GOVERNING DEEPFAKES: LEGAL INITIATIVES AND REGULATORY GAPS

YASAMAN YOUSEFI,[1, 2]
MARIA DOLORES SANCHEZ GALERA,[3]
ANGELO TUMMINELLI,[4] CALOGERO CALTAGIRONE,[4]
TOMMASO TONELLO[5]

[1] DEXAI-Artificial Ethics, Rome, Italy
yasaman.yousefi@dexai.eu
[2] University of Bologna, CIRSFID ALMA AI, Faculty of Legal Studies, Bologna, Italy
y.yousefi@unibo.it
[3] Charles III University of Madrid, Madrid, Spain
mariadsa@inst.uc3m.es
[4] LUMSA University, Department of Human Sciences, Rome, Italy
a.tumminelli@lumsa.it, c.caltagirone@lumsa.it
[5] Utrecht University, Freudenthal Institute, Utrecht, the Netherlands
t.tonello@uu.nl

This chapter examines the pervasive threat of digital disinformation, with a specific focus on AI-generated content as a paradigmatic challenge to contemporary governance. The analysis blends ethical and legal perspectives to assess existing mitigation strategies. AIGC occupies a critical intersection of advanced technical capability, complex social meaning-making, and often conflicting legal protection frameworks. Consequently, effective responses require an interdisciplinary approach that integrates conceptual clarity, technical standards, robust legal instruments, and widespread social interventions to preserve public trust and protect vulnerable individuals.

DOI
https://doi.org/10.18690/um.feri.2.2026.7

ISBN
978-961-299-109-8

Keywords:
Digital Disinformation,
EU Regulatory Approach,
GDPR,
AI Act
Harm Mitigation

University of Maribor Press

1 Conceptual Considerations

This section establishes the ethical and sociological context for disinformation, framing the problem of synthetic media in terms of relational responsibility and the material consequences of immaterial harms.

Conceptual clarity regarding the nature of digital communication is necessary to frame legal responses. Sociological critiques argue that the hyperconnected infosphere fosters a cultural state of "existential relativism," a condition where distinctions between truth and falsehood blur, rationality yields to emotionality, and communication operates under the premise that "anything goes" (Donati, 2024, p. 36). This phenomenon risks confusing technologies that support human identity with those that actively erode it, leaving individuals vulnerable to technological domination (Donati, 2024, p. 32).

The cultural diagnosis of "existential relativism" in techno-mediated contexts cannot remain a mere description of fragmented meanings. The pervasiveness of digital platforms destabilizes symbolic reference points and weakens shared norms. This sociological condition translates into normative challenges, requiring new forms of rule legitimation. At the same time, it generates moral challenges, expanding responsibility for actions whose consequences are diffuse. Subjectivity must therefore renegotiate criteria of autonomy and accountability. The shift toward ethical responsibility becomes a response to the volatility of digital environments. In sum, cultural diagnosis demands an ethical rethinking capable of guiding common practices.

In this sense, the concept of responsibility must be re-centred. Responsibility, in its deepest sense (Miano 2009; Da Re 2003), is not merely an individual legal commitment but a dialogical and ecological capacity to respond to the call of others and to care for the world as a shared home. The velocity and pervasive nature of AI challenge this relational commitment. The creation or sharing of deceptive content without reflecting on its impact constitutes a profound failure of this relational commitment.

When technological systems, such as hyperconnectivity and algorithmic amplification, overwhelm individual capacity for verification and responsible reflection, the individual alone cannot discharge the ethical duty of care. This creates

an ethical vacuum. The regulatory response, namely, the requirement under the Digital Services Act (DSA) that Very Large Online Platforms (VLOPs) manage systemic risks, is thus ethically justified. The state enforces the transfer of the burden of relational care from the overwhelmed individual to the systemic actors (platforms) that control the informational infrastructure. However, is this enough? In addition, how can we trust self-regulation and self-risk-management systems?

1.1 Privacy, Reputation, and the Materiality of Immaterial Harms

AI-generated contents pose direct threats to protected rights, notably privacy and reputation, by weaponizing personal data.

- **Privacy:** Privacy is the inherent right of an individual to control their personal information, linked intrinsically to dignity, freedom, and autonomy. Deepfakes violate this right by depicting individuals in false, compromising, and potentially harmful situations without consent, attacking the integrity of their self-presentation.
- **Reputation:** Reputation reflects the moral and social value attributed to a person, based on actions and perceived identity, functioning as a critical component of credibility within a community. Deepfakes inflict grave damage by distorting public perception, leading to exclusion, professional loss, and emotional distress.

The Cambridge Analytica scandal illustrates how the misuse of personal data can become a powerful instrument of manipulation and reputational harm. By harvesting the personal information of millions of Facebook users without their knowledge or consent, Cambridge Analytica exploited intimate details of individuals' preferences, vulnerabilities, and networks to influence electoral behaviour (Isaak & Hanna, 2018). This case underscores how data, once weaponized, undermines privacy and autonomy by stripping individuals of control over their own digital identities, while simultaneously reshaping collective reputations and public discourse in ways that erode trust in democratic institutions.

Deepfakes exacerbate these concerns by combining the mass-scale data misuse seen in Cambridge Analytica with highly persuasive falsifications of identity. Unlike simple data profiling, deepfakes do not just predict or manipulate preferences; they

fabricate "hyperreality". Comparable to revenge porn cases, where intimate images are shared without consent, or the proliferation of deepfake pornography targeting women in public life, these manipulations inflict enduring reputational damage that cannot be easily corrected once the falsified content circulates (Chesney & Citron, 2018). Similarly, instances where politicians or journalists are targeted with synthetic media, such as the 2019 deepfake video of Nancy Pelosi manipulated to make her appear intoxicated, demonstrate how fabricated content erodes public trust, polarizes societies, and destabilizes democratic debate (Reuters, 2020).

Critically, the harms inflicted by deepfakes are often immaterial: psychological distress, reputational degradation, and erosion of evidentiary trust. While these harms are not physical or pecuniary in the traditional sense, they carry severe material consequences (e.g., job loss, social ostracization). This profile presents a critical remedial gap. Current liability frameworks, including the revised Product Liability Directive (PLD), remain primarily oriented toward material or pecuniary damages, rendering the doctrinal fit for typical deepfake injuries imperfect and procedurally onerous for victims.

2 The Constitutional Balancing Exercise: Freedom of Expression, Human Rights, and Democratic Integrity

Effective mitigation strategies must navigate the tensions inherent in liberal constitutional orders, requiring a careful balance between freedom of expression and the protection of other fundamental rights, particularly the right to receive accurate information and the integrity of democratic processes. Accurate information and knowledge are necessary for citizens to make informed political decisions, as systematically deceitful content can distort the opinion-forming process, potentially leading to electoral results based on a perverted public discourse.

The challenge lies in reconciling these competing constitutional demands, a process heavily influenced by contrasting legal traditions across the Atlantic. The French approach illustrates these dilemmas vividly: the 2018 "fake news law" (Loi n° 2018-1202) empowers judges to order the removal of false or manipulated content, including deepfakes, during election periods if it is likely to affect the outcome of a vote. While designed to safeguard democratic integrity, the law has been criticized for its potential chilling effects on freedom of expression and the press, as the broad and somewhat vague definitions of "false information" risk overreach (Douek,

2025). Similar tensions arise across the EU, where regulation must remain consistent with the European Convention on Human Rights and the Charter of Fundamental Rights of the EU, both of which enshrine freedom of expression while also permitting proportionate restrictions necessary in a democratic society under the rule of law premises. This balancing act demonstrates that regulating synthetic media is a constitutional challenge as much as a technical one, requiring legislators and courts to calibrate carefully between the prevention of harm and the preservation of open discourse.

Freedom of expression in Europe, codified in Article 10 of the European Convention on Human Rights (ECHR) and Article 11 of the EU Charter of Fundamental Rights, is recognized as a *relative* right, not an absolute one. The European framework incorporates a crucial *passive dimension* of freedom: the right to receive information in a pluralistic context, explicitly linking it to the functioning of a "democratic society". European courts prioritize values such as human dignity and pluralism. Consequently, false, misleading, or deceitful information does not receive the unfettered constitutional protection afforded under the US model. The ECHR framework explicitly allows for limitations to freedom of expression when such limitations are deemed "necessary in a democratic society" (Article 10(2)). The European Court of Human Rights (ECtHR) has confirmed that the Internet environment poses a "higher risk of harm" compared to traditional media, justifying greater limitations, provided that the legislator provides the framework for reconciling competing claims. This distinction makes the European Union's resulting multi-instrumental regulatory stack (DSA, AI Act, GDPR) constitutionally permissible, as its foundation is the defence of the passive right to be informed and the preservation of pluralism against intentional disinformation. In electoral periods, freedom of political debate is paramount, but in cases of conflict, contracting states have a margin of appreciation to restrict speech to protect the "free expression of the opinion of the people in the choice of the legislature".

3 Regulatory Measures as Mitigation Strategies: The EU Architecture

The EU has developed a complex, multi-instrumental architecture, designed to govern AI and content dissemination across the entire lifecycle (design, deployment, dissemination, and remedy). These instruments operate as complementary levers, and introduce points of friction and structural limitations.

3.1 General Data Protection Regulation (GDPR): Friction, Accuracy, and the Technical Impracticability of Erasure

The General Data Protection Regulation (GDPR) is immediately relevant because deepfakes are frequently produced using personal data, including images or other associated information that can be traced back to an individual, such as someone's recognisable voice. Article 4(2) GDPR defines "processing" broadly, covering every stage from collection to dissemination, which clearly encompasses the creation and distribution of deepfakes. A key obligation here is the principle of accuracy under Article 5(1)(d), which requires controllers to take reasonable steps to ensure that inaccuracies in personal data do not cause harm. Generative models that produce fabricated likenesses or statements implicate this principle when the output is traceably linked to an identifiable individual, particularly where reputational or dignitary harm follows.

Supervisory authorities have already begun to test the GDPR's applicability in this context. In 2022, the Italian Data Protection Authority (*Garante*) launched an investigation into *FakeYou*, a platform offering synthetic voice generation of public figures, to determine how personal data were being processed and whether safeguards against misuse were in place (Garante per la protezione dei dati personalo, 2022). More recently, in October 2023, the *Garante* adopted an urgent measure against *Clothoff*, an app that generated "deep nudes" by creating pornographic content from images of real people. The authority imposed the immediate limitation of data processing for Italian users, stressing that the service allowed anyone, including minors, to create synthetic sexualized content without verifying consent and without any indication of the artificial nature of the images. These cases show that EU data protection authorities view the misuse of deepfake technologies as a clear form of unlawful processing under the GDPR, particularly when fundamental rights such as dignity, privacy, and the protection of minors are at stake (Garante per la protezione dei dati personali, 2025*)* .

Despite this, enforcement faces significant technical friction. The right to erasure (Article 17) illustrates the problem: even if a data subject requests deletion, trained AI models may retain informational traces that allow re-synthesis of a likeness. This raises the need for controllers to ensure lawful data provenance and consent before training occurs, as post hoc deletion is technically challenging if not impossible. Further complexity arises from contextual exemptions, such as the household

exemption (Recital 18), which can shield the private creation of harmful deepfakes from GDPR scrutiny until dissemination occurs, creating a regulatory gap at the point of initial harm generation.

Ultimately, effective governance of deepfakes depends on aligning controller obligations under GDPR with the transparency and traceability requirements mandated by the forthcoming AI Act. Without rigorous enforcement of data provenance and consent under GDPR, subsequent interventions under the Digital Services Act (DSA) and AI Act risk becoming reactive, addressing harm only after it has occurred rather than preventing it at the source.

3.2 EU Artificial Intelligence Act (AI Act): The Limited-Risk Paradox and the Transparency Regime

The artificial intelligence Act (AI Act Regulation (EU) 2024/1689), the world's first comprehensive legal framework on AI, represents the EU's most explicit statutory engagement with synthetic media. The AI Act provides a legal definition of deepfakes: "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful" (Art. 3(60)).

The AI Act situates the problem of deepfakes within a political and ethical frame by foregrounding the risk of manipulation. Recitals 28 and 29 explicitly identify deception and manipulation among the principal social risks arising from the misuse of generative technologies, warning that such misuse can impair democratic processes and corrode public trust. Recital 133 further reiterates the legislative purpose of enabling individual recipients to recognise synthetic content and guard against impersonation and deceit.

The AI Act employs a risk-based approach, which includes a hard prohibition under Article 5 for AI systems categorized as posing an unacceptable risk. Specifically, Article 5 prohibits AI systems that use subliminal techniques or manipulative or deceptive techniques to distort behaviour, potentially causing physical or psychological harm. It also prohibits systems that exploit the vulnerabilities of individuals or specific groups. This provision sets a critical boundary against the most dangerous forms of manipulation.

For the vast majority of deepfakes, the AI Act addresses them through a mandatory transparency regime anchored in Article 50. This article imposes a dual obligation: providers of generative systems must ensure that outputs are marked in a machine-readable way, and deployers who disseminate synthetic content must disclose to the public that the material has been generated or manipulated. This infrastructure aims to make provenance and traceability foundational elements of the digital information ecosystem.

However, deepfakes are classified primarily as a *limited-risk* category, thereby avoiding the stringent substantive and supervisory requirements imposed on high-risk systems. This policy choice, intended to protect innovation and legitimate expressive uses, risks significant under-protection in contexts where manipulation yields acute public-interest harms, such as targeted electoral interference. The Act's reliance on transparency is vulnerable to adversarial evasion, as malicious actors can deliberately strip metadata or disseminate content via decentralized channels, thereby nullifying the prophylactic intent of Article 50. Moreover, the disclosure duty, linked to the standard of the "reasonably well-informed, observant and circumspect user", risks implicitly burdening less media-literate populations with verification duties, attenuating protection for those most susceptible to manipulation.

The AI Act's reliance on transparency is thus recognized as necessary but not sufficient to counter sophisticated manipulation, particularly in high-stakes political contexts where systemic democratic harm is the risk.

3.3 Digital Services Act (DSA): Reactive Moderation, Systemic Risk, and Enforcement Gaps

The Digital Services Act (DSA) is central to content governance, placing distinct obligations upon online intermediaries for content moderation, transparency, and, crucially, systemic risk assessments. For Very Large Online Platforms (VLOPs), the DSA mandates the identification and mitigation of systemic risks, including those arising from disinformation and algorithmic amplification.

Despite its importance, the DSA's efficacy is constrained by several limitations. First, its mechanisms are largely *reactive*, operating through notice-and-action procedures after content has already been posted. While effective in mitigating ongoing harm,

reactive measures cannot restore eroded public trust or undo immediate reputational injury. Second, the DSA focuses primarily on large, regulated platforms, neglecting important vectors of dissemination such as decentralized protocols and private messaging applications frequently used to circulate deepfakes. Third, enforcement relies on platform cooperation and transparency. Compliance monitoring, particularly concerning soft-law commitments like the Code of Practice on Disinformation, has been assessed as uneven and often lacking methodology-opaque reporting (Böswald, 2025). Therefore, while the DSA complements the AI Act by addressing dissemination, it does not negate the need for proactive provenance and detection at the generation point.

3.4 Product Liability Directive (PLD)

It is also important to note that the Product Liability Directive (PLD) has been revised in parallel with these regulatory processes, introducing measures designed to mitigate the information asymmetry between producers and users of AI systems. The revised Directive treats AI software as a "product" and introduces disclosure, burden-shifting, and transparency obligations (Articles 9–13), helping victims establish liability in cases of AI-related damage (Novelli et al., 2024). The scope of the Directive has been extended to include all AI systems and AI-enabled goods (excluding open-source software unless integrated into commercial products), reflecting the EU's recognition of AI's opacity and the imbalance of information between developers and consumers. This step represents an important breakthrough in adapting liability rules to the realities of generative AI and large language models.

However, the PLD reveals marked limitations when applied to deepfakes. While it reduces evidentiary burdens for victims and acknowledges AI models as legally relevant products, its remedial focus remains oriented toward physical injury and property damage. Non-material harms, such as reputational injury, dignity violations, or psychological distress, remain undercompensated. This means that although the GDPR offers direct pathways to challenge unlawful deepfake processing, the PLD provides only partial remedies and relies heavily on the AI Act to fill liability gaps. As scholars note, further legislative refinement will be necessary to extend liability to the full spectrum of harms typically caused by generative AI, especially in cases where reputational damage and privacy violations constitute the primary injury.

3.5 Soft Law Mechanisms

Legally binding regulation plays an important role in combating AI-generated disinformation. Nonetheless, policy research and policy negotiation efforts for the legislative process represent time-consuming activities (Schepel, 2005). Soft law tools in the form of non-binding norms, guidelines, codes of practice, and so on, can help manage lengthy regulatory processes by encouraging voluntary compliance from different stakeholders. Considering the fast-paced innovation in the generative AI context, earning it a place among disruptive technologies, soft law instruments promote flexible and timely reactions to promote ethical AI governance and to collect information on the empirical effects of soft law compliance (Păvăloaia & Necula, 2023).

In a context of international diplomatic and economic tension, however, it is argued that the non-binding nature of soft law raises concerns over its ability to attract stakeholders and encourage their compliance, contributing to concerns of a crisis of global AI governance (Leslie & Perini, 2024). The risk highlighted by the two authors is more real for some than others. The EU is particularly exposed to the flaws of soft law in the AI race: since 2001, the Commission has expressed interest in externalising governance duties by fostering the involvement of private stakeholders in contributing to relevant policy through self- and co-regulatory, i.e., non-binding measures.

Additionally, the legal challenges of AI appear particularly urgent considering the Union's role as a normative power: while the EU has traditionally leveraged on his large internal market to foster international companies" adaptation to European legal standards, including in the context of the fight to online disinformation, geopolitical attrition seems to undermine the principle of voluntary compliance that makes soft law a helpful tool in protecting online information and digital citizens' rights (Manners, 2002). By stressing soft law's complementary role *vis à vis* legally binding regulation, this section argues that integration of soft law tools in hard law covenants may foster AI regulation and, more specifically, the fight against AI-generated disinformation and deepfakes.

The recent endorsement of the European Commission and of the European Board for Digital Services of the 2022 Strengthened Code of Conduct in the Digital Services Act (DSA) points in this direction. The goal of contextualising the Code of

Conduct within EU regulation is allegedly to ensure better compliance with EU law for AI service providers and, consequently, to clearly define accountability.

The persistence of an accountability gap is motivated by several factors, some of them already been referenced earlier. In the first place, there persists a struggle to regulate AI, which is in turn related to both economic competitiveness concerns and to the technology-induced legislative lag (European Commission, 2025; Kosta et al., 2025). On the other hand, issues related to the opacity of AI algorithms and to our ability to attribute agency, and therefore, accountability, to AI algorithms hinders the legislator's ability to "show that the issues have been conscientiously addressed and how the result has been reached; or alternatively alert the recipient to a justiciable flaw in the process"(Calderonio 2025; Floridi 2023; Williams et al. 2022).

A great deal of uncertainty in relation to accountability, moreover, stems from the semantic uncertainty surrounding the concept, given its fluid, i.e., context and discipline-dependent, meaning. Williams et al. suggest that abstract aspirations, such as the principle of accountability, need to be specific and enforceable, an applicability gap also highlighted by Leslie and Perini. They argue that, by mapping the semantic debate on accountability, it is possible to identify five concepts, related chronologically in terms of how these terms are related, which inform the others, and how, as well as from an "activity" perspective. By the latter, it is meant how these terms foster push-pull dynamics or, in other words, to clarify whether AI providers are required to make information available (push) or if it is end-users who seek information in each context (pull). According to the authors, accountability is the last step necessary to make the concepts listed above enforceable. At the same time, these principles allow for framing accountability differently depending on the (AI) system under inquiry, making these aspirations capable of being enforced and of managing different AI systems.

It becomes then clear that frameworks like the one proposed by Williams et al. represent a necessary step to move from principles to practice, even if further challenges posed by generative AI to delineating AI agency and accountability will require a fine-tuning of such models. Nonetheless, integrating soft law instruments against disinformation in legally binding documents represents an attempt to bolster the commitment to the fight against disinformation, as well as a necessary step to deliver the tools and the metrics to tackle the AI services providers' accountability gap.

Now, soft law tools such as the Strengthened Code of Practice envisage objectives for signatories such as the following: the release of periodic transparency reports covering volumes of synthetic content, the number, and outcomes of reports and takedowns; the use of standardized reporting templates to enable comparative evaluation; cooperation during sensitive events, e.g., electoral periods. However, cooperation from this perspective has at times been sluggish, with AI companies providing limited and incomplete information or sloppy justification for the data collection methodology that they presented (OECD, 2024).

By contextualising such soft law tools into legally binding regulation (such as the DSA), nonetheless, it would be possible to frame accountability issues within a specific policy setting. If, on the one hand, this would eventually prompt EU institutions to defend their reliance on codes of conduct *et similia* in the AI governance context, on the other hand, it also articulates those push factors that AI service providers need to be presented with, as advocated for by Williams et al.

In short, soft law tools represent an important means to foster the objectives of documents that articulate compulsory actions, such as the DSA. By being contextualized within binding documents, it becomes possible to move from ethical AI governance principles advocated for in soft law tools to their practice.

4 Structural Challenges in the Governance of Deepfakes

Despite the EU's increasingly dense regulatory ecosystem, deepfakes expose persistent structural vulnerabilities in law's capacity to safeguard democratic integrity and individual dignity. The problem is not merely the presence of malicious actors but the systemic asymmetries between rapid technological development and the slower pace of legislative adaptation, the uneven enforcement capacities across Member States, and the incomplete coverage of harms, particularly immaterial and distributive ones. This section identifies five interlinked shortcomings in the current governance framework.

1) **Technological-Legislative Asymmetry:** the foundational challenge is the inherent disparity in speed between technological innovation and regulatory response. Generative capabilities evolve rapidly, meaning detection techniques (such as inference-based methods) and provenance architectures (such as watermarking) are often one step behind.

The AI Act's reliance on transparency is vulnerable to adversarial evasion strategies. Malicious actors can deliberately strip metadata, transcode files, re-edit labelled outputs, or employ adversarial attacks to obfuscate generation signatures, effectively nullifying the prophylactic intent of Article 50. The AI Act requires to track the provenance of AI-generated media. However, it does so without requiring sustained public investments towards detection research risks. At the same time, it delegates enforcement to private stakeholders, who may lack the necessary resources or incentives. A resilient architecture requires proactive measures, including public funding for detection research and the standardization of robust, tamper-resistant provenance mechanisms that prioritize interoperability.

2) **The Honest-Actor Problem and Transnational Enforcement Deficits:** EU legal instruments principally regulate actors with a clear EU nexus, providers, deployers, and platforms operating in the Union. However, deepfakes are easily disseminated across borders, and malicious actors often operate from jurisdictions with weak enforcement capacity or through highly decentralized protocols.

The ease of cross-border dissemination enables sophisticated evasion strategies. This governance gap means that domestic legal obligations risk producing mere "protective islands" that are porous at their boundaries. Addressing this honest-actor problem requires robust international cooperation, harmonized standards for provenance and liability, and the establishment of reliable bilateral and multilateral channels for rapid content takedown and mutual legal assistance.

3) **Fragmentation and Enforcement Deficit:** The multi-instrumental nature of EU regulation, involving the AI Act, DSA, GDPR, and PLD, creates both overlap and complexity. While redundancy can increase robustness, complexity undermines clarity for regulated entities. Divergent interpretations of obligations by various enforcement bodies, national data protection authorities, digital services coordinators, and national courts exacerbate this issue.

Furthermore, uneven enforcement capacity across Member States results in varied levels of protection. This fragmentation risks creating a 'forum-shopping' environment, whereby bad actors may seek legal solutions in more

permissive countries and, by doing so, hindering seeking remedies face and increasing procedural hurdles and uncertainty for victims. Uniformity in enforcement is necessary to ensure the protective effect of the EU stack is realized consistently.

4) **Insufficient Coverage of Immaterial Harms:** A core normative lacuna remains the treatment of immaterial harms. Deepfakes frequently inflict severe non-material injuries, including the erosion of dignity, emotional distress, and reputational degradation. Existing liability frameworks, such as the PLD, are slowly adapting to AI but remain historically oriented toward material or pecuniary damages.

   To bridge this remedial gap, substantive legal reform must be complemented by procedural innovation. Because the velocity of harm propagation is high in the digital sphere, temporal responsiveness is critical. Regulatory design should incorporate expedited administrative remediation pathways, such as mandates for swift provisional injunctive relief or statutory entitlements to rapid removal of non-consensual or clearly falsified content, in addition to traditional civil damages.

5) **Uneven Impact and Distributive Vulnerability:** Deepfakes do not affect all populations equally. Empirical evidence indicates a clear differential impact, with victims of non-consensual intimate imagery overwhelmingly being women, and marginalized groups frequently targeted by political disinformation campaigns (Kira, 2024). A regulatory architecture prioritizing technological neutrality may unintentionally fail to centre distributive justice and dignity.

   Addressing differential impact requires a rights-sensitive lens in regulatory design. Legal and policy responses must prioritize protective measures for the most vulnerable groups, ensuring mechanisms such as expedited takedown are readily accessible, alongside legal aid and psychosocial support. Furthermore, platform operators and regulators must incorporate explicit distributive impact assessments as part of their systemic risk frameworks (under the DSA), ensuring that mitigation efforts do not merely shift harms to less visible spaces or less empowered communities.

5 Concluding Remarks

In this chapter, we first framed the concept of disinformation through a socio-ethical lens, looking at how relational responsibility and the material consequences of immaterial harms emerge from the "hyperconnected infosphere". Second, we highlighted different traditions in the freedom of expression and its legal understanding, in order to justify effective mitigation strategies and their reach. Third, we presented the EU's complex AI governance architecture and analysed the legal instruments on which it leverages. Fourth, we identified five shortcomings hindering EU AI governance.

The analysis of mitigation strategies demonstrates that the EU's governance architecture for synthetic media is characterized by groundbreaking, innovative intent and significant structural insufficiencies. The classification of deepfakes as limited-risk under the AI Act, combined with the inherently reactive nature of platform governance under the DSA, limits the overall efficacy of the regulatory stack.

The evidence from constitutional traditions and case studies confirms that disinformation, particularly when amplified by synthetic media, constitutes a systemic threat to democratic participation. Effective governance must be rooted firmly in the European constitutional commitment to *pluralism* and the passive right to *informed choice*, justifying intervention that transcends the US marketplace paradigm.

However, persistent challenges, the technological-legislative asymmetry, the honest-actor problem, and the insufficient legal coverage of non-material harms demand strategic recalibration. An effective, future-proof governance strategy requires a coordinated policy shift that moves resolutely beyond a transparency-only paradigm for high-consequence contexts. The priorities must include substantial public investment in interoperable detection and tamper-resistant provenance standards (as technical solutions complement legal frameworks), securing international regulatory harmonization, and creating procedural mechanisms tailored to expedite remedial action for reputational and psychological harms. These diagnostic conclusions form the essential analytical groundwork for the integrated policy recommendations that will be detailed in Chapter 8.

End notes

Yasaman Yousefi is the lead author of this Chapter. She conceptualized this Chapter, coordinated the writing and polished the final text. She wrote the introduction and the conclusion, as well as the following sections: The Constitutional Balancing Exercise: Freedom of Expression, Human Rights, and Democratic Integrity, Regulatory Measures as Mitigation Strategies: The EU Architecture, and Structural Challenges in the Governance of Deepfakes. Maria Dolores Sanchez Galera contributed to the legal analysis in Regulatory Measures as Mitigation Strategies: The EU Architecture. Angelo Tumminelli and Calogero Caltagirone wrote the conceptual considerations. Tomasso Tonello wrote the Soft Law Mechanisms. All authors reviewed and approved the final version.

References

Böswald, L.-M. (2025, November 19). Soft law, hard risks? Co-regulation and risk mitigation under the Digital Services Act. Interface. https://www.interface-eu.org/publications/dsa-co-regulatory-mechanisms

Calderonio, V. (2025). *The opaque law of artificial intelligence* (arXiv:2310.13192). *arXiv*. https://doi.org/10.48550/arXiv.2310.13192

Chesney, R., & Citron, D. K. (2018). *Deep fakes: A looming challenge for privacy, democracy, and national security*. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3213954

Da Re, A. (2003). *Filosofia morale: Storia, teorie, argomenti*. Pearson Italia.

Deep fake: Garante avvia istruttoria su app che falsifica le voci. (2022, October 12). *Garante per la protezione dei dati personali*. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9816291

Deepfake, Garante privacy: Stop a Clothoff, l"app che spoglia le persone. (2025, October 3). *Garante per la protezione dei dati personali*. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10174320

Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC (Text with EEA relevance). (2024). *Official Journal of the European Union*. http://data.europa.eu/eli/dir/2024/2853/oj

Donati, P. (2024). *Being human in a virtual society*. Peter Lang. https://www.peterlang.com/document/1461639

Douek, E. (2024). The Politics and Perverse Effects of the Fight Against Online Medical Misinformation. *Yale LJF*, *134*, 237.

European Commission. (2025, February 13). *Commission endorses the integration of the voluntary Code of Practice on Disinformation into the Digital Services Act*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/commission-endorses-integration-voluntary-code-practice-disinformation-digital-services-act

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), Article 15. https://doi.org/10.1007/s13347-023-00621-y

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Kira, B. (2024). *Deepfakes, the weaponisation of AI against women and possible solutions*. *Verfassungsblog*. https://doi.org/10.59704/9987d92e2c183c7f

Kosta, E., Hallinan, D., Hert, P. D., & Nusselder, S. (2025). *Data protection, privacy and artificial intelligence* (Vol. 17). Bloomsbury Publishing.

Leslie, D., & Perini, A. M. (2024). Future Shock: Generative AI and the international AI policy and governance crisis. Harvard Data Science Review, (Special Issue 5).

Manners, I. (2001). Normative Power Europe. A contradiction in terms, 235-258.

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity. Computer Law & Security Review, 55, 106066.

OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/facts-not-fakes-tackling-disinformation-strengthening-information-integrity_ff96d19f/d909ff7a-en.pdf

Reuters. (2020, August 3). *Fact check: "Drunk" Nancy Pelosi video is manipulated*. https://www.reuters.com/article/world/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2B1/

Schepel, H. (2005). The constitution of private governance: Product standards in the regulation of integrating markets (Vol. 4). Hart Publishing.

Păvăloaia, V. D., & Necula, S. C. (2023). Artificial intelligence as a disruptive technology – a systematic literature review. Electronics, 12(5), 1102.

Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., ... & Pang, W. (2022). From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, 4, e7