Normative Inflation and the Crying Wolf Effect in the International Governance of AI

Luciano Floridi[1,2]

[1]Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, U.S.

[2]Department of Legal Studies, University of Bologna, Via Zamboni, 27/29, 40126, Bologna, IT

Abstract

The international governance of artificial intelligence (AI) is often said to suffer from a regulatory gap. In this article, I argue that the opposite problem is equally significant. The governance field is characterised by institutional proliferation and normative inflation: an excess of principles, declarations, and advisory bodies that lack binding authority or hierarchical coordination. This 'crying wolf' dynamic results from repeated, high-urgency normative signalling that, without enforcement mechanisms, gradually diminishes the credibility, relevance, and guiding power of the norms themselves. I distinguish between iterative soft law—which advances towards binding instruments through successive approximation—and substitutive soft law, which replaces enforcement with expression. I contend that this dynamic poses a structural threat to the performative credibility of technology governance and represents an ethical failure, as it gives everyone, particularly those most vulnerable to AI's harms, the false illusion of protection while offering none. The challenge facing the international governance of AI is to cease generating more empty signals and to strengthen institutional influence through delegated authority, normative and epistemic integration, and credible consequences. I conclude by proposing three design criteria that any future consolidation effort must meet, and by asserting that the ethics of governance architecture and its design deserve as much philosophical scrutiny as the ethics of the technologies being developed and overseen.

Keywords: normative inflation, crying wolf effect, AI governance, governance illusion, regime complexity, performative credibility, soft law

1. Introduction: The Paradox of Governance Plenitude

In November 2023, twenty-eight governments signed the Bletchley Declaration, pledging historic cooperation on AI safety. Six months later, the Seoul Ministerial Statement pledged historic cooperation on AI safety. Nine months after that, the Paris AI Action Summit pledged historic cooperation on AI safety. In August 2025, the UN General Assembly unanimously established an Independent International Scientific Panel on AI and a Global Dialogue on AI Governance (UN General Assembly, 2025). Six months later, eighty-eight countries and international organisations adopted the New Delhi Declaration on AI Impact: a non-binding agreement built around principles of inclusive, human-centred AI development (India AI Impact Summit, 2026). The wolf was always real. The problem was the shepherd.

Illich, I. (1976). *Limits to Medicine: Medical Nemesis: The Expropriation of Health*. Marion Boyars.

A familiar narrative frames this sequence as evidence of a governance gap: rapidly advancing technology outpacing sluggish law (Erdélyi & Goldsmith, 2022; Cath et al., 2018). There is some truth to this old story, but it is not the focus of this article. Here, I want to argue that it conceals a countervailing phenomenon that is at least as significant and philosophically intriguing. Over recent years, international governance has become crowded. The United Nations system alone now hosts multiple AI governance mechanisms. Regional instruments—most notably the European Union's AI Act—add binding regulatory layers, but multilateral declarations endorsed by numerous states are increasingly prevalent. Civil society principles, corporate responsibility charters, and multi-stakeholder frameworks multiply alongside them. The result is many stalls, much noise, and no one in charge of the market. The main point of this article is that institutional proliferation and the accompanying inflation of soft-law norms create a specific governance pathology, a typical crying wolf effect. Repeated signals of transformative urgency, without enforcement convergence, gradually diminish the normative power of each successive signal.

The argument has an ethical dimension. A governance system that only seemingly offers protection—by creating discursive artefacts such as principles, declarations, and frameworks—without establishing the institutional conditions necessary for their enforcement, does not place its recipients in the same position as a governance system that never existed. It could even harm them by fostering a false sense of security, which reduces political pressure for real regulation because the appearance of governance replaces its actual substance. Arguably, the populations most at risk from AI face the consequences of that deception. I have examined related malpractices—ethics shopping, ethics bluewashing, ethics lobbying—elsewhere (Floridi, 2019), but their systemic interaction within a regime complex has received less focus, and the 'crying wolf' effect remains a pressing issue.

The argument unfolds in four stages. The international governance of AI forms a regime complex without hierarchy (Section 2). The soft-law instruments that dominate it—which I classify as either *iterative* or *substitutive*—are prone to normative inflation (Sections 3–4). Their interaction creates the *crying wolf dynamic*, a problem of performative credibility separate from both *legitimacy* and *legality* (Section 5), and a unique ethical failure known as *governance illusion* (Sections 6–7).

Before starting, let me add three clarifications. First, the main claim about normative inflation and the crying wolf dynamic builds on ideas I developed earlier, especially the progression from self-regulation to hard law (Floridi, 2021) and the taxonomy of ethical malpractice (Floridi, 2019). I do not mean to suggest that soft law is inherently flawed. I have supported it before because it plays important roles, especially in quickly evolving fields (Abbott & Snidal, 2000). My point is conditional: soft law's effectiveness relies on conditions that are now being weakened by the overwhelming amount of soft-law efforts.

Second, I do not claim that the crying wolf effect is unique to AI governance. On the contrary, similar dynamics have been observed in climate governance (Keohane & Victor, 2011) and international human rights monitoring (Hathaway, 2002). The most similar precedent may be nuclear governance—another general-purpose technology with dual-use features and large power imbalances—where fragmented arrangements eventually consolidated into a regime with real institutional influence (the IAEA, 1957; the Non-Proliferation Treaty, 1968), although the process took decades from the first nuclear demonstrations in 1945 and was driven by proliferation crises rather than a single event. Later in the article, I respond to the objection that normative inflation is simply the normal initial cost of governance regimes.[1]

Finally, AI governance is instructive because the proliferation has occurred over an unprecedentedly short period, and the underlying technology evolves rapidly enough to make institutional sluggishness acutely visible.

2. Regime Complexity Without Hierarchy

A *regime complex* (Raustiala & Victor, 2004) is a governance structure where multiple, partially overlapping institutions address the same issue area without a clear hierarchy. This complexity creates strategic opportunities for forum- and norm-shopping (Alter & Meunier, 2009; Keohane

[1] Two structural features distinguish the AI case from historical precedents. First, pace: the nuclear regime complex consolidated in response to technologies whose capabilities evolved over decades, giving institutions time to catch up; AI capabilities evolve on a timeline of months to years, meaning the lag between norm-formation and potential consolidation is far shorter relative to the technology's evolution. Second, the strategic dimension: the structural and strategic channels of inflation identified in Section 5 interact in a way that does not merely delay consolidation but actively prevents it since powerful actors have both the incentive and the institutional means to perpetuate fragmentation indefinitely. The crying wolf effect has the properties of a stable equilibrium: the actors best positioned to drive consolidation are those with the least incentive to do so.

& Victor, 2011). Here, I use the term in an expanded sense to include private-governance instruments, given AI governance's particular reliance on private actors. The international governance of AI fits this model with remarkable precision, as recent empirical and theoretical work confirms (Tallberg et al., 2023; Geith et al., 2025; Maas, 2025). Normative authority over AI is claimed by developing companies, human rights bodies, trade regimes, security institutions, development agencies, standards organisations, ethics advisory panels, and a growing number of ad hoc and self-appointed entities. Each operates under a different mandate, engages a different constituency, and employs a different normative vocabulary.[2] No single institution has the jurisdiction, let alone the capacity, to override or coordinate the others.

Regime complexity is not inherently problematic. In some cases, it promotes experimentation and redundancy (Raustiala & Victor, 2004). However, its benefits rely on at least some institutional interaction and mechanisms through which overlapping mandates are negotiated and normative consistency is maintained. In AI governance, such interaction remains in its early stages. Listing initiatives—as the UN's various AI resource hubs do—provides information about the regime complex without establishing hierarchy within it. The point connects to a question at the heart of the philosophy of technology, at least since Winner (1980): how does a governance system acquire practical authority over the sociotechnical systems it purports to regulate?[3] Winner's insight was that technologies are not passive objects awaiting regulation but active structuring forces that shape the conditions under which governance operates. Applied to AI, this means that the current architecture of AI development—concentrated in a small number of corporate actors, dependent on immense investments, large-scale compute infrastructure, and structured by intellectual property regimes and talent markets—creates power asymmetries that enable powerful actors to exploit governance fragmentation. A general-purpose[4] technology like AI intensifies this dynamic. Its applications are highly diverse, its capability frontier evolves faster than any institutional committee can convene, and its effects impact populations unevenly, who vary greatly in their capacity to demand regulatory attention. Governing AI is not like regulating a bridge or a dam; it is more akin to managing a financial system, an energy grid, or any other infrastructure that permeates every area of social life but can only be governed effectively through

[2] Constructivist accounts of norm diffusion in international relations—e.g., Finnemore and Sikkink (1998)—offer a complementary perspective on how informal norms constrain state behaviour. My concern here is not with how normative constraint is established but with the conditions under which it erodes.

[3] The implication, which Winner's argument supports even if he does not develop it in these terms, is that the politics of artefacts can be self-reinforcing: once a technology is designed and deployed in a particular way, it creates constituencies and entrenched patterns that make alternative arrangements difficult. The current architecture of AI development is an active force that shapes governance in its own image.

[4] In the sense of Bresnahan and Trajtenberg (1995): pervasive, capable of sustained improvement, and generative of complementary innovations.

a robust institutional framework. A further challenge is that governance of AI systems (applications deployed in specific contexts, as the EU AI Act primarily addresses) and governance of AI capabilities (the inherent properties of models, regardless of deployment, as in frontier safety discussions of dangerous capabilities) require distinct regulatory structures, expertise, and enforcement mechanisms. The regime complex conflates them.

Governance authority, as Raz (1986) argued in a different context, is not solely a result of formal delegation but also depends on the service conception: an authority is legitimate to the extent that its directives enable recipients to comply better with the reasons already applicable to them (the moral and practical demands they already face). The thesis concerns individual authorities but extends to the systemic level. Even if individual institutions satisfy the service conception in isolation, the system as a whole may fail it. A regime complex where every institution claims urgency but none commands obedience fails the service conception not because its individual norms are unjustified but because the system does not enable addressees to identify and act on a coherent set of demands. The regulated actor confronts a variety of frameworks to manage. The regime complex thus inadvertently empowers the very actors—states and corporations with strong interests in minimal regulation—it seeks to constrain.

I mentioned earlier that one might object that this is simply the growing pains of an emerging governance framework and that hierarchy will eventually crystallise, as it has in areas like trade and environmental governance, for example. Perhaps. The indicators I develop in Section 3 offer a prospective test for distinguishing growing pains from pathology. But AI governance faces a structural challenge that those domains did not confront in quite the same way. As I have argued (Floridi, 2014), the digital transforms the conditions under which governance operates. The object of AI governance is not a single pollutant, commodity, or weapons system but a technology of radical heterogeneity that resists the definitional boundaries on which traditional regulatory regimes depend. The regime complex may not be an adolescent phase AI governance will outgrow. It may be a chronic condition and, as I will argue in Section 7, a co-productive one in Jasanoff's sense: the fragmentation of governance does not merely fail to regulate AI but actively helps to make AI ungovernable. If so, the philosophical task is not to wait for hierarchy to emerge but to understand the normative consequences of its absence and to prevent it from becoming a chronic problem.

3. Soft-Law Saturation and Normative Inflation

Regime complexity, as outlined earlier, is a structural condition. Normative inflation is its dynamic outcome: the regime complex produces a steadily growing amount of governance output whose

marginal normative influence wanes. Soft law—non-binding norms expressed in declarations, guidelines, and principles—holds a significant role in the philosophy of international governance. Its advocates (including myself) highlight timeliness and flexibility: when knowledge is uncertain and technology advances quickly, hard law risks premature locking-in—this is the well-known Collingridge dilemma in governance, worsened by even faster development cycles (Collingridge, 1980)—while soft law allows for iterative learning. In the initial phases of norm development, this is persuasive. However, without some binding authority, soft law relies on reputational incentives and normative legitimacy, both of which can be easily exhausted (Shelton, 2000). The philosophical question is whether there are conditions under which more provides less. The trajectory from self-regulation to hard law in the digital industry (Floridi, 2021) serves as a cautionary example: what started as a reasonable governance approach ended, through repeated non-enforcement, as an unintentional licence for inaction.

AI governance is now largely shaped by a vast array of soft-law outputs. Principles such as transparency, accountability, fairness, and safety recur across many instruments, including UNESCO, Bletchley, Seoul, Paris, New Delhi, the G7 Hiroshima Process, and various multilateral declarations. Jobin et al. (2019) documented eighty-four documents containing AI ethics principles by 2019 alone; this number has since increased significantly. Hagendorff (2020) showed that this apparent convergence conceals notable asymmetries. The most commonly expressed principles are also the most abstract, while practical, specific commitments are rare. The convergence is misleading: there is consensus on what AI governance should address, but no agreement on who enforces what, or with what consequences. Moreover, the norms being proliferated typically operate at the highest level of abstraction (Floridi, 2008), addressing surface-level behaviour without reaching the technical design choices—at lower levels of abstraction—that actually determine how AI systems function and malfunction.

Normative inflation is a helpful label for describing this dynamic, by analogy with monetary economics.[5] When a central bank increases the money supply without a corresponding rise in productive output, each unit of currency loses its purchasing power. Similarly, when normative declarations multiply without a matching increase in enforcement capacity, each loses what might be called 'normative purchasing power': its ability to influence expectations and behaviour. The

[5] In the article, I use several related terms throughout: normative purchasing power (the capacity of a declaration to alter expectations and behaviour), normative presence (the degree to which a norm occupies practical reasoning), normative authority (the general standing of a governance system to issue binding norms), and performative credibility (the accumulated expectation that pronouncements will be followed through). These are not synonyms. Normative purchasing power is the currency; normative presence is the effect of spending it; normative authority is the credit rating of the issuer; and performative credibility is the track record on which the credit rating depends.

analogy highlights a structural feature: inflation is not caused by a single act of issuance but by the cumulative effect of individually rational behaviour under insufficient coordination.[6]

Normative inflation is not a straightforward process. The initial declaration on a subject may carry considerable influence: it solidifies a consensus, establishes an agenda, and creates a discursive framework. It might serve as a basis for subsequent legislation (more on this shortly). The second declaration may even reinforce the first. However, by the tenth or twentieth iteration, the additional impact decreases sharply, especially if no binding instrument has been introduced. New declarations can become counterproductive, signalling the system's failure to progress beyond the declaratory phase. Not all soft law holds the same weight. A distinction must be made between *iterative soft law*—which advances towards hard law through successive approximation, offering normative models that eventually crystallise into binding instruments—and *substitutive soft law*, which substitutes enforcement with expression and functions as an end in itself. For example, the GDPR was preceded by decades of non-binding data protection frameworks that clearly influenced its content; the EU AI Act drew on earlier principles developed first by the AI4People initiative, then by High-Level Expert Group.[7] In those cases, soft law acted as scaffolding for a structure that was eventually built. Moreover, substitutive soft law can produce incidental benefits—such as norm socialisation, capacity building, and discursive reference points—even if these benefits come at a cost (the erosion of performative credibility) that increases with each additional declaration. However, AI governance has crossed the point where costs outweigh benefits. The 'crying wolf' effect is unique to the substitutive type: repeated declarations without moving towards a binding instrument, without building enforcement capacity, and without a clear legislative path. Currently, the international governance of AI is mostly substitutive. This argument would be disproved if successive declarations were shown to have increased, rather than decreased, the likelihood of binding follow-on agreements; the following sections explain the structural reasons why the historical record offers little support for that counter-hypothesis. The conditions that enabled the soft-to-hard transition in European data protection—such as a dominant regulatory power, a specific triggering event, and a well-defined domain—are mostly missing where the issue is diffuse, jurisdictional authority is fragmented, and the most influential actors have strong incentives to oppose binding constraint.

[6] The analogy has its limits: monetary inflation is uniform, whereas normative inflation is partly selective, since a declaration supported by credible enforcement retains its influence. In this sense, the dynamic resembles less a uniform currency depreciation and more a market for lemons (Akerlof, 1970), where the inability to distinguish enforceable from non-enforceable declarations causes addressees to devalue all declarations, though the discount is more severe for those without visible enforcement. What prevents this selectivity from being complete is the credibility spillover I describe below as legitimacy leakage.

[7] Disclosure: I designed and led the AI4People initiative and I was a member of the HLEG.

The distinction is easier to draw retrospectively than prospectively, but some observable indicators can help determine whether a soft-law instrument functions iteratively or substitutively as it is being drafted: whether it includes a sunset clause or review mechanism, whether it is explicitly framed as preparatory to a binding instrument with a specific timeline, whether it assigns specific responsibilities to identifiable actors rather than general principles to all, and whether it establishes monitoring or reporting obligations, however minimal. An instrument that lacks all four indicators is almost certainly substitutive. An instrument that possesses all four may still fail the transition if political will collapses; the indicators are necessary, not sufficient. A policymaker who cannot point to any of these features in a declaration under negotiation should recognise it for what it is: an addition to the noise, not a step towards the signal. Applied to the most prominent instruments—UNESCO, Bletchley, Seoul, Paris, New Delhi—the indicators support a predominantly substitutive assessment; only the OECD AI Principles, embedded in existing peer-review infrastructure, arguably qualify as iterative.[8]

4. Normative Presence and Semantic Deflation

The philosophical aspect of this dynamic goes beyond institutional efficiency to the very foundations of normativity. Norms influence behaviour through their propositional content and enforcement, but also through their normative presence: how they occupy the practical reasoning of those addressed, as considerations that must be acted upon rather than just recognised. Put simply, the claim is this: a norm that no one acts on, regardless of how strictly it is enforced, ceases to be a norm in any meaningful sense for practical reasoning. The idea draws on Brandom's (1994) inferentialist account: to treat a norm as binding is not merely to agree with it verbally but to adopt the practical commitments and entitlements it entails, meaning to act accordingly and to hold oneself and others accountable for compliance. The inferentialist approach leads to a more radical conclusion than a simple claim of lost credibility. For Brandom, the inferential role is not an

[8] The UNESCO Recommendation on the Ethics of AI (2021) contains no sunset clause, no timeline for a binding follow-up instrument, addresses general principles to all member states without assigning specific responsibilities to identifiable actors, and creates no monitoring mechanism: substitutive on all four indicators. The Bletchley Declaration (2023) established an international dialogue on frontier AI risk but created no institutional body, no compliance mechanism, and no follow-up timeline: substitutive. The Seoul Ministerial Statement (2024) iterated on Bletchley by securing voluntary commitments from leading AI developers but remained explicitly non-binding and assigned no regulatory responsibilities: substitutive. The Paris AI Action Summit (2025) broadened participation and emphasised inclusive development, but again without enforcement architecture or binding follow-through: substitutive. The New Delhi Declaration (2026), adopted by eighty-eight countries and international organisations at the first AI summit hosted in the Global South, emphasised inclusive and human-centred AI development but contained no enforcement mechanism, no binding timeline, no assigned responsibilities, and no monitoring obligations: substitutive on all four indicators. The OECD AI Principles are a partial exception: they have been operationalised through peer review processes and were explicitly referenced in the EU AI Act's legislative history, making them arguably iterative. But the OECD Principles succeeded because they were embedded in an institutional context—the OECD's existing peer review infrastructure, the EU's legislative competence—that most instruments of international governance of AI lack.

incidental feature of a norm; it is essential to its semantic content. A norm from which no one draws practical conclusions undergoes semantic deflation—not of its propositional content (everyone still understands what 'AI systems shall be transparent' means)—but of its normative content in Brandom's sense: the practical commitments, entitlements, and accountability relations that make a norm binding. I am extending Brandom's framework beyond his original scope—his account does not explicitly separate propositional from normative content in this way—but this extension aligns with his broader inferentialism[9]: if the inferential role is fundamental to semantic content generally, then the same applies to normative content, where the relevant role is not propositional but the pattern of practical commitment, entitlement, and accountability that makes a norm binding. The practical commitments that underpin norm-following can collapse while speakers still understand the sentence. However, this normative content becomes empty because the practice of treating it as binding has disintegrated,[10] and, according to inferentialist logic, that practice is what constitutes the norm's true essence. This echoes Austin's (1962) analysis of performative utterances: a norm, like a promise, must be embedded within the right institutional context to have its intended effect. A promise made by someone who has broken every previous promise is propositionally identical to a reliable one, but no one would consider them to carry the same practical force. Normative presence relies on distinctiveness, credibility, and perceived consequence. Inflation destroys all three. A norm cannot maintain its normative presence unless the issuing institution retains performative credibility; when this collapses, the norm ceases to influence practical reasoning, even though its formal status remains unchanged. I have argued elsewhere that soft ethics—the post-compliance space where actors choose to exceed legal requirements (Floridi, 2018)—depends on a background of institutional credibility: it presupposes that the normative framework is taken seriously enough that going beyond it remains meaningful. Normative inflation erodes and can ultimately obliterate that presence background.

5. The Crying Wolf Effect as a Problem of Performative Credibility

Section 4 identified what is lost: *normative presence*. The task now is to identify the institutional mechanism whose failure explains the loss. Let us call this mechanism *performative credibility*. Its governance application (introduced in Section 1) can now be given a clear philosophical

[9] One might object that on a thoroughgoing inferentialism, propositional content should also deflate when inferential role collapses. I disagree with this because speakers retain the capacity to draw propositional inferences from the sentence—they can explain what transparency means—while losing the practical commitments that make it normatively binding. The deflation is selective, affecting the normative layer while leaving the descriptive layer intact.

[10] Levitsky and Ziblatt (2018) document an analogous dynamic in democratic governance: the informal norms that sustain democratic institutions can erode while their formal language persists—precisely the tension between propositional and normative content identified here.

characterisation. Aesop's fable offers the label, but the governance dynamic it describes is structurally different from the fable and more problematic. In Aesop's story, the shepherd boy gives false signals: there is no wolf, the villagers learn that the source is unreliable, and when the wolf truly appears, the truthful signal is disregarded. The credibility failure relates to the truthfulness of the signal. In the governance case I am examining, the signals are not false. The wolf is real and has been real each time. What the signalling institutions fail to produce is not truth but consequence (enforcement, sanctions, material follow-through). The villagers stop responding, not because they doubt the wolf's existence but because they have learned that no one will act on the warning. The credibility failure concerns the institutional capacity to respond, not the epistemic reliability of the alarm. The result is the same—the wolf remains unaddressed—but the mechanism is different, and the moral issue is arguably more serious: addressees are rationally justified in their inaction given the institutional track record, even though the threat they are ignoring is genuine. Translated into governance terms: repeated urgency signalling that fails to produce observable institutional consequences weakens the action-guiding force of future signals, regardless of their normative validity. The shepherd is telling the truth. The problem is that no one expects him to do anything about it.

Risk-communication research confirms common sense. Breznitz (1984) showed that repeated warnings not followed by experienced consequences decrease attention and behavioural responsiveness. This finding has been replicated across disciplines, from tornado warnings to cybersecurity alerts, and the mechanism is well understood. Agents learn that the warning system is unreliable each time a warning is issued without any observable consequence. In governance, the mediating factor is not the direct experience of a threat but the observation of an institutional response. When declarations are not followed by enforcement or material consequences, addressees quite rationally infer that the system lacks either the capacity or the will to act on its own pronouncements.[11] This dynamic can be articulated more precisely in philosophical terms by distinguishing three properties often conflated in governance theory: *legitimacy*, *legality*, and *performative credibility*:

a) *Legitimacy* concerns the right to govern, that is, the normative justification for an institution's authority, grounded in consent, procedural fairness, or output quality (Buchanan & Keohane, 2006).

[11] The model applies most directly to informed addressees—regulated actors, civil-society organisations, policymakers who follow institutional responses closely enough to register the gap between declaration and enforcement. For the general public, the mechanism operates differently: the inference is not from observed non-enforcement to reduced credibility but from observed institutional activity to assumed effectiveness. The two pathways produce opposite outcomes—scepticism among the informed, misplaced confidence among the uninformed—and both contribute to the governance pathology, though through different channels.

b) *Legality* concerns the formal bindingness of a norm within a recognised legal system.

c) *Performative credibility* concerns the extent to which an institution's pronouncements are treated as action-guiding signals that will be followed. A governance declaration is a performative utterance. We saw that Austin is clear that a promise by a known liar is an abuse, not a misfire. The illocutionary act succeeds; it is the perlocutionary uptake that fails. Depleted-credibility declarations succeed as speech acts while failing to alter behaviour or generate compliance. Performative credibility is therefore a *perlocutionary concept*: the accumulated expectation, grounded in institutional track record, that a governance pronouncement will produce the behavioural uptake it purports to demand. What Searle (1995) calls the system's capacity to generate *status functions*—to make things count as obligations, violations, or compliance—depends not only on collective acceptance but on demonstrated willingness to maintain those obligations.

From (c), it follows that the crying wolf effect is a crisis of performative credibility. Institutions may possess considerable convening power and legitimacy—multilateral endorsement, expert composition, inclusive process—and their soft-law norms may even be moving towards customary legal status. But if the system repeatedly frames AI governance as historic and urgent while failing to consolidate enforcement, performative credibility diminishes with each intervention. Three mechanisms drive this erosion, and they interact to make the dynamic self-reinforcing:

i) *Attention fatigue*. Agenda saturation diminishes the cognitive importance given to any single governance signal. When every summit is declared decisive and every declaration is described as historic, the language of urgency becomes exhausted. Far from being a communication failure, this is a structural result of a regime complex where each institution is motivated to amplify the prominence of its own output. The infosphere itself is a commons, and its normative aspect is no exception: what I called the tragedy of the digital commons (Greco & Floridi, 2004)—the deterioration of a shared digital environment through individually rational exploitation—has its exact analogue in the governance arena, where the shared resource being depleted is not bandwidth but normative credibility.

ii) *Legitimacy leakage*. Institutional activism that results in no tangible change in the behaviour of powerful actors fosters a perception of performative governance, as a spectacle rather than a genuine constraint (Jasanoff, 2016). Once this perception is established, it is difficult to reverse and may spill over, undermining confidence in governance institutions more broadly. The leakage is not contained; scepticism about AI governance can erode trust in technology governance overall, which is a dangerous spillover in an era when public trust in institutions is already under severe strain.

iii) *Incentive stagnation.* In the absence of binding obligations, the payoff structure faced by regulated actors remains unchanged. Rational actors respond accordingly: they participate in declaratory processes—since participation is free and refusal is reputationally risky—while making only minimal adjustments to their substantive behaviour. The outcome is an equilibrium where everyone talks the language of responsible AI governance, but the actual distribution of power, risk, and benefit stays the same.

The dynamic is not hypothetical. The digital industry's broader encounter with self-regulation followed the same sequence over 15 years (Floridi, 2021). Between 2004 and 2019, a series of self-regulatory initiatives—Facebook's codes of conduct, Google's Advisory Council on the right to be forgotten in 2014 (Floridi, 2015), the proliferation of corporate AI ethics boards culminating in Google's short-lived Advanced Technology External Advisory Council (2019)—generated attention fatigue, legitimacy leakage, and incentive stagnation in exactly the order described above. Each initiative attracted attention; unfortunately (and much to my regret[12]), none really changed corporate behaviour at the C-suite level. By the time the industry had produced hundreds of codes, guidelines, and manifestos, self-regulation had revealed itself in all its embarrassing vacuity. The remedy came not from better principles but from hard law: the GDPR, the Digital Markets Act, the Digital Services Act, and the AI Act. *Dura lex, sed lex digitalis*. The international governance of AI is now following the same pattern, but at the intergovernmental level and on a compressed timeline. The disanalogy is real: states face different incentive structures from corporations, and the enforcement pathway that worked domestically within the EU may not transfer to the international level. But the mechanism—attention fatigue, legitimacy leakage, incentive stagnation—operates at both levels, even if the exit routes differ.

The loop is self-reinforcing and operates through two distinct channels, which are worth analysing separately. The first is structural normative inflation: the unintended, emergent result of individually rational institutional behaviour. Each organisation has incentives to demonstrate relevance by producing normative output, driven by FOMO and the 'me too' effect. No single actor aims to undermine governance credibility. However, collectively, they generate the saturation that damages it. The second channel is strategic normative inflation: the intentional exploitation of regime fragmentation by powerful actors who benefit from a weak normative environment. The ethics of shopping, bluewashing, and lobbying identified in the introduction (Floridi, 2019; 2021) are not innocent side effects of institutional complexity but strategies through which well-resourced actors actively sustain the conditions where soft law replaces binding constraints.interested in its repair. The relationship between the two channels matters for the

[12] More disclosure: I was an advisor or a member in all the named initiatives. *Mea culpa.*

prescriptive argument. Institutional consolidation addresses the structural channel directly, by reducing the incentive for competitive normative output, and the strategic channel indirectly: fragmentation is a necessary condition for forum-shopping and ethics washing, and a consolidated architecture with credible consequences is harder to circumvent.[13] These two channels interact in a way that makes the dynamic self-concealing. The structural pathology provides cover for the strategic one: a non-binding declaration with broad principles and no enforcement mechanism appears identical, regardless of whether it results from genuine collective-action failure or strategic manipulation. The same stall may be selling honest goods or counterfeit ones, and the buyer cannot tell the difference. The result is a system that benefits those least

The analysis would be incomplete without noting that the crying wolf dynamic operates in private governance with at least as much force as in public. A great deal of *de facto* AI governance is conducted by private actors—model-release policies, safety protocols, responsible-scaling commitments—whose instruments share the structural features of their public counterparts (voluntary adoption, self-assessed compliance, no external enforcement) and are vulnerable to the same inflationary logic. Corporate responsibility charters may be just as substitutive as intergovernmental declarations, and their proliferation may generate the same false sense of security. The interaction between the two layers compounds the problem: private governance can substitute for public regulation in much the same way that soft law replaces hard law, and the actors most actively engaged in strategic normative inflation—frontier AI companies and the industry associations that represent them—are those with the greatest capacity to shape private governance in their own interests while citing it as evidence that public regulation is unnecessary. A governance architecture that fails to specify the relationship between public regulatory authority and private governance leaves the most consequential decisions to the actors with the strongest incentives to avoid constraint.

6. The Governance Illusion: Ethical Dimensions of Normative Inflation

Let me now turn to what I see as the most significant aspect of the issue: its ethical dimension. The crying wolf effect does not leave the world unchanged. It actively worsens the situation and unevenly distributes costs. I refer to this as the *governance illusion*. A governance system that produces the superficial symbols of regulation—endorsed principles, convened committees, signed declarations—without establishing the necessary institutional conditions for enforcement alters the political landscape. The mechanism is displacement: governance activities absorb

[13] The dynamic described here has structural affinities with regulatory capture (Stigler, 1971; Carpenter & Moss, 2013): normative inflation creates conditions favourable to capture by enabling governance shopping and regulatory arbitrage, allowing the regulated to give the impression of compliance while avoiding binding constraint.

political energy and advocacy resources that could otherwise be channelled into binding instruments, while the declarations serve as rhetorical cover, allowing policymakers to report progress in governance without delivering substantive change. Its effects vary depending on the audience. For political leaders and the general public—those who do not closely follow AI governance—the vigorous output of governance measures sustains the reasonable belief that someone, somewhere, is overseeing the issue. The inference that institutional activity equates to effectiveness is valid in most governance areas, but it is false in this particular case. For expert communities, civil-society organisations, and those most directly affected by algorithmic decision-making, the mechanism functions differently. These groups are often acutely aware that declarations offer no real safeguards. For them, the governance illusion produces not belief but displacement: governance processes deplete their limited advocacy resources—time on advisory panels, energy spent drafting principles, political capital used for endorsements—and the resulting declarations serve as rhetorical cover for inaction, since demands for binding regulation are met with the response that governance is already in progress. The proliferation of governance activity acts as a political analgesic: it dulls the pain of regulatory absence without addressing the root problem. The dynamic is iatrogenic in a precise sense: a governance system that produces the very harm it purports to prevent—not through malice but through the structural counterproductivity of its own proliferation.[14] The centre's belief in this illusion is maintained because the costs of non-enforcement fall on communities far from governance centres and production nodes. The illusion is not uniform; it is upheld where power resides and perceived as absence where power is needed.

This point has a structural parallel in Hathaway's (2002) contested but instructive finding that ratification of human rights treaties, under specific conditions, can be associated with worse human rights outcomes—a paradox she explains by the absence of domestic enforcement mechanisms—precisely the institutional deficit I diagnose in this paper. The expressive benefits of ratification (signalling commitment, enhancing reputation) substitute for the instrumental obligations. The same holds for AI governance declarations: they express governance without enacting it.

The distributive aspect of this ethical failure deserves emphasis. The costs of the governance illusion fall disproportionately on those with the least ability to protect themselves through private measures—whether vulnerable communities within wealthy nations or less powerful countries lacking the regulatory capacity and market leverage to shape governance in their own interests—such as forum shopping, contractual safeguards, and the market power

[14] The concept of institutional counterproductivity originates with Illich (1976), who argued that institutions beyond a certain threshold of growth begin producing the harm they were designed to prevent. His examples were medicine, education, and transport; AI governance fits the pattern with uncomfortable precision.

available to well-resourced actors. A multinational technology corporation confronting a fragmented regime complex can choose the most lenient jurisdiction, employ regulatory strategists, and influence norms through self-regulation. Conversely, a community in the Global South subjected to algorithmic decisions in credit, policing, or public service distribution has no such option. The governance illusion suggests that principles of fairness, transparency, and accountability apply there. However, the institutional reality is that no one will ensure they do. This pattern exists both within wealthy nations and between them. For example, Australia's Robodebt scheme—an automated debt recovery system whose income-averaging method was declared unlawful and which caused 'severe distress' to hundreds of thousands of welfare recipients (Rinta-Kahila et al., 2022; Royal Commission into the Robodebt Scheme, 2023)—the Dutch SyRI case, where a welfare fraud detection system using nationality and postcode as risk indicators was struck down by the Hague District Court in 2020 for breaching the European Convention on Human Rights (Rachovitsa & Johann, 2022; van Bekkum & Borgesius, 2021), and predictive policing systems in the United States and the United Kingdom, whose feedback loops systematically direct enforcement towards low-income and minority communities (Lum & Isaac, 2016; Richardson et al., 2019), all took place in jurisdictions that had formally adopted AI ethics principles. These principles are on paper; real protection is absent. One could argue that unenforced norms still offer discursive leverage, or at least provide a vocabulary of accountability that civil society can use in litigation. Discursive leverage is real. But it is not equivalent to governance, and considering it as such is the very substitution this paper diagnoses.

There is a justice aspect to normative inflation. The literature on algorithmic fairness has rightly focused on how AI systems can reproduce and reinforce structural inequalities (Binns, 2018; Floridi, 2023; Selbst et al., 2019). However, the governance framework meant to address these inequalities can itself become a tool of structural injustice—structural in Young's (2011) sense: injustice that arises not from individual misconduct but from institutional processes that systematically put specific groups at a disadvantage—if it offers protection in name but fails to deliver it in practice. The proliferation of principles might actually worsen the original problem: initially, the technology creates the disparity; subsequently, the governance system gives the illusion of addressing it.

The ethical analysis raises a harder question: moral responsibility in collective action problems. If the "crying wolf" effect is an emergent outcome of individually rational institutional behaviour, who is responsible for its harms? The usual tools for moral responsibility—intention, foreseeability, causation—struggle with emergent outcomes, even though the collective effect is foreseeable (Shelton, 2000; Hathaway, 2002). The appropriate model is collective moral

responsibility (Kutz, 2000; Floridi, 2016): the institutions within the regime complex share a joint responsibility for maintaining the conditions under which their collective normative output continues to carry authority. The claim rests on a premise: the institutions in question—and, more precisely, the member states that mandate them—retain discretionary authority over the volume, specificity, and timing of their pronouncements; that they choose to contribute rather than restrain grounds their responsibility. The objection that each institution's individual contribution is negligible does not defeat the claim: the responsibility is collective precisely because the harm is emergent, and the duty is to coordinate restraint rather than to refrain individually.

I should add that the responsibility extends to the epistemic communities that lend intellectual credibility to the governance apparatus. Their presence lends epistemic authority to instruments that lack enforcement capacity, and their willingness to engage in successive rounds of principle-drafting helps sustain the impression that the governance system is making progress. In the iterative mode, expert involvement can shape norms that eventually bind. In the substitutive mode, it risks complicity by lending the prestige of independent scholarship to governance appearances. The charge is strongest in retrospect, once the substitutive character of a process has become apparent; prospectively, it imposes a duty of vigilance: a responsibility to monitor whether one's participation is contributing to iterative progress or to governance theatre, and to withdraw when the evidence favours the latter. The indicators from Section 3—sunset clauses, binding timelines, assigned responsibilities, and monitoring obligations—provide a practical test: an expert who finds none incorporated after successive drafts has grounds to conclude the process is substitutive. The uncomfortable implication is that the ethics community's eagerness to be 'at the table' may itself be a mechanism of the governance illusion it should be diagnosing.

7. Toward Institutional Gravity: Design Criteria for Consolidation

If the central challenge in the international governance of AI is not normative innovation but the erosion of normative authority—and if this erosion carries ethical costs falling most heavily on the least powerful—then the question becomes what a non-inflationary governance architecture would require. I call this requirement institutional gravity: the capacity of a governance architecture to command compliance by concentrating authority rather than dispersing it. The metaphor is physical: gravity is not a property of any individual body but an emergent effect of sufficient mass. A governance system acquires institutional gravity when it accumulates enough delegated competence, epistemic integration, and enforcement credibility to bend the behaviour of actors into its orbit. Without it, norms drift: beautiful, well-formulated, and weightless. The polycentric alternative—the claim, associated with Ostrom (2010), that overlapping governance can be more

adaptive than hierarchical consolidation—fails in the AI case because it depends on conditions the AI governance complex does not satisfy.[15] Polycentric governance succeeds when participants share compatible goals and when mechanisms of shared monitoring and graduated sanctions are established. In AI governance, the asymmetry involves not only interests but also the capacity to shape the governance framework itself, with polycentricity serving as a strategic resource for the powerful rather than an adaptive feature of the system.[16]

Institutional gravity, as I understand it, requires at least three properties. These are not an institutional blueprint, which would require a confidence in institutional forecasts I do not possess, but philosophical design criteria. The substantive content of such instruments—capacity-building provisions, dispute resolution mechanisms, compliance review procedures—lies beyond the scope of this article but represents the natural next step for institutional design.

The first is delegated competence. Some institution or coordinated set of institutions must have a clear, recognised mandate to act: to set standards with practical consequences for non-compliance. This need not be a single international regulator; federated models with mutual recognition are possible and perhaps more realistic. But some locus of decisional authority must be identifiable. The difference between an advisory panel and a regulatory body is not one of expertise; it is one of illocutionary force. When an advisory panel recommends, it expresses a view; when a regulator mandates, it imposes an obligation: a distinction that the philosophy of speech acts shows to be constitutive of institutional reality (Searle, 1995).

The second is epistemic integration. Given AI's cross-disciplinary nature, governance must be capable of integrating knowledge across domains—e.g., safety, rights, economics, security—without fragmenting into silos that each set their own norms in isolation. This is partly an institutional design challenge and partly an epistemological one: it requires governance structures that can combine diverse forms of expertise under conditions of irreducible uncertainty (Jasanoff, 2004). The temptation to create a separate expert body for each dimension of AI governance—one for safety, another for rights, a third for economic impact, a fourth for military applications, and so forth—is understandable but ultimately counterproductive: silo governance of a general-purpose technology produces the fragmentation that the crying wolf dynamic exploits. What is

[15] Ostrom's (2010) analysis presupposes a degree of institutional cooperation that the AI regime complex conspicuously lacks. Where common-pool resource governance can rely on repeated local interactions to sustain trust, the international governance of AI involves actors whose interactions are episodic, asymmetric, and mediated by incompatible institutional mandates.

[16] The metaphor has a further limitation worth recognising: unlike physical gravity, which affects all bodies equally, institutional gravity must be differentiated and directional. It must exert a stronger pull on powerful actors with the greatest ability to evade restraint, and a lighter pull on vulnerable populations and smaller actors that require protection rather than compliance burdens. An institutional gravity that oppresses the weak while leaving the strong untouched would be worse than weightlessness.

needed instead is closer to a synthetic approach to governance (Floridi, 2014)—one that treats the governed system as an integrated whole rather than a set of separable risk categories—which understands the object of governance not as a collection of separable risks but as a sociotechnical system whose properties emerge from the interaction of its components. A deeper point connects to Jasanoff's (2004) idiom of co-production: governance norms and the technologies they regulate are interconnected. Fragmenting governance into separate risk categories—bias here, privacy there, safety elsewhere—actively fosters a fragmented understanding of AI. The regime complex helps to establish AI as ungovernable. The circularity creates a self-reinforcing loop: fragmentation leads to a fractured understanding, which makes integrated governance seem impossible, which in turn justifies further fragmentation. The crying wolf effect thus has a co-productive aspect: it undermines not only governance credibility but also the epistemic conditions necessary for effective governance.

The third is credible consequences. Norms must be linked to observable outcomes: monitoring, reporting, review, and where necessary, sanctions. The specific form of consequence is less important than its credibility: addressees must have reason to believe that the system will act on its own pronouncements.[17] The crying wolf effect goes beyond mere empty talk: governance declarations are normative acts, not just informational signals, and their failure is normative, not simply epistemic. Technical governance mechanisms—model evaluations, safety benchmarks, responsible-scaling policies—operate on a different logic: their credible consequence is embedded within the mechanism rather than relying on institutional follow-through. However, they address capability risks rather than the broader concerns (fairness, accountability, democratic control) that normative instruments aim to cover. The eloquence of the declaration is irrelevant; what matters is whether the declaration has teeth, or at least the plausible prospect of developing some. Without credible consequences, the governance system remains ethically deficient: its ambitions are unmatched by its capacity to deliver.

These criteria can be applied to existing institutional proposals. For example, the UN's newly created International Scientific Panel on AI (UN General Assembly, 2025) meets the epistemic integration criterion fairly well: it aims to synthesise knowledge across fields and deliver authoritative assessments. However, it notably lacks delegated competence (its mandate is advisory, not regulatory) and credible consequences (it has no enforcement mechanism and cannot compel disclosure, let alone impose sanctions for non-compliance). In short, it is yet another advisory body within a system already overwhelmed with advice. The New Delhi Declaration

[17] Compute governance—export controls on advanced semiconductors and reporting requirements for large-scale training runs—represents a rare example of AI governance that operates through material constraints rather than normative declarations, and its rapid development illustrates what credible consequences look like in practice.

(India AI Impact Summit, 2026)—the first such instrument shaped by the Global South—scores no better: no sunset clause, no binding timeline, no assigned responsibilities, no monitoring obligations. Conversely, the EU AI Act has the legal framework for credible consequence—binding obligations, designated authorities, material penalties—though its enforcement track record remains uncertain, and grants delegated competence to the European AI Office. Yet, its jurisdictional scope is limited, its extraterritorial effects predominantly work through market access rather than regulatory authority, and regulatory protections are concentrated in one region while risks are globally distributed. All three are genuine steps forward, but none fully meets all three criteria, highlighting the gap between current reality and what is required. The fact that the only practical example of successful iterative soft law discussed here—European data protection—relied on a dominant regulatory actor, a specific catalyst, and a bounded domain is not incidental. It suggests that the iterative approach may be structurally unavailable at the international level for AI. The design criteria outlined are thus offered not as a blueprint to replicate the EU model but as abstract conditions any governance effort must meet, regardless of the institutional form it ultimately takes.

These criteria are demanding. A policymaker might argue that fragmentation stems from genuine disagreements among major powers—about state-market relations, surveillance, innovation versus caution—not from poor institutional design. This objection is serious. If the 'crying wolf' effect were simply a coordination failure, institutional design could resolve it. Under strategic divergence, it is only a partial solution: selective consolidation is still possible. Narrow agreements on specific risks—such as catastrophic AI safety—where even rival powers share an interest in preventing existential threats are more feasible than overall frameworks, and each can help generate the institutional gravity that this paper identifies as lacking. Partial gravity is better than none. The analysis of collective action indicates that consolidation is unlikely to arise from within the complex of regimes itself: the institutions that need to control their output are the same ones that benefit from producing it. The most probable catalyst is a state or coalition with enough market power to make regulatory compliance unavoidable and enough institutional capacity to establish enforcement infrastructure. The EU's regulatory influence through the Brussels effect (Bradford, 2020) provides one example, though its jurisdictional limits are real. Alternatively, consolidation prompted by a catalysing event—such as the nuclear case—is more probable but less desirable.

Institutional gravity need not be embodied in a single international organisation. It might develop through structured interoperability—mutual recognition, regulatory equivalence, coordinated enforcement—among national regimes that already impose credible consequences.

The challenge I have faced before remains: governance of digital environments exists in a logical space that defies the territorial assumptions of the Westphalian model (Floridi, 2014), and AI governance fully inherits this mismatch. The raw material exists; what is missing is the commitment to connecting it. The philosophical principle remains independent of political prediction: a governance system unable to meet these criteria will deepen normative inflation. The real question is not whether the international community desires effective AI governance—its rhetoric indicates as much; repeated declarations attest to that—but whether it is willing to build the institutional infrastructure required. The difference between aspiration and actual architecture is akin to the difference between drawing a bridge and constructing one. Only the latter is truly functional weight.

8. Conclusion

The international governance of AI is suffering from normative inflation within a regime complex that lacks hierarchy and enforcement capacity. The crying wolf effect—a progressive erosion of performative credibility through repeated urgency signalling absent enforcement—is a structural threat and an ethical failure whose costs fall on the least powerful. The path forward does not lie in further multiplication of principles, panels, or declarations. It lies in the unglamorous but essential work of institutional consolidation: concentrating authority, integrating expertise, and linking norms to credible consequences. Whether the international system is capable of such consolidation amid geopolitical fragmentation is an open question. That it needs to attempt it is not. I argued in 2021 that the digital industry's self-regulation era had ended and that hard law had to replace it (Floridi, 2021). The argument of this paper extends that lesson to the international level: the crying wolf effect is not confined to one industry's encounter with voluntary codes; it is the systemic pathology of any governance architecture that substitutes declaration for enforcement.

The critique that this conclusion is both obvious and unhelpfully vague may not be entirely unfair. However, the response is that the contribution lies in the diagnosis and in the conceptual framework that makes it precise. Normative presence, performative credibility, and the iterative/substitutive distinction provide tools for analysing governance issues beyond AI. The tendency in academic and policy discourse is to treat each new AI governance initiative as a step forward. My argument is that, under current conditions, each such step may actually push the system backwards by depleting the normative resource—credibility—on which future governance authority relies. Norms do not disappear; they lose their normative presence, remaining as propositions but ceasing to function as considerations in practical reasoning. Ethical analysis

reinforces this point: the governance illusion becomes a mechanism by which structural injustice worsens, those at risk are told they are protected, while the powerful remain unrestrained. The broader suggestion is that the ethics of governance design deserve as much philosophical attention as the ethics of the technologies being regulated. There is a sophisticated literature on algorithmic fairness, transparency, and accountability (Floridi, 2023). But there is not yet an equally sophisticated literature on the conditions under which governance systems tasked with upholding those values actually operate effectively. The philosophical framework is not ornamental. The inferentialist approach makes the claim of norm erosion precise rather than merely descriptive; the governance illusion analysis reframes non-enforcement as active harm rather than missed opportunities; the collective-action framework identifies the structural beneficiaries of fragmentation. It might seem ironic that a paper diagnosing normative inflation is itself a normative intervention, but I would argue that the diagnosis is a different kind of speech act from the declarations it critiques, since it does not propose further principles but advocates restraint in the creation of principles, which is a second-order intervention that does not contribute to the first-order inflation.

Recognising the crying wolf dynamic does not tell us what to build, but it does indicate what to stop doing, and sometimes that is the more urgent lesson. It also provides a practical diagnosis: any governance mechanism lacking a sunset clause, a binding timeline, responsibilities assigned to identifiable actors, and monitoring obligations is almost certainly substitutive. The apparent paradox with which this paper began — that more governance can lead to less governance — is not an actual paradox. It is the predictable result of institutional structures that incentivise normative output without consolidating normative authority. Flawed philosophy mistakes the map for the territory. Flawed governance mistakes the declaration for the deed.

Acknowledgements: I am grateful to Jessica Morley and Neo Hui Yuan for their insightful and informative comments on an earlier draft. I am the only person responsible for any remaining mistakes.

Conflict of interest

The author declares no conflict of interest.

References

Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization*, 54(3), 421–456.

Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.

Alter, K. J., & Meunier, S. (2009). The politics of international regime complexity. *Perspectives on Politics*, 7(1), 13–24.

Austin, J. L. (1962). *How to Do Things with Words*. Clarendon Press.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: 'Engines of growth'? *Journal of Econometrics*, *65*(1), 83–108.

Breznitz, S. (1984). *Cry Wolf: The Psychology of False Alarms*. Lawrence Erlbaum Associates.

Buchanan, A., & Keohane, R. O. (2006). The legitimacy of global governance institutions. *Ethics & International Affairs*, *20*(4), 405–437.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics*, 24(2), 505–528.

Carpenter, D., & Moss, D. A. (Eds.). (2013). *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press.

Collingridge, D. (1980). *The Social Control of Technology*. Frances Pinter.

Erdélyi, O. J., & Goldsmith, J. (2022). Regulating artificial intelligence: Proposal for a global solution. *Government Information Quarterly*, 39(4), 101748.

Finnemore, M., & Sikkink, K. (1998). International norm dynamics and political change. *International Organization*, 52(4), 887–917.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines*, 18(3), 303–329.

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press.

Floridi, L. (2015). 'The right to be forgotten': A philosophical view. *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics*, 23, 163–179.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A*, *374*(2083), 20160112.

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A*, *376*(2133), 20180081.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 32(2), 185–193.

Floridi, L. (2021). The end of an era: From self-regulation to hard law for the digital industry. *Philosophy & Technology*, 34(3), 619–622.

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.

Geith, J., Lundgren, M., & Tallberg, J. (2025). The emerging regime complex for artificial intelligence. SSRN Working Paper. https://doi.org/10.2139/ssrn.5722202

G7 Hiroshima Process. (2023). Hiroshima Process International Guiding Principles for Advanced AI Systems and Code of Conduct for Organizations Developing Advanced AI Systems. https://www.mofa.go.jp/files/100573466.pdf

Greco, G. M., & Floridi, L. (2004). The tragedy of the digital commons. *Ethics and Information Technology*, 6(2), 73–81.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hathaway, O. A. (2002). Do human rights treaties make a difference? *Yale Law Journal*, 111(8), 1935–2042.

India AI Impact Summit. (2026). New Delhi Declaration on AI Impact. https://impact.indiaai.gov.in/

Jasanoff, S. (Ed.). (2004). *States of Knowledge: The Co-production of Science and Social Order*. Routledge.

Jasanoff, S. (2016). *The Ethics of Invention: Technology and the Human Future*. W. W. Norton.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. *Perspectives on Politics*, 9(1), 7–23.

Kutz, C. (2000). *Complicity: Ethics and Law for a Collective Age*. Cambridge University Press.

Levitsky, S., & Ziblatt, D. (2018). *How Democracies Die*. Crown.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19.

Maas, M. M. (2025). *Architectures of Global AI Governance: From Technological Change to Human Choice*. Oxford University Press.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3), 641–672.

Paris AI Action Summit. (2025). Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet. https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review*, 22(2), ngac010.

Raustiala, K., & Victor, D. G. (2004). The regime complex for plant genetic resources. *International Organization*, 58(2), 277–309.

Raz, J. (1986). *The Morality of Freedom*. Clarendon Press.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94, 15–55.

Rinta-Kahila, T., Someh, I. A., Gillespie, N., Indulska, M., & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems*, 31(3), 313–338.

Royal Commission into the Robodebt Scheme. (2023). Report. Commonwealth of Australia.

Searle, J. R. (1995). *The Construction of Social Reality*. Free Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM.

Seoul Ministerial Statement. (2024). Seoul Ministerial Statement for Advancing AI Safety, Innovation and Inclusivity. https://www.gov.uk/government/publications/seoul-ministerial-statement-for-advancing-ai-safety-innovation-and-inclusivity

Shelton, D. (Ed.). (2000). *Commitment and Compliance: The Role of Non-binding Norms in the International Legal System*. Oxford University Press.

Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(1), 3–21.

Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence: Next steps for empirical and normative research. *International Studies Review*, *25*(3), viad040. https://doi.org/10.1093/isr/viad040

UK Government. (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

UN General Assembly. (2025). Resolution A/RES/79/325: Independent International Scientific Panel on Artificial Intelligence and Global Dialogue on AI Governance. United Nations.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO.

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security*, 23(4), 323–340.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136.

Young, I. M. (2011). *Responsibility for Justice*. Oxford University Press.