake of conciseness, will not be thoroughly analyzed here [26]. Generally speaking, centralized modes of platform governance operate with lower levels of transparency, whereas decentralized models exhibit greater openness [27]. Under centralized models, executive authority is concentrated within a singular core, resulting in a high degree of restrictive oversight managed directly by the platform entity. Concurrently, the operational processes in these settings lack visibility. In contrast, decentralized governance frameworks are defined by the dispersion of decision-making authority across a wider matrix of stakeholders, which yields a much higher degree of systemic transparency [26, 27].

AI and ethics in the context of digital platforms

The lack of a universally recognized definition specifying the goals, boundaries, and constraints of AI is a primary driver behind the opacity observed in AI-driven networks. Academic consensus indicates that the vague, overly broad, and conflicting conceptualizations held by different parties—such as software engineers, regulators, and consumers—hinder the creation of practical ethical frameworks capable of resolving current accountability problems [14, 28-30]. Two main factors explain this lack of conceptual clarity.

First, establishing a standard definition for human intellect has historically proven elusive [31], and artificial intelligence is no exception to this trend. The issue is further complicated because AI does not refer to a single tool with a narrow application; rather, it is integrated across diverse commercial sectors and advanced systems, ranging from targeted advertisement distribution to predictive policing models.

Despite the absence of a uniform definition for intelligence, Legg and Hutter [31] analyzed approximately 70 different interpretations to map out significant commonalities and conceptual linkages. They isolated three universal traits of intelligence:

1. The capacity to interact with an external environment.
2. The ability to successfully achieve designated targets.
3. The capacity to adapt flexibly to shifting goals and surroundings.

Reflecting these shared dimensions, Legg and Hutter [31] offered a practical, informal baseline: "Intelligence measures an agent's ability to achieve goals in a wide range of environments" (p. 9).

Second, the domain of AI and its ethical governance involves diverse interest groups—such as developers and end-users—whose perspectives often diverge or conflict. For instance, by compiling a diverse set of AI definitions for a technical assessment, Samoili et al. [32] distinguished three unique viewpoints driving definition design: institutional/policy bodies, academic research, and the commercial market. Regulatory and institutional frameworks use these definitions to assess how AI affects society and to drive technological growth. The research orientation treats AI primarily as a scientific field of study. Meanwhile, the market orientation evaluates AI strictly through the lens of economic value.

By evaluating this gathered selection of definitions, Samoili et al. [32] isolated four core dimensions that characterize AI:

“1. Perception of the environment, including the consideration of the real world complexity, 2. Information processing, 3. Decision making, and 4. Achievement of specific goals” (p. 8).

Notably, these four pillars display substantial thematic overlap with the core components of intelligence outlined by Legg and Hutter [31]. Beyond these definitional challenges, Crawford et al. [14] emphasized that existing AI ethical codes have been overwhelmingly produced by organizations situated in the Global North, which are historically dominated by white males. Consequently, perspectives from the Global South remain severely underrepresented in this discourse, even though these regions are frequently the most vulnerable to the magnified socio-economic disparities generated by the AI sector.

In a similar vein, the absence of a structured definition for AI within digital environments impedes the deployment of enforceable, measurable ethical standards. Establishing precise conceptual definitions is a prerequisite for drafting ethical codes in any professional domain. To illustrate this point, Mittelstadt [15] contrasted the mature domain of medical ethics against the emerging field of AI ethics across four structural pillars:

- (1) common aims and fiduciary duties
- (2) professional history and norms
- (3) proven methods to translate principles into practice
- (4) robust legal and professional accountability mechanisms

First, whereas medicine is united by the singular, cohesive objective of safeguarding human health and well-being, AI development is fragmented by the conflicting objectives of engineers, users, and external bodies. End-users typically prioritize the enhancement of human welfare, whereas system developers are frequently driven by commercial considerations. Concurrently, state actors and governments may seek to leverage AI architectures to enhance state surveillance and regulatory tracking. Second, unlike the long-standing field of medicine, AI represents a contemporary discipline lacking a deeply entrenched institutional history.

While the tenets of bioethics and medical practice were refined over multiple centuries, the origins of AI ethics only extend back to the 1950s [33]. This brief institutional history has prevented the field of AI from integrating the collective insights that typically accumulate over extended periods. Third, this same historical depth in medicine has yielded reliable methodologies for operationalizing abstract ethical principles into concrete clinical practices—mechanisms that remain underdeveloped within AI. Finally, whereas medical practice is highly governed by comprehensive legal and statutory frameworks, the field of AI operates under highly restricted and fragmented regulatory oversight.

As a baseline conceptual foundation, the definition of AI employed and advocated throughout this study mirrors the core framework utilized by Samoili et al. [32], which was originally formulated by the High-Level Expert Group on Artificial Intelligence (HLEG):

“Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions” [34] (p. 6).

In the realm of digital platforms, AI clearly serves as a foundational component of everyday operational workflows. Historically, AI systems were restricted to localized, narrow applications on digital networks. However, driven by rapid technological progression, AI has transformed into an indispensable, structural component of modern platforms [35]. At present, AI architectures are deployed to execute functions critical to platform survival, including target advertisement

distribution, search engine optimization and ranking, computer vision and image classification, automated translation, consumer demand forecasting, and algorithmic content moderation [9, 35].

AI ethics and transparency

Throughout the past several decades, AI-powered digital platforms have exerted a profound influence on pivotal global political events and societal disruptions. While the precise magnitude of their impact remains a subject of ongoing debate, their underlying influence is universally recognized. Notable instances reflecting the critical nature of platform transparency include the 2016 United States presidential election, the Brexit referendum, the orchestration of genocide in Myanmar [7], and the 2019 mosque shootings in Christchurch, New Zealand [9]. Consequently, diverse stakeholders—including researchers, legislators, and engineering firms—have channeled significant resources into resolving transparency deficits across digital networks. From the perspective of academia and public policy, this has catalyzed new conceptual paradigms and legislative proposals; from the developer perspective, firms have increasingly positioned transparency as a core accountability tool designed to rebuild public trust damaged by historical failures [7, 10].

This investigation focuses heavily on transparency because it has consistently emerged as the most ubiquitous principle embedded within international AI ethical codes and policy guidelines [36]. Within digital platform ecosystems, Suzor, Van Geelen, and Myers West [37] conceptualize transparency as encompassing “the right of users to know when, why, and by whom decisions that affect them are being made” (p. 392). However, although transparency seems intuitive and direct on the surface, Gorwa and Ash [7] observe that it functions as a highly popular mechanism of oversight precisely due to its inherent “flexibility and ambiguity.” In a complementary critique, Leone de Castris [10] characterizes contemporary initiatives aimed at bolstering the transparency and accountability of digital platforms as fundamentally “vague and superficial.”

Zones of Non-Transparency in Platforms

Within platform ecosystems, transparency serves as a foundational element. For instance, empirical research conducted by Deng et al. [38] highlights transparency as a core priority for crowdworkers. The essential values mapped out by Deng et al. [38] encompass access, autonomy, fairness, transparency, communication, security, accountability, making an impact, and dignity. In order to expose gaps in current understanding and illuminate overlooked areas of opacity within the creation and deployment of platforms, this study applies an AI ethics lens to three specific, under-investigated zones of non-transparency. The first area of opacity concerns the question of ‘who contributes to platforms?’. The second area investigates ‘who is working behind platforms and what are their work conditions?’. Lastly, the third zone evaluates ‘how algorithms are developed and governed’.

Non-transparency on who contributes to platforms

The digital platform environment relies on a varied assembly of participants, which includes owners, software architects, and consumers; specifically, “...a digital platform may need its platform owners to manage its core technology and governance systems, third-party developers to build innovative applications on top of its core technology, and end-users to utilize its core and complementary services” [18] (p. 1308). In contrast to widely recognized contributors such as engineers and end-users, platform laborers constitute a segment of contributors who have remained largely unacknowledged for an extended duration. These laborers are vital to the operational viability of digital networks; a platform worker can be characterized as “those who have ever gained income from providing services via online platforms, where the match between provider and client is made digitally, payment is conducted digitally via the platform, and work is performed either (location-independent) web-based or on-location” [39] (p. 3). Brancati et al. [39] categorize these laborers into three distinct tiers depending on their level of engagement:

- **Main platform workers** commit to more than 20 hours weekly or generate at least 50% of their total revenue through digital platform assignments.

- **Secondary platform workers** commit to more than 10 hours weekly or generate 25–50% of their revenue through digital platform assignments.
- **Marginal platform workers** commit to fewer than 10 hours weekly or generate 25% or less of their revenue through digital platform assignments.

Across Europe, the population of platform laborers is expanding consistently, alongside a growing demographic of female participants. While the typical demographic profile points toward a young male, data indicates that “platform workers tend to be younger, more educated, and more likely to live in a larger household and have dependent children” [39] (p. 4). Nevertheless, obtaining precise census data regarding this workforce is remarkably challenging due to several constraints. One primary hurdle is the absence of a standardized definition for what constitutes platform labor. In a 2018 assessment, collected data became distorted because “many respondents poorly understood the definition, such as answering yes if they merely made use of a computer or mobile app in their job” [40] (p. 18). Furthermore, conceptual definitions for platform laborers vary significantly across international boundaries [40] (p. 19). Another complicating factor is the scarcity of institutional data managed directly by the platforms; due to regulatory ambiguities surrounding digital labor networks, these workers are frequently excluded from official datasets. For instance, ride-hailing applications obscure the distinctions between hailing a taxi on the street and booking a private vehicle in advance, with numerous applications leveraging regulatory gaps in current employment frameworks [40] (p. 20).

These policy and regulatory challenges are examined in greater detail later in this paper. Given the scarcity of data on precise workforce numbers, the obscured identity of those contributing to these networks, and the conflicting global definitions of platform labor, this domain represents the initial zone of non-transparency affecting the ethical framework of digital platforms.

Non-transparency on contributions of workers and their work conditions

The second area of opacity in digital networks involves the actual contributions and operational environments of platform laborers. The asymmetrical power dynamic between laborers and corporations heavily influences how networks deploy AI-driven methodologies to distribute tasks, monitor efficiency, and establish compensation rates to optimize corporate profitability [14]. As an illustration, the output and efficiency of laborers in Amazon fulfillment centers are tracked via computerized routines; even the mechanisms governing disciplinary procedures are fully automated, remaining indifferent to contextual variables that might impact worker output [14]. Within this computerized tracking framework, warehouse personnel are monitored continuously by an algorithm, and falling below an automatically generated productivity benchmark triggers immediate automated disciplinary measures ranging from formal reprimands to termination [14, 41]. The operational metrics underlying these AI-driven platforms are withheld from public view, rendering this a critical domain requiring enhanced transparency.

A prominent reflection of this power asymmetry is found in the sector of commercial content moderation. Commercial content moderators (CCMs) are the personnel tasked with screening user-produced material on digital networks. These individuals perform a vital function for digital platforms because the complex decisions involved in approving or censoring content currently exceed the functional boundaries of AI. Roberts [42] outlines four primary organizational models for moderation labor: in-house setups, boutique firms, call centers, and microlabor platforms.

For in-house CCMs, employment is predominantly coordinated through third-party labor contractors. These moderators are typically denied healthcare coverage, excluded from the standard incentives provided to permanent corporate staff, given low hourly wages instead of fixed salaries, and left without long-term job security [42]. Regarding call-center operations, Roberts [42] notes that CCM personnel based in Manila, Philippines received approximately \$400 monthly. For microlaborers, compensation is typically structured on a “pay-per-task” model, with rates starting as low as \$0.01 per completed assignment [43]. In 2020, the typical hourly compensation rate hovered “between approximately EUR 7 for microtasking and EUR 23 for software development” [39].

Across all these organizational models, a recurrent pattern emerges that mirrors the externalized, low-skill employment frameworks conceptualized by Huws [44]. Within these deregulated labor markets, positions are characteristically temporary or part-time, offering minimal employment stability. This vulnerability is exacerbated for microlaborers, whose digital workspaces operate with even less regulatory oversight than the other three operational categories [45].

Non-transparency on how algorithms are developed

The third dimension of institutional opacity centers on the deficit of regulatory oversight across digital platforms, which fundamentally obscures how platform-based processing models are constructed and deployed. Automated machine learning and artificial intelligence frameworks are now deeply embedded within day-to-day routines. They dictate both tangible and intangible outcomes in modern decision-making pipelines, managing tasks as varied as commercial advertisement delivery, multimedia recommendation streams, credit application reviews, and even high-stakes arenas like public bail evaluations and predictive policing models [11, 12, 46]. Despite this ubiquitous presence, contemporary deep learning mechanisms operate primarily as proprietary, uninterpretable "black boxes" shielded by intellectual property protections. Large language models (LLMs) serve as a prominent example of this lack of clarity.

To ensure LLMs are applied responsibly, users must subject these systems to continuous critical evaluation. The specific parameters of this evaluation naturally vary by domain. For instance, a structured rubric designed to critically assess the clinical implementation of LLMs within healthcare has been formulated [47] (p. 185). Under the framework developed by Ilicki [47], the initial phase of ethical verification requires identifying the primary data repository used to train or power the model. The second phase involves defining the explicit target audiences or recipients of the model's outputs. Once these foundational elements are established, evaluating the inherent limitations of the system becomes paramount. Analysts must interrogate the system with targeted ethical queries: what constraints exist within the data repository relative to the target population?; was the source data aggregated through ethical collection methods?; does the generated output precisely and efficiently meet user requirements?; and are there potential ethical harms embedded within the information provided to the end-user? Ilicki [47] emphasizes that this baseline critique should never serve as a substitute for exhaustive, comprehensive validation protocols prior to live industry deployment [186]. Nevertheless, within this discussion, the framework offers a practical tool to map out system limitations and systematically appraise the ethical soundness of an LLM.

Furthermore, the foundational data utilized to train these predictive systems frequently contains historical biases that distort subsequent downstream conclusions. For example, regarding automated facial analysis frameworks, a systematic literature review by Khalil, Ahmed, Khattak, and Al-Qirim [12] documented pervasive algorithmic bias directed "against a specific group or category". While these training disparities are well-documented, applying opaque machine learning models to life-altering legal proceedings—such as criminal recidivism risk profiling—remains highly problematic. Within the United States, judicial systems across nine states—Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin—have integrated automated risk scores into courtrooms to assist judges during sentencing determinations [48, 49]. Because AI now functions as a core component of modern digital infrastructure, maximizing transparency surrounding algorithmic design has become an urgent public interest.

Compounding this problem, empirical research by Nguyen, Yosinski, and Clune [50] demonstrates that deep neural networks (DNNs) remain highly susceptible to adversarial manipulation, processing visual stimuli in a manner fundamentally distinct from human perception. The authors proved that

"discriminative DNN models are easily fooled in that they classify many unrecognizable images with near-certainty as members of a recognizable class" [50] (p. 434).

This vulnerability introduces significant risks given the accelerating reliance on automated logic systems in public administration. Because digital platform operators increasingly offload core maintenance and administrative duties onto AI systems, resolving the opacity surrounding code development is vital. Under current opaque practices, machine learning

algorithms are highly vulnerable to internal and external manipulation, making it exceptionally difficult to isolate, audit, or correct the root causes of systemic bias.

Conclusion

The analytical scope of this study is twofold. First, by synthesizing current literature, evaluating the ethics of platformization, and utilizing established AI ethical frameworks, this paper maps out previously obscured zones of opacity within digital networks. The second objective is to offer actionable structural alternatives to resolve these accountability deficits. While the preceding sections analyzed the core vulnerabilities, these closing observations detail potential pathways forward.

Identify contributors

Platform architectures must implement clear reporting standards to explicitly identify the human and automated entities driving their success, alongside the specific volume of their contributions. Because international labor standards currently lack a uniform definition for online work, establishing a precise, standardized conceptualization of platform labor is an essential step toward recognizing worker rights and vulnerabilities. This definition must encompass a broad spectrum of digital occupations to remain comprehensive. A standardized baseline would enable platform laborers to better self-identify, helping dismantle the pervasive culture of anonymity that isolates the digital workforce. Furthermore, internal corporate policies and external legislative mandates must be expanded to safeguard these workers. Historical regulatory deficits have resulted in a severe shortage of administrative data and institutional reporting on platform labor. Enhancing data collection and reporting mechanisms will naturally bring worker compensation and operational environments into public view.

Identify governance/self-governance

Given the market consolidation currently enjoyed by major technology firms, the absence of robust policy frameworks capable of auditing platform administration represents a significant systemic risk. As argued by Golumbia [51], this vulnerability is exacerbated by corporate attempts to implement centralized self-regulation free from public oversight, a strategy that inevitably results in increasingly opaque data management. Power remains highly concentrated among platform owners within centralized governance models, leaving secondary stakeholders with virtually no leverage. Conversely, Chen *et al.* [18] note that semi-decentralized governance models distribute greater authority and administrative control to independent developers and external stakeholders, a structural shift strongly correlated with superior operational performance. Consequently, semi-decentralization can incentivize developer engagement, stimulate collective problem-solving, and accelerate technological refinement. Ultimately, Chen *et al.* [18] position shared, semi-decentralized administration as a superior alternative to both completely closed or fully decentralized systems.

Broad-scale transparency can be achieved through multiple avenues, several of which rely directly on legislative action. For instance, the Transparency Working Group of the Global Platform Governance Network actively evaluates structural opacity on digital networks to draft targeted policy recommendations for international legislators (MacCarthy [52]). Alternative strategies rely on open data initiatives, comprehensive mapping of the broader platform ecosystem, and institutional codes of conduct. In the context of crowdworking, where statutory protections remain weak, public information sharing serves as a viable interim mechanism to improve accountability. For example, the independent web platform <http://faircrowd.work/> (accessed on 25 September 2023) systematically audits digital labor platforms, publishing independent performance reviews for public consumption.

Voluntary codes of conduct and institutional ethics protocols offer another mechanism to improve transparency across the three dimensions examined in this paper: the identity of platform contributors, the hidden labor and conditions of the workforce, and the governance of automated algorithms. A major limitation of voluntary ethical guidelines, however, is their tendency to rely on broad, abstract principles that are incredibly difficult to legally enforce. Similarly, formulating

quantifiable metrics to audit AI ethics remains an immense challenge. The field of AI governance is fragmented by conflicting perspectives, with researchers criticizing purely principle-based models, human-rights-centric frameworks, and design-oriented ethical methodologies. These theoretical frameworks frequently fail to influence real-world corporate environments [53] (p. 1429). As observed by Joris Krijger [53]:

“Ethical decisions regarding model development and deployment are ultimately made within contexts of organizations that have to align the ethical principles with vested interests such as the organizational culture, mission and goals” (p. 1429).

Consequently, while a universal metric for evaluating AI ethics does not yet exist, current industry trends indicate that ethical standards and their corresponding verification metrics must be tailored to the specific demands of each industry sector.

Regulatory frameworks can be instituted by state authorities, independent industry watchdogs, or the platform owners themselves. This paper has detailed the necessity of updating existing statutory safe harbors—such as Section 230—while demonstrating that purely internal corporate self-regulation allows firms to maintain excessive, unchecked operational control. For models of independent, industry-wide self-regulation, successful precedents can be found in adjacent markets. A notable example is the Entertainment Software Rating Board (ESRB) within the interactive entertainment sector. Formed in 1994 in direct response to US Congressional hearings regarding violent video game content, the ESRB operates as an independent regulatory body [54]. In an empirical study of the system, Lacznia, Carlson, Walker, and Brocato [55] confirmed that the ESRB framework effectively assists consumers in managing media exposure. Independent, self-regulatory frameworks of this nature provide valuable blueprints for reforming the governance and accountability of modern digital platforms.

Identify datasets and explanation of decisions

As previously indicated, the data repositories utilized for training machine learning algorithms are frequently characterized by inherent biases. Additionally, these computational models currently operate much like uninterpretable “black boxes” that fail to offer rationale for their inferences and final outputs. Given that AI has integrated fundamentally into digital platform operations, it is critical to resolve the transparency and equity dilemmas surrounding this sector. Although the AI community has access to certain open-source data repositories, the misrepresentation within these datasets remains a widely recognized obstacle [12]. Challenges surrounding data availability and the clarity of data analysis are not unique to AI ethics. Indeed, other academic disciplines with much longer institutional histories have actively confronted the structural hurdles linked to these exact issues. In contemporary scientific publishing, researchers face increasing pressure to maintain absolute transparency regarding their foundational data, methodologies, institutional funding, and potential conflicts of interest.

Regarding research data, numerous academic journals now strongly encourage or strictly mandate that authors deposit their evaluated datasets into open-access repositories [56–58]. These initiatives are designed to empower independent researchers to replicate findings for verification or repurpose the data for subsequent investigations. For example, within psychological research, Hardwicke *et al.* [58] attempted to replicate 1,324 specific target values derived from 35 publications that provided reusable datasets; their analysis revealed that 64 of those values could not be successfully reproduced even within a 10% margin of error. Investigations of this nature underscore the vital role of data availability mandates. Accordingly, we suggest that implementing comparable data-sharing frameworks would highly benefit the domain of AI governance. If training data repositories are made public, subjected to algorithmic bias audits by oversight bodies, and distributed transparently among engineers, systemic biases can be uncovered and systematically corrected over time.

Beyond the challenges of training data distortion, the systemic bias of the algorithm itself represents a critical vulnerability that must be managed. At present, it remains methodologically unfeasible to reconstruct the precise logical pathway an algorithm follows to arrive at a specific conclusion. Compounding this issue, empirical evidence indicates that these

mathematical models remain highly susceptible to adversarial manipulation and deception [50]. Driven by rapid technological progress and a growing consensus demanding higher levels of equity and openness, innovative paradigms like Explainable AI (XAI) have been introduced [46]. The core objective of XAI is to demystify and clarify the internal decision-making pipelines of automated models. While a universal standard for this paradigm has not yet been finalized, the XAI taxonomy established by Gilpin et al. [46] categorizes approaches based on

“whether the method tries to explain the processing of the data by a network, explain the representation of data inside a network or to be a self-explaining architecture to gain additional meta predictions and insights about the method” (p. 86).

Consequently, we recommend that, whenever technically viable, the explanatory insights generated by XAI mechanisms should be disclosed to the general public rather than remaining exclusively restricted to software developers.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Sep 2024

Revised: 19 Oct 2024

Accepted: 17 Dec 2024

Published online: 15 March 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Nieborg DB, Poell T. The platformization of cultural production: theorizing the contingent cultural commodity. *New Media Soc.* 2018;20(11):4275-92.
<https://doi.org/10.1177/1461444818769694>.
- Juris JS. Reflections on #Occupy Everywhere: social media, public space, and emerging logics of aggregation. *Am Ethnol.* 2012;39(2):259-79.

<https://doi.org/10.1111/j.1548-1425.2012.01362.x>.

Tremayne M. Anatomy of protest in the digital era: a network analysis of Twitter and Occupy Wall Street. *Soc Mov Stud*. 2014;13(1):110-26.

<https://doi.org/10.1080/14742837.2013.830969>.

Rane H, Salem S. Social media, social movements and the diffusion of ideas in the Arab uprisings. *J Int Commun*. 2012;18(1):97-111.

<https://doi.org/10.1080/13216597.2012.662168>.

Ghannam J. Social media in the Arab world: leading up to the uprisings of 2011. Washington, DC: Center for International Media Assistance; 2011. p. 1-44.

Allcott H, Gentzkow M. Social media and fake news in the 2016 election. *J Econ Perspect*. 2017;31(2):211-36.

<https://doi.org/10.1257/jep.31.2.211>.

Gorwa R, Ash TG. Democratic transparency in the platform society. In: Tucker JA, Persily N, editors. *Social media and democracy: the state of the field, prospects for reform*. Cambridge: Cambridge University Press; 2020. p. 286-312.

Stevenson A. Facebook admits it was used to incite violence in Myanmar. *New York Times*. 2018 .

Gorwa R, Binns R, Katzenbach C. Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data Soc*. 2020;7:1-15.

<https://doi.org/10.1177/2053951719897945>.

Leone de Castris A. Types of platform transparency: an analysis of digital platforms and policymakers discourse on Big Tech governance and transparency. Chicago: University of Chicago; 2022.

Chouldechova A, Roth A. A snapshot of the frontiers of fairness in machine learning. *Commun ACM*. 2020;63(5):82-9.

<https://doi.org/10.1145/3376898>.

Khalil A, Ahmed SG, Khattak AM, Al-Qirim N. Investigating bias in facial analysis systems: a systematic review. *IEEE Access*. 2020;8:130751-61.

<https://doi.org/10.1109/ACCESS.2020.3009844>.

Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach*. 2020;30(1):99-120.

<https://doi.org/10.1007/s11023-020-09517-8>.

Crawford K, Dobbe R, Dryer T, Fried G, Green B, Kaziunas E, et al. *AI Now 2019 report*. New York: AI Now Institute; 2019.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019;1:501-7.

<https://doi.org/10.1038/s42256-019-0114-4>.

Van Alstyne MW, Parker GG, Choudary SP. Pipelines, platforms, and the new rules of strategy. *Harv Bus Rev*. 2016;94(4):54-62.

Nishikawa BT, Orsato RJ. Professional services in the age of platforms: towards an analytical framework. *Technol Forecast Soc Change*. 2021;173:121131.

Chen Y, Richter JI, Patel PC. Decentralized governance of digital platforms. *J Manag*. 2021;47(5):1305-37.

Asadullah A, Faik I, Kankanhalli A. Digital platforms: a review and future directions. In: *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*; 2018. p. 1-14.

Tapscott D, Williams AD. *Wikinomics: how mass collaboration changes everything*. London: Penguin; 2006.

Duguay S. Running the numbers: modes of microcelebrity labor in queer women's self-representation on Instagram and Vine. *Soc Media Soc.* 2019;5(4):1-11.
<https://doi.org/10.1177/2056305119894002>.

Gillespie T. Governance of and by platforms. In: Burgess J, Poell T, Marwick A, editors. *The SAGE handbook of social media*. London: SAGE Publications; 2017. p. 254-78.

United States Congress. Communications Decency Act of 1996, 47 USC §230. 1996.

McKnelly M. Untangling SESTA/FOSTA: how the Internet's "knowledge" threatens anti-sex trafficking law. *Berkeley Technol Law J.* 2019;34:1239-76.

Daub A. *What tech calls thinking: an inquiry into the intellectual bedrock of Silicon Valley*. New York: FSG Originals; 2020.

Gol ES, Stein MK, Avital M. Crowdwork platform governance toward organizational value creation. *J Strateg Inf Syst.* 2019;28(2):175-95.
<https://doi.org/10.1016/j.jsis.2019.01.001>.

Hein A, Schreieck M, Wiesche M, Krcmar H. Multiple-case analysis on governance mechanisms of multi-sided platforms. In: *Proceedings of Multikonferenz Wirtschaftsinformatik*; 2016.

Greene D, Hoffmann AL, Stark L. Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*; Grand Wailea, HI; 2019. p. 2126-35.

Larsson S, Heintz F. Transparency in artificial intelligence. *Internet Policy Rev.* 2020;9(2).
<https://doi.org/10.14763/2020.2.1469>.

Whittlestone J, Nyrupe R, Alexandrova A, Cave S. The role and limits of principles in AI ethics: towards a focus on tensions. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*; Honolulu, HI; 2019. p. 195-200.

Legg S, Hutter M. A collection of definitions of intelligence. In: Goertzel B, Wang P, editors. *Advances in artificial general intelligence: concepts, architectures and algorithms*. Amsterdam: IOS Press; 2007. p. 17-24.

Samoli S, Cobo ML, Gomez E, De Prato G, Martinez-Plumed F, Delipetrev B. *AI Watch. Defining artificial intelligence: towards an operational definition and taxonomy of artificial intelligence*. Luxembourg: Publications Office of the European Union; 2020.
<https://doi.org/10.2760/019901>.

Borenstein J, Grodzinsky FS, Howard A, Miller KW, Wolf MJ. AI ethics: a long history and a recent burst of attention. *Computer.* 2021;54(1):96-102.
<https://doi.org/10.1109/MC.2020.3034950>.

Mucha T, Seppälä T. *Artificial intelligence platforms: a new research agenda for digital platform economy*. Helsinki: Research Institute of the Finnish Economy; 2020. ETLA Working Papers 76.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell.* 2019;1:389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.

Suzor N, Van Geelen T, Myers West S. Evaluating the legitimacy of platform governance: a review of research and a shared research agenda. *Int Commun Gaz.* 2018;80(4):385-400.
<https://doi.org/10.1177/1748048518757142>.

Deng X, Joshi KD, Galliers RD. The duality of empowerment and marginalization in microtask crowdsourcing: giving voice to the less powerful through value sensitive design. *MIS Q.* 2016;40(2):279-302.

Urzi Brancati MC, Pesole A, Fernandez-Macias E. New evidence on platform workers in Europe: results from the second COLLEEM survey. Luxembourg: Publications Office of the European Union; 2020.
<https://doi.org/10.2760/459278>.

O'Farrell R, Montagnier P. Measuring digital platform-mediated workers. *New Technol Work Employ*. 2020;35(1):130-44.
<https://doi.org/10.1111/ntwe.12155>.

Lecher C. How Amazon automatically tracks and fires warehouse workers for 'productivity'. *The Verge*. 2019 .

Roberts ST. *Behind the screen: content moderation in the shadows of social media*. New Haven: Yale University Press; 2019.

Ross J, Irani L, Silberman MS, Zaldivar A, Tomlinson B. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Extended Abstracts*; 2010. p. 2863-72.
<https://doi.org/10.1145/1753846.1753873>.

Huws U. *Labor in the global digital economy: the cybertariat comes of age*. New York: NYU Press; 2014.

Marvit MZ. How crowdworkers became the ghosts in the digital machine. *The Nation*. 2014.

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*; Turin, Italy; 2018. p. 80-9.
<https://doi.org/10.1109/DSAA.2018.00018>.

Ilicki J. A framework for critically assessing ChatGPT and other large language artificial intelligence model applications in health care. *Mayo Clin Proc Digit Health*. 2023;1(2):185-8.
<https://doi.org/10.1016/j.mcpdig.2023.04.001>.

Angwin J, Larson J. Bias in criminal risk scores is mathematically inevitable, researchers say. *ProPublica*. 2016 Dec 30.

Yapo A, Weiss J. Ethical implications of bias in machine learning. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*; Waikoloa Village, HI; 2018. p. 2205-12.
<https://doi.org/10.24251/HICSS.2018.299>.

Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; Boston, MA; 2015. p. 427-36.
<https://doi.org/10.1109/CVPR.2015.7298640>.

Columbia D. Do you oppose bad technology, or democracy? *Medium*. 2019 Apr 24.

MacCarthy M. *Transparency recommendations for regulatory regimes of digital platforms*. Waterloo, ON: Centre for International Governance Innovation; 2022.

Krijger J. Enter the metrics: critical theory and organizational operationalization of AI ethics. *AI Soc*. 2022;37:1427-37.
<https://doi.org/10.1007/s00146-021-01374-2>.

Kocurek CA. Night Trap: moral panic. In: *How to play video games*. New York: New York University Press; 2019. p. 309-16.

Laczniak RN, Carlson L, Walker D, Brocato ED. Parental restrictive mediation and children's violent video game play: the effectiveness of the Entertainment Software Rating Board (ESRB) rating system. *J Public Policy Mark*. 2017;36(1):70-8.
<https://doi.org/10.1509/jppm.15.099>.

Federer LM, Belter CW, Joubert DJ, Livinski A, Lu YL, Snyders LN, et al. Data sharing in PLOS ONE: an analysis of Data Availability Statements. *PLoS One*. 2018;13(5):e0194768.
<https://doi.org/10.1371/journal.pone.0194768>.

Gherghina S, Katsanidou A. Data availability in political science journals. *Eur Polit Sci*. 2013;12(3):333–49.
<https://doi.org/10.1057/eps.2013.8>.

Hardwicke TE, Mathur MB, MacDonald K, Nilsson G, Banks GC, Kidwell MC, et al. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R Soc Open Sci*. 2018;5:180448.
<https://doi.org/10.1098/rsos.180448>.

Gunning D. Broad agency announcement explainable artificial intelligence (XAI). Arlington, VA: Defense Advanced Research Projects Agency; 2016.