ARTICLE

ECO-DISINFORMATION AND AI DEEPFAKES: DEVELOPING A LEGAL FRAMEWORK AGAINST DIGITAL SABOTAGE OF ENVIRONMENTAL POLICIES

Most. Rezwana Nishat Leana [*] & Md Abu Rayhan [**]

Abstract

Artificial intelligence has transformed the production and circulation of synthetic media, creating new possibilities for environmental disinformation. Deepfakes can fabricate apparently authentic statements, events and evidence concerning climate science, environmental regulators, public officials and renewable-energy projects. Although existing legal regimes address aspects of defamation, fraud, environmental information, platform governance and artificial-intelligence transparency, they do not expressly connect malicious synthetic deception with interference in environmental decision-making. This article develops the concept of "Digital Environmental Sabotage" and proposes a narrowly tailored legal framework for addressing it. Using doctrinal and comparative legal research, the article examines environmental-information principles under the Aarhus Convention and Paris Agreement, intermediary and systemic-risk regulation under United States and European Union law, emerging transparency obligations under the EU Artificial Intelligence Act, and human-rights-based principles for digital-platform governance. It argues that liability should depend upon cumulative elements: material deception, knowledge or intent, an identifiable environmental nexus, and material harm or a serious and foreseeable risk of interference with environmental governance. The article then proposes a model Prevention of Digital Environmental Sabotage Act (PDESA) covering definitions, jurisdiction, civil and criminal liability, platform responsibilities, provenance and labelling, notice-and-action procedures, institutional oversight and safeguards for freedom of expression. The proposed framework is designed to protect the integrity of environmental decision-making without converting legitimate scientific disagreement, political criticism, satire, journalism or academic inquiry into unlawful conduct.

Keywords:- Eco-disinformation; artificial intelligence; deepfakes; environmental law; Digital Environmental Sabotage; platform liability; synthetic media; climate governance.

I. INTRODUCTION

Environmental governance increasingly depends upon the integrity of the digital information environment. Governments announce climate policies online; environmental organisations mobilise public participation through social media; scientific institutions disseminate research digitally; and citizens increasingly encounter information concerning climate change, pollution, biodiversity and renewable energy through algorithmically curated services. Environmental disputes are therefore no longer confined to legislatures, administrative agencies and conventional media. They are also

[*] Apprentice Lawyer, Dhaka Judge Court, Bangladesh.
[**] Student, Dhaka International University

contested within digital infrastructures capable of producing and distributing persuasive synthetic content at unprecedented speed.

This Paper identifies this problem through the concept of "eco-disinformation," describing the possibility that AI-generated material can fabricate environmental statements, events and evidence in ways capable of misleading the public and influencing policy.

The development of generative AI changes the character of that risk. A false environmental claim can be written, illustrated, voiced or rendered as a realistic video with comparatively little technical expertise. Deepfake technology can make a scientist appear to retract a finding, a regulator appear to admit corruption, or a renewable-energy facility appear to have suffered an accident that never occurred. The resulting deception can affect public confidence, investment, administrative hearings and political pressure.

Recent climate-misinformation research shows that misinformation is evolving beyond straightforward denial toward more sophisticated forms of delay, deflection and uncertainty, while generative AI is expected to increase the speed and sophistication of information manipulation.[1]

The legal question is not whether every false environmental statement should be prohibited. Environmental governance necessarily contains scientific uncertainty, political disagreement and competing policy preferences. The relevant problem is narrower: what should the law do when synthetic media is deliberately created or disseminated to make a false representation appear authentic and thereby interfere with an identifiable environmental policy, regulatory decision or environmentally significant project?

This article proposes the concept of Digital Environmental Sabotage to describe that conduct. The term is normative and proposed; it does not claim that an independent offence with this title already exists in international or domestic law. The objective is to identify a legal gap and develop a model framework capable of addressing it while preserving freedom of expression.

The article makes three principal contributions. First, it connects the information-integrity problem identified in climate-misinformation scholarship with the procedural values of environmental governance. Second, it develops a cumulative legal test distinguishing malicious synthetic deception from legitimate environmental disagreement. Third, it proposes the Prevention of Digital Environmental Sabotage Act (PDESA) as a model statute combining liability, platform governance, provenance and procedural safeguards.

II. RESEARCH OBJECTIVES AND METHODOLOGY

The research is doctrinal and comparative. It examines primary legal instruments and authoritative regulatory materials concerning access to environmental information, public participation, online intermediary governance and AI-generated content. It also draws upon recent scholarly literature addressing climate misinformation, deepfake regulation and freedom of expression.

The research addresses four questions: (1) how can synthetic environmental deception produce legally cognisable harm; (2) why do existing legal doctrines leave a regulatory gap; (3) what elements should constitute Digital Environmental Sabotage; and (4) how can such a regime be designed without suppressing legitimate expression?

[1] James Kennon Rice, Beyond Post-Truth: Projecting the Future Trajectory of Climate Misinformation, 5 PLOS Climate e0000916 (2026); see also Yotam Ophir, Misinformation and Society (Wiley-Blackwell 2025).

The analysis is deliberately jurisdictionally comparative rather than tied to a single domestic legal system. The European Union provides particularly relevant examples through the Digital Services Act and AI Act; United States law provides a contrasting intermediary-liability model through section 230 of the Communications Decency Act; and international environmental and human-rights instruments provide normative principles concerning information, participation and freedom of expression.

III. ECO-DISINFORMATION AND THE DEEPFAKE PROBLEM

A. From Misinformation to Deliberate Synthetic Deception

A useful legal distinction exists between misinformation and disinformation. Misinformation may be false or misleading without proof of an intention to deceive. Disinformation, by contrast, ordinarily denotes deliberately deceptive or manipulated information. The distinction is particularly important for criminal or civil liability because a legal rule based solely on falsity would risk punishing honest error, uncertainty or evolving scientific knowledge.

Recent research describes climate misinformation as encompassing denial, cherry-picking, false causation, misleading comparisons and other forms of distorted framing.[2]

Generative AI introduces an additional layer: synthetic authenticity. The problem is no longer limited to whether a proposition is true. A fabricated audiovisual representation can create the appearance that a particular person made a particular statement or that a particular event occurred. The evidentiary force of the medium itself becomes part of the deception.

Empirical research on climate-related deepfakes has found substantial difficulty among viewers in identifying manipulated material, with vulnerability varying according to the content and the characteristics of the viewer.[3]

B. Environmental Applications

The environmental context presents several plausible categories of malicious synthetic media. Identity-based deepfakes may impersonate scientists, environmental officials, activists or corporate executives. Event-based deepfakes may fabricate pollution incidents, ecological disasters or failures of renewable infrastructure. Policy-based deepfakes may falsely depict ministers, regulators or public agencies announcing decisions that have not occurred. Financially oriented synthetic claims may falsely represent that an environmental project has failed, received an unlawful subsidy or caused an economic catastrophe.

The harm can be institutional as well as reputational. A fabricated video distributed immediately before an environmental hearing may affect public participation. A false image circulated during an investment decision may influence financing. A fabricated statement by a regulator may generate unnecessary administrative responses. The legal significance therefore lies in the relationship between the digital act and the environmental decision-making process.

---

[2] Zeinab Shahbazi & Sara Behnamian, Using Large Language Models to Detect and Debunk Climate Change Misinformation, 10 Big Data & Cognitive Computing, no. 1, 34 (2026), https://doi.org/10.3390/bdcc10010034.

[3] Anthony C. C. Lee et al., Deepfakes and Scientific Knowledge Dissemination, 13 Sci. Rep. 12959 (2023), https://doi.org/10.1038/s41598-023-39944-3.

IV. ENVIRONMENTAL GOVERNANCE AND INFORMATION INTEGRITY

A. The Aarhus Convention

The Aarhus Convention is structured around three pillars: access to environmental information, public participation in environmental decision-making, and access to justice in environmental matters.[4]

The Convention does not create a specific offence of environmental disinformation. It nevertheless establishes an important normative connection between environmental governance and reliable access to information. If public participation is formally available but participants are systematically exposed to fabricated evidence, the quality of participation may be impaired even though the procedural right exists on paper.

This suggests that digital information integrity should be treated as an enabling condition of environmental participation rather than as a substitute for substantive environmental regulation. A proposed anti-sabotage framework should therefore supplement, not replace, existing environmental-information rights.

B. The Paris Agreement

Article 12 of the Paris Agreement provides for cooperation, as appropriate, on climate-change education, training, public awareness, public participation and public access to information.[5]

Again, Article 12 does not establish a deepfake offence. Its relevance is structural. Effective climate governance requires an informed public capable of participating in decisions and evaluating competing claims. Digital manipulation that deliberately creates false evidence may therefore undermine an important precondition of participatory climate governance.

V. LIMITATIONS OF EXISTING LEGAL FRAMEWORKS

A. Defamation and Tort

Defamation is potentially applicable when a deepfake falsely attributes misconduct to a scientist, official, activist or company. It becomes less satisfactory when the primary injury is not reputational but institutional. For example, when the objective is to delay a public environmental project or distort an administrative decision.

This paper proposes a "civil tort of policy defamation." The idea is innovative, but the terminology risks confusing reputational injury with interference in public decision-making. Environmental policy does not possess reputation in the same doctrinal sense as a natural person. A statutory environmental digital-harm action would therefore be conceptually cleaner.

Such an action should require proof of the protected interest, the defendant's mental state, the environmental nexus, causation and actual or legally cognisable risk of harm. Without those limitations, public bodies could use environmental litigation to suppress criticism of policy.

[4] Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters arts. 1, 4–9, June 25, 1998, 2161 U.N.T.S. 447.
[5] Paris Agreement art. 12, Dec. 12, 2015, T.I.A.S. No. 16-1104, 3156 U.N.T.S. 79.

B. Intermediary Liability

Section 230 of the Communications Decency Act generally provides that an interactive computer service shall not be treated as the publisher or speaker of information supplied by another information-content provider, while also containing protection for certain good-faith content-moderation activity.[6]

The U.S. approach illustrates why platform liability cannot simply be transferred from users to intermediaries. A platform may host enormous quantities of user content and may lack the factual context necessary to determine whether an environmental claim is false. Overbroad liability can encourage removal of lawful expression.

C. The European Union's Risk-Based Approach

The Digital Services Act adopts a different architecture for very large online platforms and search engines. Article 34 requires assessment of systemic risks arising from the design and functioning of services, while Article 35 requires reasonable, proportionate and effective mitigation measures tailored to identified systemic risks.[7]

This risk-based approach is more suitable than strict liability for individual pieces of content. It recognises that harm can arise from platform design, recommender systems and amplification patterns rather than solely from the original speaker.

D. AI Transparency and Deepfakes

The EU Artificial Intelligence Act provides a particularly direct example. Article 50(4) requires deployers of AI systems generating or manipulating image, audio or video content constituting a deepfake to disclose that the content has been artificially generated or manipulated, subject to specified exceptions, including an adapted rule for artistic, creative, satirical and fictional works.[8]

This transparency model supports an important distinction for PDESA: synthetic content should not be unlawful merely because it is synthetic. The regulatory trigger should be the combination of deception, intent and environmental interference.

VI. THE PROPOSED DOCTRINE OF DIGITAL ENVIRONMENTAL SABOTAGE

This article proposes the following definition:

> *"Digital Environmental Sabotage" means the intentional creation, manipulation, coordination, or dissemination of materially deceptive synthetic media undertaken for the principal purpose of delaying, obstructing, altering, or unlawfully influencing an identifiable environmental policy, regulatory decision, or environmentally significant project.*

The definition contains four cumulative elements.

A. Materially Deceptive Synthetic Media

The content must be materially false or manipulated in a manner capable of creating a false impression concerning a person, statement, event or circumstance. The definition should be

[6] 47 U.S.C. § 230(c)(1), (2).

[7] Regulation (EU) 2022/2065 of the European Parliament and of the Council of Oct. 19, 2022, on a Single Market for Digital Services (Digital Services Act), arts. 34–35, 2022 O.J. (L 277) 1.

[8] Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13, 2024, Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 50(4), 2024 O.J. (L 1689) 1.

technologically neutral. It should therefore cover future forms of synthetic media without tying liability to a particular AI architecture.

B. Knowledge or Intent

Criminal liability should ordinarily require knowledge of material falsity or an intention to cause an audience to believe that the synthetic representation is authentic. This element prevents liability based on innocent sharing, reasonable mistake or unresolved scientific disagreement.

C. Environmental Nexus

The conduct must be connected to an identifiable environmental policy, regulatory proceeding, institution or environmentally significant project. A generic political deepfake with no environmental nexus should not become an environmental offence simply because the subject matter happens to mention climate change.

D. Material Harm or Serious Foreseeable Risk

Finally, there must be either material harm or a serious and foreseeable risk of interfering with the identified environmental decision. Mere speculative possibility should be insufficient. This requirement protects against expansive liability while recognising that some environmental decisions are time-sensitive and that harm may occur before a full physical consequence becomes measurable.

The four elements create a meaningful boundary between protected expression and unlawful manipulation. Saying that a renewable-energy policy is economically misguided is political opinion. Fabricating an apparently authentic recording of a minister admitting that the project has secretly failed, while knowing the recording is false and intending to delay approval, falls within the proposed category.

VII. MODEL PREVENTION OF DIGITAL ENVIRONMENTAL SABOTAGE ACT

A. Purpose and Jurisdiction

The Prevention of Digital Environmental Sabotage Act (PDESA) should state that its purposes are to protect the integrity of environmental decision-making, prevent malicious synthetic-media campaigns, promote trustworthy environmental information, establish proportionate remedies and preserve freedom of expression.

Jurisdiction should exist where the conduct occurs within the territory, where a substantial portion of the targeted audience is located within the jurisdiction, or where substantial environmental or regulatory harm occurs there. Cross-border application should operate through ordinary mutual legal assistance and judicial cooperation rather than unlimited extraterritorial enforcement.

B. Offence and Civil Liability

A criminal offence should arise only where the prosecution proves beyond the applicable standard that the defendant knowingly created or disseminated materially deceptive synthetic media, intended it to be understood as authentic, and intended materially to interfere with an identifiable environmental policy, decision or project.

Civil remedies should include injunctions, correction or retraction orders, compensation for demonstrable loss, preservation of evidence and disclosure orders relating to coordinated campaigns. Public authorities and qualifying environmental organisations may be given standing where they can

establish a statutory or legally protected interest. Mere disagreement with government policy should never be sufficient.

C. Platform Responsibilities

PDESA should impose heightened duties on large platforms whose systems can materially amplify environmental disinformation. These duties should include risk assessment, reasonable detection and labelling systems, transparency regarding recommendation systems, preservation of evidence following lawful notices, expedited review of verified high-risk campaigns and meaningful user appeals.

UNESCO's Guidelines for the Governance of Digital Platforms advocate human-rights due diligence, transparency, accountability and a multistakeholder approach to online information governance, while cautioning against arbitrary or disproportionate state measures.[9]

The statute should not impose an absolute duty to detect every deepfake. Detection technology is imperfect, context is often ambiguous and excessive automation can create false positives. A risk-based duty to implement reasonable systems is more consistent with proportionality.

D. Provenance, Labelling and Content Credentials

The original manuscript proposes cryptographic and immutable watermarking.[7] The objective is valuable, but a watermark alone cannot establish truth. Metadata may be stripped and content may be transformed after generation.

The Coalition for Content Provenance and Authenticity (C2PA) provides a relevant technical model through Content Credentials and cryptographically verifiable provenance. The system is designed to preserve information concerning how digital assets were created or transformed and to make provenance tamper-evident within a defined trust model.[10]

PDESA should therefore encourage layered provenance: cryptographic content credentials, disclosure labels, platform-level signals and human review. The statute should avoid presenting provenance as conclusive evidence that a claim is true; provenance establishes information about origin and transformation, not the substantive truth of every assertion contained in an asset.

E. Notice, Review and Removal

A competent authority should be permitted to issue a verified notice where synthetic media creates an immediate and substantial risk to an identifiable environmental decision. Platforms should conduct prompt review and may label the material, reduce algorithmic amplification, preserve evidence or remove the material where the statutory threshold is met.

A fixed 24-hour removal requirement should not apply automatically to every government notice. Emergency procedures may be justified in exceptional circumstances, but removal should be reasoned, reviewable and accompanied by an appeal mechanism. Government requests should be logged and subject to transparency reporting.

F. Independent Digital Environmental Security Council

The proposed Digital Eco-Security Council in the original manuscript can be retained in revised form as an independent multidisciplinary authority. Its membership could include legal scholars,

[9] UNESCO, Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information Through a Multistakeholder Approach 12–25 (2023).

[10] Coalition for Content Provenance and Authenticity, C2PA Specifications 2.4, Content Credentials, https://spec.c2pa.org/specifications/.

climate scientists, digital-forensics experts, technologists, civil-society representatives and environmental-policy specialists.

Its functions should include developing verification standards, receiving complaints, coordinating investigations, issuing non-binding technical guidance, publishing annual transparency reports, supporting courts and regulators, and facilitating cross-border cooperation. It should not have authority to determine which environmental opinions are politically acceptable.

VIII. FREEDOM OF EXPRESSION AND HUMAN-RIGHTS SAFEGUARDS

Any law targeting synthetic media must confront the risk of censorship. Deepfakes may be harmful, but not all false or fictional content warrants prohibition. Satire, parody, political criticism, journalism, academic research, artistic expression and whistleblowing can use manipulated or fictional material for legitimate purposes.

The literature on deepfake regulation emphasises that the regulatory problem requires balancing expressive interests against serious harms rather than assuming that falsity alone removes constitutional or human-rights protection.[11]

UNESCO likewise recommends that restrictions on expression satisfy legality, legitimate aim, necessity and proportionality, and calls for transparency concerning state requests to platforms.[12]

PDESA should therefore include explicit exemptions for political opinion, scientific disagreement, satire and parody, good-faith journalism, academic research, artistic and fictional works, and public-interest whistleblowing. The exemptions should not protect a person who merely labels a deceptive commercial or political deepfake "satire" after the fact; context, presentation and intent must remain relevant.

Procedural safeguards are equally important. Government notices should identify the allegedly false content, explain the environmental nexus, state the legal basis for intervention and be subject to independent review. Platforms should provide reasons for significant restrictions and maintain an accessible appeal process. These safeguards reduce the risk that environmental information regulation becomes a tool for suppressing dissent.

IX. INTERNATIONAL COOPERATION

Digital Environmental Sabotage is inherently transnational. A campaign may be generated in one state, hosted through infrastructure in another, amplified by a platform headquartered elsewhere and directed toward a regulatory proceeding in a fourth jurisdiction. Domestic law therefore requires mechanisms for preservation and exchange of digital evidence, mutual legal assistance and regulatory cooperation.

International cooperation should focus on minimum standards rather than a single global offence. States could coordinate around common definitions of synthetic media, interoperable provenance standards, transparent government requests, high thresholds for criminal liability and rights-protective procedures.

---

[11] See Bart van der Sloot & Yvette Wagensveld, Deepfakes: Regulatory Challenges for the Synthetic Society, 46 Computer L. & Sec. Rev. 105716 (2022), https://doi.org/10.1016/j.clsr.2022.105716; Alex Barber, Freedom of Expression Meets Deepfakes, 202 Synthese 40 (2023), https://doi.org/10.1007/s11229-023-04266-4.

[12] UNESCO, supra note 11, at 22–25; see also International Covenant on Civil and Political Rights art. 19, Dec. 16, 1966, 999 U.N.T.S. 171.

Environmental institutions should also cooperate with digital regulators. Environmental agencies possess substantive expertise concerning climate policy and ecological risk, while digital regulators possess expertise concerning platform systems and content governance. The proposed multidisciplinary model is intended to bridge that institutional divide.

X. CONCLUSION

The rise of generative AI creates a new information environment in which fabricated statements, images, audio and video can appear authentic and circulate rapidly. When such material is deliberately deployed against environmental policies or institutions, the resulting harm can extend beyond reputational injury to the integrity of public participation, administrative decision-making, investment and environmental governance.

Existing legal frameworks are relevant but fragmented. Defamation addresses certain personal harms; criminal law can address fraud and impersonation; environmental law protects substantive and procedural interests; platform regulation addresses systemic risks; and AI regulation increasingly introduces transparency obligations for synthetic content. None, however, necessarily supplies a coherent legal category for deliberate synthetic deception aimed specifically at environmental decision-making.

Digital Environmental Sabotage is proposed as that connecting concept. Its legitimacy depends upon narrow elements: material deception, knowledge or intent, an identifiable environmental nexus, and material harm or a serious and foreseeable risk of interference. Those requirements distinguish malicious manipulation from environmental disagreement.

The proposed PDESA should consequently be understood as an information-integrity framework rather than a law against environmental criticism. Its strongest features are a risk-based approach to platform responsibility, layered provenance rather than reliance on a single watermark, independent oversight, judicially reviewable procedures and explicit protection for lawful expression.

The central legal challenge is therefore not to eliminate synthetic media. Synthetic media has legitimate artistic, educational, journalistic and technological uses. The challenge is to prevent its deliberate use as fabricated evidence within environmental governance while preserving the democratic conditions under which environmental policy can be contested. A proportionate, rights-based and technologically adaptive framework can protect that distinction.

XI. RECOMMENDATIONS

The proposed framework should adopt a technology-neutral statutory definition of Digital Environmental Sabotage, based on material deception, intent, an identifiable environmental nexus, and material harm or a serious foreseeable risk of harm. It should employ a risk-based model of platform liability, drawing on systemic-risk principles rather than imposing strict liability for user-generated content. The framework should also require the disclosure or labelling of qualifying deepfake content, while preserving appropriate exemptions for artistic, satirical, fictional, journalistic, and academic uses. Furthermore, it should promote interoperable provenance systems, including cryptographically verifiable Content Credentials, without treating provenance information as conclusive proof of the substantive truth or falsity of the underlying content. An independent, multidisciplinary authority should be established with transparent procedures and effective judicial review, supported by expedited procedures for serious and verified threats to imminent environmental decisions, while ensuring

reasons, appeal mechanisms, and transparency reporting. At the same time, the framework should expressly protect political criticism, scientific disagreement, satire, parody, journalism, academic research, artistic expression, and good-faith whistleblowing from overbroad enforcement. Effective implementation should also strengthen cross-border cooperation among environmental regulators, digital regulators, online platforms, scientific institutions, and law-enforcement bodies. Finally, legal enforcement should be complemented by sustained investment in climate literacy, digital literacy, and public awareness, thereby strengthening societal resilience against synthetic environmental deception.