Universidad Andina Simón Bolívar

Sede Ecuador

Área de Derecho

Maestría en Derecho de la Economía Digital

Legal impact of Artificial Intelligence (AI) hallucinations

Darío Efraín Echeverría Muñoz

Tutor: Luís Fernando Enríquez Álvarez

Quito, 2025

Universidad Andina Simón Bolívar
Ecuador Headquarters

Area of Law

Master's Degree in Digital Economy Law

Legal impact of Artificial Intelligence (AI) hallucinations

Darío Efraín Echeverría Muñoz
Tutor: Luís Fernando Enríquez Álvarez

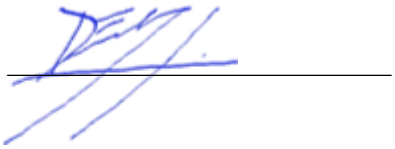
Quito, 2025

Publication rights assignment clause

I, Darío Efraín Echeverría Muñoz, author of the thesis entitled “Legal impact of Artificial Intelligence (AI) hallucinations”, hereby state that the work is of my exclusive authorship and production, and that I have prepared it to fulfill one of the prerequisites for obtaining the degree of Master in Digital Economy Law at the Universidad Andina Simón Bolívar, Ecuador headquarters.

1. I hereby grant to the Universidad Andina Simón Bolívar, Ecuador headquarters, the exclusive rights of reproduction, public communication, distribution and dissemination, for 36 months after my graduation, and therefore the University may use and use this work by any means known or to be known, as long as it is not done for financial gain. This authorization includes the total or partial reproduction in virtual, electronic, digital, optical formats, as well as local network and internet uses.
2. I declare that in the event of any claim by third parties with respect to the copyright of the aforementioned work, I will assume all liability to third parties and to the University.
3. On this date, I hereby deliver to the General Secretariat the respective copy and its annexes in printed and digital or electronic format.

May 12, 2025.

Signature: 

Summary

This thesis examines the legal impact of hallucinations produced by artificial intelligence (AI) systems, defining them as inaccurate responses generated by AI models. It details how this phenomenon arises from reliance on training data and erroneous extrapolation, which poses significant challenges in terms of legal and ethical liability. Through an analysis of the current regulatory framework, including the European Union's AI Act, measures are proposed to ensure transparency and protection of fundamental rights, emphasizing the need for AI developers and operators to assume clear responsibilities to mitigate the associated risks.

Keywords: Artificial Intelligence, Hallucinations, AI Act, Risk Management, AI Regulation

Resumen

Esta tesis examina el impacto legal de las alucinaciones producidas por sistemas de inteligencia artificial (IA), definiéndolas como respuestas inexactas generadas por modelos de IA. Detalla cómo este fenómeno surge de la dependencia de los datos de entrenamiento y de la extrapolación errónea, lo que plantea desafíos significativos en términos de responsabilidad legal y ética. A través de un análisis del marco regulatorio actual, incluyendo la Ley de IA de la Unión Europea, se proponen medidas para garantizar la transparencia y la protección de los derechos fundamentales, enfatizando la necesidad de que los desarrolladores y operadores de IA asuman responsabilidades claras para mitigar los riesgos asociados.

Palabras clave: inteligencia artificial, alucinaciones, Ley de IA, gestión de riesgos, regulación de la IA

This project is dedicated to Myriam, my mother, Roberto, my younger brother, to Freyja, my loyal pet partner companion in adventure and the rest of my family, who had been my strength and encouragement in every step of this development process. Their constant support, even in the most complex moments, has been fundamental to achieving these goals with success and determination, for which I express my most sincere respect and gratitude.

To all those who are going through difficult times, this work is a testimony that, with the support of those who accompany and value us, it is possible to overcome any obstacle.

May this dedication be a reminder of the importance of perseverance and the power of the people who truly belong to us.

My deepest gratitude to those who have been part of this journey.

Acknowledgments

To my tutor, Prof. Luís Enríquez, for his invaluable time, dedication, and guidance in the development of this project, and for being a fundamental reference in this path.

To Guillermo Manuel Zamora, who, from Argentina, gave me an invaluable opportunity to project my perspective in this specialty and open new doors in the international arena.

To Paulina Casares Subía, for her constant support and for being a role model who motivated me to continue with this work. To all the members of the Ibero American Network "El Derecho Informático," whose example, support and work have been fundamental to strengthen in me the spirit of research and collaboration.

To the other teachers for their teachings and experiences. To my friends, for the good and pleasant moments.

To all of them, my deepest gratitude for their influence in this stage, which has been key to reach every achievement embodied in this thesis.

Table of contents

Introduction

The rise of Artificial Intelligence (AI) systems in recent years has ushered in unprecedented technological capabilities while simultaneously presenting novel legal challenges that demand careful examination. Among these challenges, AI hallucinations instances where AI systems generate false, misleading, or decontextualized information have emerged as a particularly concerning phenomenon with far-reaching legal implications.

While Large Language Models (LLMs) represent a prominent current example of systems prone to hallucinations, this research examines the legal impact of this phenomenon across various types of AI systems and applications where inaccurate outputs can occur.AI hallucinations occur when artificial intelligence models produce outputs that deviate significantly from their training data or intended functions, often resulting in the generation of convincing but entirely fabricated information. These hallucinations can manifest in various forms, from subtle inaccuracies in language models to potentially dangerous misidentifications in autonomous systems. The legal ramifications of such phenomena extend beyond mere technical curiosities, potentially impacting fundamental rights including the right to life, personal integrity, privacy, and honor.

The urgency of addressing AI hallucinations stems from their potential to cause tangible harm in increasingly AI-dependent sectors. Consider an autonomous vehicle misinterpreting road conditions due to hallucinations, or a medical diagnostic system generating false positives that lead to unnecessary treatments. These scenarios illustrate how AI hallucinations can directly impact human lives and well-being, raising critical questions about liability, responsibility, and the adequacy of existing legal frameworks.

This research aims to analyze the complex interplay between AI hallucinations and legal rights, focusing particularly on three fundamental rights: the right to life, personal integrity, and honor. The investigation seeks not only to identify potential violations of these rights but also to examine the legal mechanisms available for prevention and mitigation. Furthermore, this study will evaluate the effectiveness of current regulatory frameworks, with special attention to the European Union’s Artificial Intelligence Act and its approach to managing AI-related risks.

The methodological approach combines qualitative and quantitative analysis to provide a comprehensive understanding of the phenomenon. Through examination of case law, legislative developments, and technical documentation, alongside empirical data on AI system failures and their consequences, this research aims to bridge the gap between technological reality and legal protection mechanisms.

This investigation is particularly timely given the rapid deployment of AI systems across critical sectors of society. As these systems become more prevalent in decision-making processes that affect fundamental rights, understanding and addressing their potential for hallucination becomes crucial for maintaining legal certainty and protecting individual rights. The findings of this research will contribute to the development of more effective legal frameworks and risk management strategies for AI systems, ultimately promoting their responsible development and deployment.

The structure of this thesis progresses from a detailed examination of AI hallucinations and their legal implications to practical considerations for risk management and regulatory compliance. This progression allows for a systematic analysis of both the theoretical underpinnings and practical challenges of addressing AI hallucinations within legal frameworks, culminating in concrete recommendations for legal and technical measures to protect fundamental rights in an AI-driven world.

Beyond analyzing the current landscape, this research aims to lay a foundational stone for future work by identifying critical gaps, evaluating existing solutions, and proposing actionable insights. The conclusions and recommendations presented herein are intended to serve as a valuable resource and starting point for policymakers, developers, researchers, and legal professionals seeking to navigate the complexities introduced by AI hallucinations and build a more responsible and legally sound future for Artificial Intelligence.

First Chapter
Legal Implications of AI Hallucinations

1. Definition and characteristics of AI hallucinations

1.1. Brief history of artificial intelligence

The desire to create entities capable of thinking and acting like humans has deep roots in cultural and scientific history. In Ancient Egypt, ushebtis, small funerary figures, were considered magical servants working for the deceased in the afterlife, an idea that symbolizes the delegation of tasks to non-human entities. In Greek mythology, Talos, a bronze giant created by Hephaestus, patrolled Crete obeying programmed orders; a clear conceptual precursor to the automatic guardians.[1] Likewise, in Jewish tradition, the Golem, a clay figure animated by mystical incantations exemplified the aspiration to confer artificial life under human control.[2]

Modernity transformed these myths into scientific and philosophical projects. During the Scientific Revolution, Descartes compared the mind to a replicable mechanism, and in the 18th century, Jacques de Vaucanson designed automata that simulated human functions. However, it was Alan Turing who is widely regarded as the founder of modern artificial intelligence, thanks to his groundbreaking 1950 paper, "Computing Machinery and Intelligence". In this influential work, he introduced the concept that would later become known as the Turing Test. This test was designed to evaluate whether a machine could exhibit intelligent behavior indistinguishable from that of a human, effectively shifting the focus of AI discourse from abstract speculation about machine "consciousness" to tangible assessments of observable capabilities. By doing so, Turing not only redefined the way we approach AI but also laid the foundation for future advancements in the field, marking a pivotal moment in the history of technology and philosophy.

---

[1] Teun Koetsier, "A Note on Adrienne Mayor's Gods and Robots", Advances in Mechanism and Machine Science 73, n.º 1 (2019): 1187–96, doi:10.1007/978-3-030-20131-9_11..

[2] Adrienne Mayor, "Gods and Robots: Ancient Dreams of Technology", YouTube video, presented by Long Now Foundation, 2020, 07:03, https://www.youtube.com/watch?v=czj-7G6JzbQ&ab_channel=LongNowFoundation

The term “artificial intelligence” was formally introduced at the Dartmouth conference in 1956, at which time John McCarthy defined AI as “the science and engineering of creating intelligent machines”.[3] Since then, AI has evolved significantly, with advances in machine learning and data processing expanding its applicability.[4]

1.2. Definition of artificial intelligence

AI is an interdisciplinary field whose definition has evolved significantly, reflecting both its technical complexity and its practical applications. The diversity of approaches to conceptualizing AI responds to the multiple disciplines that nurture it, such as computer science, philosophy, engineering, and law.

The following are relevant definitions from authoritative sources that highlight different perspectives of the term:

a) *European Commission:* The High-Level Expert Group on Artificial Intelligence from the European Commission defines AI as “a set of systems designed to simulate human behaviors such as learning,adaptation and problem solving”.[5] This definition emphasizes the ability of AI systems to mimic essential human characteristics, highlighting the practical utility and inherent limitations of technology.

b) *UNESCO:* The UNESCO Recommendation on the Ethics of Artificial Intelligence defines AI as “technologies that use complex algorithms to analyze data, identify patterns and make predictions or decisions without direct human intervention”.[6] This definition emphasizes the role of data and algorithms in modern AI systems, linking technology with its ethical and social implications.

c) *Kaplan and Haenlein:* According to Andreas Kaplan and Michael Haenlein, “AI refers to the ability of a system to correctly interpret external data, learn from it, and use that learning to achieve specific goals through flexible

---

[3] Dartmouth Conference, “Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, Dartmouth College, August 31, 1955, num. 117.

[4] F. Escolano Ruiz y Rizo Aldeguer, “Fundamentos de inteligencia artificial”, Digitalia Publishing 1, n.o 661 (2024): 1-10, https://www.digitaliapublishing.com/a/661/fundamentos-de-inteligencia-artificial

[5] European Commission: High-Level Expert Group on Artificial Intelligence, “A Definition of AI: Main Capabilities and Scientific Disciplines”, European Commission, December 18, 2018, num. 9, https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf.

[6] UNESCO, *Recommendation on the Ethics of Artificial Intelligence,* 2021, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

adaptation".[7] Here the idea of flexibility is introduced as a distinctive attribute of AI, relevant in dynamic applications.

Although definitions of AI vary according to context, they share fundamental elements:

a) Emulation of human functions such as learning, perception and decision making.
b) The use of algorithms and data as technological pillars.
c) Autonomy as a desirable characteristic in the performance of specific tasks.

These variations reflect not only the technical evolution of the field, but also the different expectations and applications attributed to AI. Based on these definitions, it can be stated that AI is the interdisciplinary field dedicated to the development of autonomous and adaptive systems capable of processing information, learning from data and making informed decisions that emulate human cognitive capabilities, all aimed at solving complex problems in diverse environments.

1.3. Pillars of artificial intelligence

The contemporary development of AI is based on conceptual and technological architecture that has evolved significantly since its inception in the 1950s. The convergence of advances in computational capacity, availability of massive data and refinement of algorithms has allowed the consolidation of four fundamental pillars that define the current capabilities of AI systems.

These elements do not operate in isolation, but are interrelated and mutually reinforcing, creating a complex and dynamic technological ecosystem.

The development of modern AI is based on four fundamental technological pillars as Echeverría Muñoz (2020) classifies:[8]

a) *Machine Learning:* This field relies on algorithms to analyze data, recognize patterns, and improve performance in specific tasks without requiring explicit programming for every conceivable situation. By enabling systems to make data-driven decisions based on historical information, machine learning has become a fundamental pillar of modern artificial intelligence, powering

---

[7] Andreas Kaplan y Michael Haenlein, "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence", *Business Horizons* 62, n.º 1 (2019): 15-25, doi:10.1016/j.bushor.2018.08.004.

[8] Darío Echeverría Muñoz, "Derecho al honor e Inteligencia Artificial", *El Derecho Informático*, August 31, 2020, 31-5, https://issuu.com/elderechoinformatico.com/docs/revista_35.

applications ranging from recommendation systems to predictive analytics and beyond.

b) *Deep Learning:* As a specialized branch of machine learning, deep learning leverages multi-layered artificial neural networks to analyze and process vast datasets. This powerful technique enables systems to uncover intricate patterns and relationships within data, serving as the foundation for advanced technologies like facial recognition, natural language processing, and generative models that create realistic images, text, or audio.

c) *Neural networks:* Inspired by the human brain, these data structures and algorithmic processes allow systems to learn and generalize from examples. They have revolutionized areas such as speech recognition and machine translations.

d) *Natural Language Processing (NLP):* Focuses on the interaction between machines and human language, enabling systems to understand, generate and respond in natural language. This pillar is central to generative models that experience hallucinations.

1.4 Large Language Models (LLM) in AI

Large Language Models (LLMs) represent a groundbreaking advancement in the field of Artificial Intelligence (AI), situated at the intersection of Natural Language Processing (NLP) and Neural Networks. These advanced models utilize deep learning architectures to decode, interpret, and generate human language with extraordinary precision and fluency. By integrating vast amounts of textual data, LLMs can discern complex syntactic structures, semantic relationships, and pragmatic contexts, enabling them to produce coherent, contextually accurate, and grammatically refined outputs. The advent of LLMs signifies a transformative milestone in computational linguistics and AI-driven automation, redefining the capabilities of machines to understand and generate human-like text.[9]

LLMs operate on the principle of predictive text generation, where they are trained to anticipate the next word or sequence of words in each context. This is achieved through a process called training, during which the model is exposed to extensive corpora of

[99] "What Are Large Language Models (LLMs)?", *IBM*, November 2, 2023, para. 4, https://www.ibm.com/think/topics/large-language-models.

textual data. The foundation of most LLMs is the transformer architecture, which uses self-attention mechanisms to evaluate the significance of input data and produce highly contextually relevant outputs. Unlike traditional recurrent neural networks (RNNs), transformers process entire sequences of data in parallel, significantly enhancing their efficiency and capacity to handle large-scale datasets.

The purpose of LLMs is multifaceted[10]:

a) *Text Generation:* LLMs excel at producing coherent and contextually appropriate text, making them indispensable for tasks like content creation, summarization, translation, and even creative writing or dialogue generation.

b) *Information Retrieval:* These models can extract and synthesize information from vast repositories of data, facilitating tasks like document summarization and question-answering systems.

c) *Conversational AI:* LLMs drive virtual assistants and chatbots, facilitating seamless and intuitive interactions between humans and machines. These models enable natural dialogue, making technology more accessible and user-friendly for everyday tasks like customer support, personal assistance, and interactive learning.

d) *Knowledge Extraction:* By processing vast amounts of text, LLMs can uncover patterns, trends, and insights that would be either extremely time-consuming or nearly impossible for humans to detect manually.

The training process of LLMs involves several key steps[11]:

a) *Data Collection:* Gathering large and diverse datasets from sources such as books, websites, and academic papers to ensure the model is exposed to a wide range of linguistic structures and topics.

b) *Pre-training:* The model is initially trained on these datasets using unsupervised learning techniques, where it learns to predict masked words or next sentences in the text. This helps the model understand the basic structure and semantics of the language.

c) *Fine-tuning:* After pre-training, the model is further refined for specific tasks using smaller, task-specific datasets. This process employs supervised

---

[10] "What is LLM (Large Language Model)? - AWS", Amazon Web Services, February 11, 2025, par. 1, https://aws.amazon.com/es/what-is/large-language-model/.

[11] "What Is a Large Language Model (LLM)," GeeksforGeeks, January 22, 2025, https://www.geeksforgeeks.org/large-language-model-llm/.

learning, where the model is trained to excel in particular functions such as sentiment analysis, machine translation, or text classification, enhancing its precision and adaptability for targeted applications.

d) *Evaluation and Iteration:* The model's performance is evaluated using benchmark datasets, and the training process is iteratively refined to improve accuracy and robustness.

The proliferation of LLMs has led to the development of several state-of-the-art architectures, each with distinct design principles and capabilities. Notable models include:

a) *LLaMA (Large Language Model Meta AI):* Developed by Meta, LLaMA is engineered for efficiency and scalability, optimizing computational resources while maintaining high performance in natural language tasks. It is designed to handle complex reasoning tasks and generate step-by-step solutions, mimicking human problem-solving processes.

b) *DeepSeek:* This model is designed for precision in information retrieval and content generation. It enhances semantic accuracy, contextual comprehension, and factual consistency, making it ideal for applications that require high levels of precision and reliability.

c) *Flux:* Known for its dynamic response generation and adaptability, Flux excels in handling ambiguous and complex conversational scenarios. It demonstrates superior performance in maintaining coherent and contextually relevant dialogues over extended interactions.

The evolution of LLMs is characterized by a continuous refinement of architectural paradigms, algorithmic innovations, and optimization techniques. As research in this domain progresses, these models will increasingly shape the future of NLP applications, pushing the boundaries of human-machine communication.

1.5. Classification of artificial intelligence

The exponential evolution of artificial intelligence in recent decades has given rise to various taxonomies and classificatory frameworks that seek to categorize its capabilities and scope.

In the current context, where AI has permeated practically all sectors of society, from medicine to entertainment, it is essential to understand the different manifestations of this technology and their implications. This classification not only has academic relevance but is also crucial for the development of regulatory frameworks and public policies that adequately respond to the challenges that each type of AI presents.

AI has the following classification as Lateef says:[12]

a) *Weak or narrow AI:* Designed to perform specific tasks, such as virtual assistants (Siri, Alexa) or recommendation algorithms. This category currently dominates most commercial applications and lacks general awareness and understanding.

b) *Strong or general AI:* A theoretical artificial intelligence that could match the cognitive capacity of a human being, including reasoning, learning and adaptive abilities in any context. Although it is the subject of research, its development remains speculative.

c) *Artificial Superintelligence (ASI):* A hypothetical form of AI that vastly surpasses human intelligence in all respects, posing significant ethical and control challenges.

1.6. Hallucinations in AI

The term "hallucinations" describes outputs produced by AI systems that are either false, misleading, or entirely fabricated, even though they may seem coherent and convincing. These responses often lack a factual foundation and can range from invented references and inaccurate historical details to nonsensical or irrelevant answers.[13] However, the phenomenon of hallucination goes beyond mere factual inaccuracies; it includes the generation of outputs based on unreal or logically impossible parameters, which can mislead users about the capabilities of the system itself and provoke adverse real-world effects. For example, generative AI models might invent non-existent academic citations or provide erroneous medical advice.[14]

---

[12] Zulaikha Lateef, "Types of AI: Understanding Different Types of Artificial Intelligence in 2024", *Edureka* (blog), June 18, 2019, https://www.edureka.co/blog/types-of-artificial-intelligence/.

[13] Ankit, "What Are AI Hallucinations? The Complete Guide", *GeeksforGeeks*, January 24, 2025, https://www.geeksforgeeks.org/what-is-ai-hallucination/.

[14] "What Are AI Hallucinations? IBM", September 1, 2023, https://www.ibm.com/think/topics/ai-hallucinations.

The occurrence of hallucinations in AI stands as one of the most pressing challenges in the development and implementation of intelligent systems, especially in fields where precision and reliability are critical. Unlike human bias, which typically arises from cultural biases, personal experiences or identifiable cognitive limitations, hallucinations in AI emerge from the complex interplay between algorithms, training data and neural network architecture.

Human bias, by its nature, is usually consistent and can be anticipated or mitigated through specific protocols and controls. In contrast, AI hallucinations can manifest themselves unpredictably and generate results that, while consistent in their presentation, lack factual basis. This distinction is crucial for several reasons:[15]

a) *Predictability and control:* While human bias follows recognizable patterns and can be addressed through training and awareness, AI hallucinations can arise even in seemingly well-calibrated systems.
b) *Verification mechanisms:* Human biases can be contrasted with direct experience and shared knowledge. However, AI hallucinations can create completely fictitious narratives that are difficult to verify without extensive fact-checking.
c) *Systemic impact:* Human bias, while problematic, is limited by the scale of human interaction. AI hallucinations, in contrast, can propagate through automated systems, simultaneously affecting millions of users or decisions.

The credibility of these responses poses unique challenges, especially when used in contexts such as medicine, law, or government decision making. For example, errors in natural language processing can lead to serious misunderstandings in automated judicial decisions.[16]

A prominent example is State v. Loomis, in which the Wisconsin Supreme Court determined that the use of an algorithmic risk assessment tool (COMPAS) during sentencing did not infringe upon the defendant's due process rights. Eric Loomis was sentenced considering the COMPAS risk score along with other factors, which

---

[15] Claire Naughtin and Sarah Vivienne Bentley, "Both Humans and AI Hallucinate — but Not in the Same Way", *The Conversation*, June 16, 2023, http://theconversation.com/both-humans-and-ai-hallucinate-but-not-in-the-same-way-205754.

[16] Jeff Larson Mattu Julia Angwin, Lauren Kirchner, Surya, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

highlighted the importance of transparency and verification in the use of algorithmic tools in the judicial system. This case highlights how human biases can be contrasted with shared knowledge and direct experience, unlike AI hallucinations that can generate fictional narratives that are difficult to verify.[17]

In the current legal and ethical context, the distinction between human bias and AI hallucinations becomes particularly relevant in areas such as judicial decision-making, medical diagnostics, and financial analysis. The implementation of AI systems in these sectors requires not only robust verification mechanisms, but also a regulatory framework that contemplates the unique nature of algorithmic hallucinations and their potential consequences.

This involves creating robust cross-validation systems, incorporating algorithmic safeguards, and striking a balance between automation and human oversight in critical processes.

Several interrelated factors contribute to AI hallucinations, as highlighted in recent research on Large Language Models (LLMs):[18]

a) *Training Data Integrity:* The quality of training data directly influences the reliability of model outputs. If the underlying corpus contains biases, misinformation, or inconsistencies, the model may inadvertently propagate these inaccuracies.

b) *Contextual Coherence Constraints:* While LLMs exhibit impressive fluency, they often struggle with maintaining long-range contextual dependencies, leading to logical inconsistencies in extended discourse.

c) *Overconfidence in Output Generation:* LLMs frequently assign high confidence levels to incorrect or speculative responses, presenting fabrications as authoritative facts.

d) *Extrapolation from Sparse Data:* In the absence of complete information, models attempt to infer patterns, occasionally producing factually incorrect or entirely speculative statements.

e) *Logical and Semantic Inconsistencies:* Instances where models contradict themselves or produce erroneous logical conclusions, such as asserting that

---

[17] United States, State v. Loomis, No. 2015AP157-CR (Supreme Court of Wisconsin, July 13, 2016), https://law.justia.com/cases/wisconsin/supreme-court/2016/2015ap000157-cr.html.

[18] Ankit, "What Are AI Hallucinations? The Complete Guide," GeeksforGeeks, January 24, 2025, https://www.geeksforgeeks.org/what-is-ai-hallucination/.

"a pentagon has six sides," exemplify the underlying limitations in model reasoning.

This inherent unreliability in generating accurate or contextually appropriate information directly contributes to legal risks when these systems are deployed, particularly in professional domains where accuracy is paramount. A pertinent example illustrating the legal consequences arising from systems whose capabilities may not align with the claims made about them is the action taken by the United States Federal Trade Commission (FTC) against DoNotPay, Inc.[19] The FTC initiated an investigation into DoNotPay, Inc., a Delaware corporation, regarding certain acts and practices related to its service, referred to as the "DoNotPay Service" or a "Covered Product or Service," which purported to provide "Professional Services."

The FTC's Decision and Order specifically prohibited DoNotPay, Inc., and its affiliates, from making representations, express or implied, that its service operates like a human lawyer. This included claims about applying relevant laws to subscribers' situations, relying on legal expertise to avoid complications when generating legal demand letters or initiating small claims court cases, or detecting legal violations on business websites and providing advice on how to fix them. The order also prohibited misrepresentations regarding the ability of the service to analyze or evaluate documents for federal and state law violations or claims that the service would save consumers legal fees. The FTC alleged that DoNotPay lacked competent and reliable evidence to substantiate these representations.

As a result of the proceedings, DoNotPay, Inc. was ordered to pay a monetary sum to the Commission, relinquish legal and equitable rights to assets transferred, and comply with various other provisions, including notifying customers. This case highlights how the deployment and marketing of AI systems with capabilities that may be undermined by issues like hallucinations or simply lack the advertised reliability can lead to regulatory intervention, financial penalties, and mandatory changes in business practices.

It serves as a concrete illustration of the legal and ethical liability challenges that providers face when their AI systems generate outputs or are marketed with claims that cannot be reliably substantiated, directly connecting technical limitations to real-world legal consequences.

---

[19] United States, DoNotPay, Inc vs. FTC No. 232-3042 (Federal Trade Commission, January 14, 2025), https://www.ftc.gov/legal-library/browse/cases-proceedings/donotpay.

AI hallucinations raise legal and ethical liability challenges. In the context of law, the key question is: Who is liable for damages caused by a decision based on hallucination? This problem becomes relevant in sectors such as justice, where automated assessment systems may affect the fundamental rights of individuals.[20] The consequences of AI hallucinations are far-reaching, particularly in high-stakes domains:[21]

a) *Misinformation and Erosion of Trust:* Hallucinations can spread false information, undermining trust in AI systems. For example, AI-generated legal citations in court cases have led to professional and legal repercussions.
b) *Amplification of Bias:* Hallucinations in AI systems can reinforce and even magnify biases present in the training data, potentially resulting in discriminatory outcomes across areas such as hiring practices, loan approvals, and law enforcement decisions.
c) *Systemic Risks:* Unlike human bias, which is limited in scale, AI hallucinations can propagate through automated systems, affecting millions of users simultaneously. This scalability makes hallucinations particularly dangerous in applications like healthcare, finance, and autonomous vehicles.

While both AI hallucinations and human bias lead to flawed outputs, they differ fundamentally in their origins and predictability:[22]

a) *Human Bias:* Rooted in cultural, social, and cognitive factors, human bias is often consistent and can be mitigated through awareness and training.
b) *AI Hallucinations:* Stemming from technical limitations and data issues, hallucinations are unpredictable and can occur even in well-calibrated systems. They are harder to detect and verify, as they often lack any factual basis.

The need for transparency, explainability and oversight in these systems is not simply an ethical ideal, but a regulatory obligation. As AI becomes integrated into critical

---

[20] Lucrecio Rebollo Delgado, "Inteligencia artificial y Derechos fundamentales", *Digital I.A. Publishing*, 2023, https://www.digitaliapublishing.com/a/128997/inteligencia-artificial-y-derechos-fundamentales.

[21] Liz Elfman, "What Are AI Hallucinations? Examples & Mitigation Techniques", *Data world*, September 10, 2024, https://data.world/blog/ai-hallucination/.

[22] Naughtin and Bentley, "Both Humans and AI Hallucinate — but Not in the Same Way", *The Conversation*, June 16, 2023, https://theconversation.com/both-humans-and-ai-hallucinate-but-not-in-the-same-way-205754.

processes, it is imperative to anticipate and mitigate these risks, establishing a framework that ensures respect for legal principles and human rights.

2. Impact on fundamental rights such as life, honor, and privacy

AI has become a disruptive technology that, in addition to bringing significant benefits, poses concrete risks to fundamental rights. This chapter examines the risks posed by AI hallucinations —defined as incorrect or misleading outputs generated by automated systems— to fundamental rights such as life, personal integrity, and honor. These hallucinations, which emerge from AI models processing insufficient or biased data, are not merely a technical problem, but reflect profound ethical and legal challenges that demand regulatory attention.

2.1. Impact on the right to life and personal integrity

The right to life and personal integrity is a fundamental pillar in international legal systems, recognized in instruments such as the United Nations Universal Declaration of Human Rights, which states in Article 3: “Everyone has the right to life, liberty and security of person”.[23]

This right, which encompasses both the protection of physical existence and emotional and psychological integrity, faces significant risks with the integration of AI systems in critical contexts. AI hallucinations—understood as gross errors or results disconnected from objective reality—can have catastrophic consequences in situations where accuracy and safety are essential. Such risks are particularly acute in high-risk AI systems (as defined under Annex III of the EU AI Act), including autonomous vehicles, medical diagnostic tools, and military applications.[24]

Although autonomous vehicles are not explicitly listed in Annex III of the AI Act, the classification of an AI system as “high-risk” also depends on its intended purpose and regulatory context. According to the European Commission’s Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)[25],

---

[23] UN General Assembly, *Universal Declaration of Human Rights*, December 10, 1948, A/RES/217(III), Art. 3, https://www.un.org/es/about-us/universal-declaration-of-human-rights.

[24] European Union, "Artificial Intelligence Act, Regulation (EU) 2024/1689", *Official Journal of the European Union*, 2024, annex III.

[25] European Commission, “Approval of the Content of the Draft Communication from the Commission - Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act),” February 4, 2025, https://ec.europa.eu/newsroom/dae/redirection/document/112367.

AI systems that materially impact health, safety, or fundamental rights can be considered high-risk if they fall under Union harmonization legislation listed in Annex I, such as vehicle safety frameworks. Therefore, autonomous driving systems, especially those with real-time decision-making capacity—may be deemed high-risk when deployed in domains that affect public safety and fundamental rights.

This interpretation is supported by the Commission's own words:

> The prohibition in Article 5(1)(a) AI Act only applies if the harm caused by the subliminal, manipulative and deceptive techniques is 'significant'. The AI Act does not provide a definition for the concept of 'significant harm', but it should be understood as implying significant adverse impacts on physical, psychological health or financial and economic interests of persons and groups of persons. The determination of 'significant harm' is fact-specific, requiring careful consideration of each case's individual circumstances and a case-by-case assessment, but the individual effects should be always material and significant in each case. [26]

From this foundation, it is reasonable to conclude that AI systems deployed in autonomous vehicles—whose failures due to hallucinations could result in fatal accidents—clearly fall within the scope of potential significant harm. These systems, operating in real-time, are susceptible to misinterpretations of their environment (e.g., traffic signs, pedestrians, or objects), and therefore represent a tangible threat to the right to life and physical integrity.

Such systems should, consequently, be evaluated under a high-risk framework not only based on Annexes I or III, but also under Article 5 when hallucinations are involved in manipulative or deceptive outputs that impair safe decision-making, especially in dynamic physical environments like traffic.

AI systems evaluating people and determining if they are entitled to receive essential public assistance benefits and services, such as healthcare services and social security benefits, are classified as high-risk. By analogy, systems involved in the evaluation of physical environments for navigation—like those in autonomous vehicles—may fall under similar scrutiny when public safety is involved.

Under the EU AI Act, providers of these systems are legally obligated to implement risk management systems (Article 9) to identify, assess, and mitigate errors that could harm individuals' rights.[27] For example, autonomous vehicles—classified as high-risk under Annex III(1)(a)—must undergo rigorous testing to prevent algorithmic

---

[26] Ibid, 30.
[27] Ibid.

misinterpretations of traffic signs or pedestrians.[28] Similarly, Article 10 mandates data governance practices to ensure training datasets are free of errors, directly addressing the root causes of AI hallucinations in medical imaging systems.[29] These provisions aim to safeguard the right to life and personal integrity by holding developers accountable for systemic flaws that could lead to fatal accidents or misdiagnoses.

The 2025 Guidelines reinforce this obligation by clarifying that high-risk systems must not only comply technically but must also prevent deployment scenarios where fundamental rights may be impaired, particularly through hallucinations or misperceptions by the AI.

One of the most alarming contexts where AI hallucinations affect the right to life is that of autonomous vehicles. These systems, designed to make real-time decisions based on sensory data, can generate misinterpretations that endanger the safety of occupants and third parties. Documented examples include cases where systems such as Tesla's have confused traffic signs, pedestrians, or parked vehicles, causing accidents with fatal consequences.[30]

Burrell's research highlights that algorithmic opacity makes it difficult to understand the causes of these errors, as machine learning systems operate as "black boxes".[31] This lack of transparency not only complicates incident prevention but also increases the risks of automated decisions with irreversible consequences for human life. Thus, an algorithmic error resulting in a fatal accident is evidence of how a promising technology can become a direct threat to the right to life.

Accidents caused by AI hallucinations also reveal a loophole in the allocation of responsibilities. Jorqui Azofra argues that AI systems should comply with safety standards equivalent to or higher than those required of human operators, given their direct influence on fundamental rights.[32] However, in practice, the current regulatory frameworks present gaps in terms of the imputation of responsibilities: who is liable when

---

[28] Ibid.

[29] Ibid.

[30] Tom Krisher Associated Press, "EEUU investiga sistema de conducción autónoma de Tesla tras muerte de peatón arrollado", *Los Angeles Times en Español*, October 18, 2024, https://www.latimes.com/espanol/eeuu/articulo/2024-10-18/eeuu-investiga-sistema-de-conduccion-autonoma-de-tesla-tras-muerte-de-peaton-arrollado.

[31] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society* 3, n.º 1 (2016): 3-5, doi:10.1177/2053951715622512.

[32] María Jorqui Azofra, "Responsibility for Damages Caused by Artificial Intelligence Products and Systems", *Digitalia Publishing* (2023), https://www.digitaliapublishing.com/a/131264/responsabilidad-por-los-danos-causados-por-productos-y-sistemas-de-inteligencia-artificial.

an algorithmic error compromises the right to life: the vehicle manufacturer, the software developer, or the end user? This regulatory gap highlights the urgent need for a regulatory framework adapted to autonomous AI.

In addition to autonomous vehicles, AI hallucinations threaten the right to life in areas such as healthcare and the military context. In healthcare, an AI system used for diagnostics, such as radiology, can issue erroneous interpretations of medical images, leading to inappropriate treatments and putting the health and integrity of patients at risk. In extreme cases, these errors can lead to irreversible damage or death of the patient.

In the military context, the deployment of AI-based autonomous weapons increases the risks. These technologies, which make split-second decisions, lack the ability to evaluate complex ethical or legal factors, which can result in indiscriminate or disproportionate attacks. By relying on opaque algorithms, these machines create a liability gap: who responds to human losses when AI makes mistakes? The absence of clear regulations on the use of these technologies highlights the need for international agreements to regulate their design and application.

The protection of the right to life and personal integrity requires the implementation of sound legal frameworks and ethical principles governing the design and use of AI systems. The development of AI must start from the exhaustive evaluation of its potential impacts to ensure that these technologies fulfill their purpose of improving the quality of life, preventing them from becoming an added risk factor.

It is also imperative to move towards the creation of specific regulations that establish clear responsibilities in the event of fatal AI errors. These regulations must be accompanied by continuous monitoring mechanisms and standards that promote transparency in the operation of algorithms, reducing the risks associated with their opacity.

The design of secure systems must consider not only technological innovation, but also the centrality of human rights, ensuring that people's lives and integrity are protected in the face of technological failure. Only through a balance between technical progress, ethical requirements and a robust legal framework will it be possible to mitigate the inherent dangers of AI, making it an effective tool in the service of humanity.

2.2. Threats to the right to honor, image, and good reputation

The hallucinations generated by AI constitute a serious threat to the right to honor, image, and reputation, particularly when they translate into false or defamatory content

that directly affects individuals. This phenomenon, analyzed from a legal perspective, has a profound impact not only on honor, understood as a subjective assessment of one's own dignity, but also on honor and reputation, which reflects how a person is perceived socially. According to Echeverría Muñoz (2020), these three concepts, although related, have essential nuances: honor is linked to the internal perception of one's own worth, while honor and reputation transcend the social sphere and shape the public recognition of a person in his or her environment.[33]

In the context of AI, hallucinations, such as the generation of fake news or deep fakes, amplify the risks through digital platforms that facilitate the mass dissemination of false information. For example, the case in which Sam Altman, creator of OpenAI, developed a GPT voice model that realistically imitated actress Scarlett Johansson, without her consent, illustrates how these technologies can infringe on the rights of individuals, even those with high public exposure.[34]

Such a simulation, although technically impressive, poses significant risks to the image and reputation of the actress, as her voice could be associated with statements or content that do not belong to her, affecting both her personal and professional sphere. This case highlights the potential of AI tools to create fictitious scenarios that are indistinguishable from reality, increasing the damage to the reputation and honor of the affected people.

A particularly revealing example in the legal sphere is the case Mark Walters v. OpenAI, L.L.C.[35], in which the plaintiff, a prominent radio host and Second Amendment advocate, was affected by a hallucination generated by ChatGPT. Walters, known for reaching more than one million listeners per broadcast, was indirectly implicated when a journalist, Frederick Riehl, used ChatGPT to summarize a lawsuit filed by the Second Amendment Foundation (SAF) against the Attorney General of Washington. Although ChatGPT initially produced accurate summaries when provided with direct excerpts from the complaint, it subsequently delivered a different and false narrative when the same journalist provided only a URL link to the publicly available document. The system,

---

[33] Darío Echeverría Muñoz, "The Right to Honor, Dignity, and Good Reputation: Background and Constitutional Regulation in Ecuador", *Ius Humani. Revista de Derecho* 9, n.º 1 (2020): 209–30, https://doi.org/10.31207/ih.v9i1.228.

[34] Juan Ríos, "El gran engaño que vivió Scarlett Johansson con la IA y el uso de su voz con ChatGPT", *Infobae*, August 28, 2024, https://www.infobae.com/tecno/2024/08/28/el-gran-engano-que-vivio-scarlett-johansson-con-la-ia-y-el-uso-de-su-voz-con-chatgpt/.

[35] United States, *Walters v. OpenAI, L.L.C.,* No. 1:23-cv-03122 (District Court, N.D. Georgia, December 31, 2024), https://www.courtlistener.com/docket/67617826/walters-v-openai-llc/.

despite disclaimers regarding its limitations, generated fictitious factual content that misrepresented elements of the lawsuit and its participants, thereby introducing serious reputational risks.

This example is particularly significant because the defamatory or incorrect information was not based on manipulated media (such as voice or image imitation), but rather on false factual assertions produced by the AI itself. The fact that ChatGPT altered the nature of the output depending on the format of the input—textual versus URL—demonstrates not only the volatility of AI-generated content, but also the unpredictability of its narrative authority in contexts of legal relevance. Such inaccuracies, when disseminated by credible figures or platforms, can mislead the public and damage the professional credibility of the individuals involved. As Walters was not even the subject of the original lawsuit but was introduced incorrectly into its context by the AI's hallucination, the case reveals the expanded scope of potential victims and the indirect pathways through which reputational damage can occur.

Moreover, this case underscores the challenges associated with algorithmic opacity. As the court noted, it remains technically unclear why ChatGPT produced a completely different result from the URL input compared to the pasted text. This inability to trace or audit the generative process not only obstructs accountability but also complicates legal recourse for those harmed. The case exposes a legal and ethical vacuum in which victims of AI hallucinations may find themselves unable to correct the record or seek compensation, especially when platform disclaimers shift the burden of verification to end users.

The viral nature of these defamations aggravates the problem. Through social networks and digital media, reputational damage spreads exponentially, generating immediate consequences that are difficult to counteract. In these cases, traditional legal mechanisms, such as actions for libel or slander, are not sufficiently agile to offer effective protection. This creates an imbalance between the speed of digital attacks and the capacity of existing laws to guarantee the defense of the honor and reputation of affected individuals.

In addition, the specific regulatory vacuum to address these issues makes it difficult to redress damage caused by AI hallucinations. Although the EU's IA Act establishes risk categories for AI systems, concrete measures to prevent and mitigate the effects of hallucinations in the honor setting have not yet been developed. According to Burrell's analysis, the algorithmic opacity of these systems intensifies violations, as it

makes it difficult to trace the source of false information and to hold developers or users of tools that generate harmful content accountable.

In this context, the protection of honor, image and reputation requires a comprehensive strategy that combines preventive and corrective measures. On one hand, it is crucial to introduce robust auditing and transparency mechanisms for generative systems, alongside establishing clear regulatory frameworks to curb the creation and spread of false or misleading content. On the other hand, it is urgent to develop agile legal procedures that allow affected individuals to act swiftly to stop the spread of false information and repair the damage caused.

Likewise, the ethical approach promoted by institutions such as UNESCO stresses the need for technologies to respect fundamental values, prioritizing human dignity over uncontrolled innovation. This ethical approach must be complemented by a technological culture that fosters responsibility in the development and use of AI tools, ensuring that individual freedoms and fundamental rights are not compromised.

AI hallucinations represent a significant challenge for the protection of the right to honor in the digital age. Combating these threats requires a multidisciplinary approach that contemplates both legal regulation and the implementation of ethical and technological principles. Only through this combination will it be possible to effectively protect the dignity and social recognition of individuals from the risks inherent in artificial intelligence technologies.

2.3. Hallucinations and the right to privacy

The right to privacy faces critical challenges in the face of errors derived from hallucinations in AI systems, posing risks that go beyond technical inaccuracy, directly affecting data protection, privacy, and the dignity of individuals. In this context, international regulations such as the General Data Protection Regulation (GDPR) and the Artificial Intelligence Regulation (AI Act) of the European Union become relevant to establish control and correction parameters.

Article 22 of the GDPR enshrines the right of individuals to avoid being subjected to decisions made solely through automated processing, particularly those that result in legal consequences or significantly impact them in a comparable manner. This principle underscores the importance of ensuring that automated decisions are transparent, explainable, and based on accurate data. However, AI hallucinations can distort this

balance, as automated decisions derived from erroneous data compromise the fairness of the process.

Article 9 of the AI Act mandates that developers of high-risk systems adopt measures to ensure the quality of data used in model training. This provision is especially critical in contexts where automated decisions may impact fundamental rights, as models trained on inaccurate or biased data are more likely to produce discriminatory outcomes or hallucinations.[36]

Article 14 builds on this obligation by mandating that AI systems be designed to be transparent and verifiable. This means they must be capable of explaining the logic behind their automated decisions, thereby enabling effective human oversight and clearer accountability.[37] This approach aligns with the principles of the GDPR, particularly Article 22, which governs decisions based solely on automated processing and emphasizes the need for additional safeguards to protect individuals' rights.

On the other hand, Article 15 of the AI Act requires effective human supervision during the use of high-risk systems, which seeks to mitigate the risks derived from hallucinations. This supervision is essential in sectors such as health, justice, or financial services, where erroneous decisions can have irreparable consequences.[38]

For example, a hallucination generated by an AI model in a medical context could lead to a misdiagnosis, affecting a person's health and violating the accuracy principle set forth in Article 5(1)(d) of the GDPR. This type of situation demonstrates the intersection between the two regulations, highlighting the need to integrate common principles to address the risks posed by hallucinations.

Moreover, AI hallucinations frequently undermine the principle of "data minimization" outlined in Article 5(1)(c) of the GDPR, as the generation of inferred or fabricated data can lead to the accumulation of excessive and irrelevant information. In contexts such as recruitment, where AI systems are employed to evaluate candidates, hallucinations can generate false assumptions about an individual's qualifications, leading to unfair hiring decisions that impact their privacy and dignity. Such errors raise questions about whether current safeguards are sufficient to prevent the unintended amplification of biases or the misuse of personal information.

---

[36] European Union, Artificial Intelligence Act, Regulation (EU) 2024/1689, *Official Journal of the European Union*, 2024, art. 15.

[37] European Union, Regulation (EU) 2016/679, *Official Journal of the European Union*, 2016, art. 5.

[38] Ibid.

Recent cases show how hallucinations generate real consequences for individuals. For example, in 2023, a lawyer in the United States used ChatGPT to prepare a case, but the system cited non-existent precedents, affecting the lawyer's reputation, and compromising the quality of the legal defense. This event evidence how reliance on unsupervised AI systems can undermine professional and personal rights, affecting both privacy and procedural fairness.[39]

Similarly, hallucinations have been shown to create issues in the journalism sector, where AI-generated misinformation has misattributed quotes, leading to defamation claims and the dissemination of false information to the public. Such incidents not only harm individual reputations but also undermine public trust in AI-mediated content. This erosion of trust directly impacts on the perception of privacy, as individuals feel less secure about how their personal information might be manipulated in the public sphere.

In this regard, Section 61 of the AI Act introduces a framework for regular auditing and ongoing monitoring of AI systems classified as high risk. This provision is a significant development, as it recognizes that AI systems are not static and may generate unexpected results due to changes in data or model degradation over time.

However, while the technical approach of the AI Act is essential, it needs to be complemented by more robust and specific legal measures for redressing harm. Current regulations require transparency and verifiability, but do not address how to hold AI developers, users, or providers accountable when hallucinations are generated that violate fundamental rights. This regulatory vacuum generates legal uncertainty, hindering the effective protection of affected rights.

For example, hallucinations can also affect the right to privacy by inferring sensitive data that individuals have not provided. Amnesty International has documented how the use of facial recognition technologies by Israeli authorities, specifically through the system known as Red Wolf, has led to mistaken identifications resulting in unjustified arrests of innocent individuals in the occupied Palestinian territories.[40] This system, designed to surveil the Palestinian population, not only compromises privacy, but also

---

[39] Carlos Prego, "Un abogado usó ChatGPT en un juicio. Ahora es él quien debe dar explicaciones a un juez por incluir citas falsas", *Xataka*, May 29, 2023, https://www.xataka.com/legislacion-y-derechos/abogado-uso-chatgpt-juicio-ahora-quien-debe-dar-explicaciones-a-juez-incluir-citas-falsas.

[40] Amnesty International, "Report: Israeli Authorities Are Using Facial Recognition Technology to Entrench Apartheid", Amnesty International Australia, May 1, 2023, https://www.amnesty.org.au/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/.

perpetuates an environment of coercive and discriminatory surveillance, exposing individuals to undue control by the authorities.

Beyond surveillance, hallucinations also risk creating "chilling effects" on free expression, as individuals may self-censor or avoid certain digital platforms for fear that their interactions could be misinterpreted or inferred inaccurately by AI systems. These effects are particularly concerning in authoritarian regimes, where governments may exploit AI-generated hallucinations to silence dissent or justify invasive monitoring. Amnesty International underscores that "the overreach of AI-driven surveillance systems has introduced new coercive tools that erode both freedom of expression and the right to privacy, particularly in heavily surveilled regions

These situations illustrate how hallucinations can exacerbate inequalities and erode trust in AI systems. While regulations such as the GDPR and the AI Act seek to mitigate these risks, it is crucial to adopt a comprehensive approach that provides for effective redress mechanisms, continuous monitoring, and regular audits. This will not only ensure respect for fundamental rights but also foster shared responsibility among stakeholders.[41]

It is imperative to strengthen coordination between the GDPR and the AI Act to close the identified regulatory gaps. Creating a joint framework that integrates effective oversight, principles of fairness and clear accountability mechanisms will enable the risks associated with hallucinations to be addressed more effectively. Only through a coherent and balanced regulatory approach will it be possible to ensure that technological advances do not translate into new threats to fundamental rights, but rather into tools that reinforce the dignity and privacy of individuals.

[41] Ibid.

Second Chapter
Legal risk management of AI hallucinations

1. Regulatory framework: EU Artificial Intelligence Act and comparative law

Artificial intelligence (AI) systems' hallucinations, defined as unexpected, false or fabricated outputs by algorithmic models, present significant risks in the legal field including misinformation, reputational damage and decisions based on incorrect data, etc. Given this reality, the development of an appropriate regulatory framework is essential. In this context, the European Union's Artificial Intelligence Regulation (AI Act) establishes a regulatory benchmark that seeks to mitigate these risks. However, a comparative analysis with other international regulatory initiatives shows complementary and divergent approaches in the management of these challenges.

1.1. The EU Artificial Intelligence Regulation

The EU Artificial Intelligence Regulation (AI Act) marks a pivotal step in the governance of emerging technologies, introducing the first comprehensive legal framework to tackle the challenges and risks posed by artificial intelligence (AI). This landmark regulation sets out foundational principles, targeted regulatory tools, and a robust enforcement mechanism designed to prevent and mitigate the risks stemming from AI hallucinations and other potential issues.

1.1.1. Parties Involved

Effective implementation of the AI Act depends on close collaboration among several key stakeholders:

a) *Provider:* The provider refers to the individual or organization responsible for developing an AI system or introducing it to the market under their name or trademark, as well as overseeing its deployment. As the primary party accountable for ensuring compliance with the AI Act, the provider plays a critical role in upholding regulatory standards.

Its main responsibilities include:

– Adhering to the essential safety and fundamental rights requirements outlined in the AI Act, tailored to the risk level of the system.

- Conducting conformity assessments prior to market introduction or system deployment.
- Implementing and maintaining a robust risk management framework.
- Creating and preserving comprehensive technical documentation for the system.
- Ensuring the traceability of system operations and outputs throughout its lifecycle.
- Establishing a post-market monitoring mechanism to address ongoing compliance and performance.
- Collaborating effectively with relevant national authorities to ensure regulatory alignment.
- For high-risk systems, appointing an authorized representative within the EU if the provider is not established in the Union.

b) *Distributor:* A distributor refers to any individual or entity within the supply chain, other than the provider, responsible for making an AI system available on the market. While the distributor does not modify the system, they play a key role in ensuring its compliance with regulatory requirements.

Its responsibilities are as follows:

- Verifying that the AI system displays the CE marking and is accompanied by the required documentation.
- Confirming that the provider has fulfilled all necessary obligations under the AI Act.
- Retaining the required documentation for the specified period as mandated by the regulation.
- Cooperating fully with the relevant national authorities to ensure compliance.
- Notifying the provider or competent authorities if there is reason to believe that an AI system fails to meet the requirements of the AI Act.

c) *Importer:* An importer is defined as any individual or entity established within the European Union who introduces an AI system originating from a third country into the EU market. The importer plays a critical role in ensuring that these systems meet all applicable regulatory requirements before they are made available to users.

Its main responsibilities are the following:

- Ensuring that the provider, if not established in the Union, has fulfilled all their obligations under the AI Act.
- Verifying that the AI system carries the CE marking and is accompanied by the required documentation.
- Retaining a copy of the technical documentation and the EU declaration of conformity for the specified period as required by law.
- Cooperating actively with the relevant national authorities to ensure compliance with regulatory standards.
- Clearly indicating their name, registered trade name or trademark, and contact address on the AI system, its packaging, or in accompanying documentation.

d) Deployer: The deployer is the entity responsible for implementing an AI system within a specific context, integrating it into a particular process, service, or activity. This role is crucial in ensuring that the system operates effectively and safely in real-world applications.

Its main attributes include:

- Ensuring the system is used for its intended purpose and in strict accordance with the provider's instructions.
- Implementing appropriate technical and organizational measures to mitigate identified risks, tailored to the specific context of use.
- Monitoring the system's performance during deployment, including providing human oversight when necessary to address potential issues.
- Maintaining the system and applying updates or patches provided by the provider to ensure continued functionality and security.
- Ensuring compliance with personal data protection regulations, such as those outlined in the GDPR, to safeguard user privacy.
- For high-risk systems, adhering to Article 26 of the AI Act, which emphasizes the importance of proper human supervision and the relevance of input data during the deployment phase.[42]

e) User: A user refers to any individual or entity that operates an AI system under their authority, excluding cases where the system is employed for personal,

[42] European Union, *Artificial Intelligence Act, Regulation* (EU) 2024/1689, Official Journal of the European Union, 2024, art. 26.

non-professional activities. Users play a critical role in ensuring the proper and responsible use of AI systems.

His/her main responsibilities are the following:

- Using the AI system in strict accordance with the instructions provided by the provider.
- Monitoring the system's operation and promptly reporting any incidents or malfunctions to the provider and/or relevant authorities.
- For high-risk AI systems used in workplace settings, adhering to specific obligations related to informing and consulting workers, as outlined in applicable regulations.

f) Market Supervisory Authorities: These are the national bodies appointed by each Member State to oversee the implementation and enforcement of the AI Act within their respective territories. They play a pivotal role in ensuring compliance and addressing violations effectively.

Their main responsibilities include:

- Monitoring the market to ensure adherence to the provisions of the AI Act.
- Conducting investigations and carrying out inspections to identify potential non-compliance.
- Imposing appropriate sanctions in cases where violations of the Act are detected.
- Collaborating with authorities from other Member States and with the European Commission to ensure a coordinated and consistent approach to regulation.

g) European AI Board: The European AI Board is the central body responsible for coordinating the implementation of the AI Act at the European level. Its role is essential in fostering consistency, collaboration, and alignment across Member States in regulating artificial intelligence.

Its main attributes are the following:

- Issuing guidelines and recommendations to ensure the uniform application of the AI Act.
- Coordinating the supervisory activities of national authorities to promote a cohesive regulatory framework.

- Advising the European Commission on matters related to artificial intelligence and its regulation.
- Contributing to the development of harmonized technical standards in collaboration with standardization organizations.
- Engaging in international cooperation to address global AI-related challenges.
- Publishing annual reports detailing its activities and providing opinions on specific topics of relevance.
- Maintaining a public registry of high-risk AI systems that have undergone conformity assessments.
- Collaborating with other bodies, institutions, and agencies of the European Union to ensure alignment and effectiveness.
- Facilitating cooperation and the exchange of information among Member States to enhance regulatory consistency.

The AI Act introduces a comprehensive regulatory framework that assigns specific responsibilities to various stakeholders, ensuring the safe and ethical development and deployment of artificial intelligence across the EU. This multilevel governance system is reinforced by national authorities overseeing market compliance and the European AI Board coordinating efforts at the EU level. The successful implementation of the AI Act hinges on each actors understanding and fulfilling their obligations, striking a delicate balance between fostering innovation and safeguarding fundamental rights.

The AI Act adopts a stratified approach based on risk levels, classifying AI systems into four main categories: unacceptable risk, high risk, limited risk, and minimal risk. This methodology allows regulations to be adjusted according to the potential impact of each technological application, promoting the protection of fundamental rights without stifling innovation.

a) *Unacceptable risk:* According to Article 5 of the IA Act, this category includes those AI systems whose mere existence or application is considered incompatible with the values and fundamental rights recognized in the European Union. Preemptive prohibition is based on the high probability that these systems generate irreparable or systematic damage to human dignity,

freedom, equality, and non-discrimination.[43] It refers to AI practices that "exploit any of the vulnerabilities of a specific group of people due to their age, physical or mental disability" or that "are used for social scoring by public administration authorities."

An example of this type of risk involves an AI system used by an insurance company to determine premiums based on the analysis of applicants' social media posts, discriminating against those with political views deemed "dissident" or belonging to certain social groups. This violates the right to non-discrimination and freedom of expression.

b) *High risk:* Article 6 defines the general criteria for the classification of an AI system as high risk, while Annex III lists the specific areas where AI systems are high risk.[44] This category comprises AI systems that, while not prohibited per se, present a high potential for significant harm to critical areas of social life. Regulation focuses on prevention by requiring pre-market conformity assessment and implementation of an ongoing risk management system.

An example within this category involves an AI system used for the widespread use of facial recognition cameras in a city to monitor citizens' movements, without clear and proportional justification. This violates the right to privacy and personal data protection.

c) *Limited risk:* Under the provisions of Article 52, transparency obligations are detailed for AI systems that interact with natural people or generate manipulated content. [45] This category focuses on information asymmetry between the user and the AI system. Transparency obligations are put in place to ensure that users are aware of the interaction with an automated system and can make informed decisions.

An example within this category involves the use of a chatbot used by a customer service company, who must clearly inform the user that they are interacting with an automated system and not a human agent.

d) *Minimal or No Risk:* Recital 14 of the AI Act clarifies that most current AI systems fall into this category, which largely encompasses AI systems that do

[43] Ibid, art. 5.
[44] Ibid, art. 6.
[45] Ibid, art. 52,

not pose significant risks to fundamental rights or security.[46] They are allowed to be used freely, although they are still subject to horizontal legislation on data protection, intellectual property, etc.

An example of this type of risk can be seen in tools like spam filters for email inboxes or movie recommendation systems on streaming platforms. These AI-driven systems are designed to enhance the user experience, yet they do not pose a significant threat to users' rights or security.

The AI Act adopts a risk-based approach to regulating artificial intelligence, tailoring its requirements according to the potential impact of each system. This framework ranges from outright bans on unacceptable practices that directly violate fundamental rights to the unrestricted use of systems classified as having minimal risk. This gradation, which ranges from requiring compliance assessments and ongoing risk management for high-risk systems, through transparency obligations for those of limited risk, to no specific restrictions for those of minimal risk, seeks a balance between protecting society and encouraging innovation, tailoring regulatory requirements to the likelihood and severity of potential harm.

1.1.2. Conformity Assessment

To ensure compliance of high-risk AI systems, the AI Act introduces a comprehensive framework for conducting conformity assessments tailored specifically to these systems. This process transcends "initial testing" and is comprised of tired and differentiated procedures depending on the type of system and the use of harmonized standards spanning several stages:

a) *Evaluation Procedures:* Article 43 of the IA Act details the conformity assessment procedures.[47] If a provider has utilized harmonized standards or common specifications in the development of a high-risk AI system listed in Annex III, they may choose from the following options:

– *Internal control (Annex VI):* A self-assessment process conducted by the provider itself.

[46] Ibid, recital 14.
[47] Ibid, art. 43.

- *Assessment of the quality management system and technical documentation with the participation of a notified body (Annex VII):* A more rigorous process involving an independent accredited third party. The latter is mandatory when the provider has not implemented the harmonized standards or has done so partially, or if the IA system is intended for use by law enforcement or immigration authorities.

The role of the notified body is fundamental, as it assesses the system's conformity with the essential requirements of Chapter II, Section 2 of the IA Act. These requirements cover critical aspects such as risk management, data governance, technical documentation, traceability, accuracy, robustness, and cybersecurity. This external assessment ensures an objective and expert review, which is essential to build confidence in the system.

In addition, the AI Act establishes a post-market monitoring system (market surveillance), which is not limited to oversight to maintain an acceptable level of accuracy and safety. Competent national authorities are authorized to carry out investigations, request relevant information, perform tests, and, in cases of non-compliance, implement corrective measures. These measures may include withdrawing the system from the market or imposing a ban on its distribution. This mechanism ensures that AI systems remain within established safety and ethical parameters, even after they have been launched on the market.

A central focus of the AI Act is its strong emphasis on risk management, which must be actively addressed at every stage of an AI system's lifecycle. This involves identifying, analyzing, evaluating, and mitigating potential risks associated with the system, including issues such as "hallucinations" in language models, algorithmic bias, privacy violations, and security vulnerabilities. By adopting this proactive approach, the regulation aims to minimize the potential adverse impacts of AI deployment while fostering trust and safety in its use.

Finally, traceability is a fundamental requirement, as stated in Section 11 of the AI Act. A comprehensive record of system development, validation and operation is required, allowing for accountability and incident investigation. Audits, conducted primarily by notified bodies, ensure that these requirements are met, promoting transparency and accountability.

b) *Adoption Measures:* The adoption measures outlined in the AI Act require precise implementation aligned with its provisions:

- *Technical Documentation (Annex IV):* The AI Act mandates comprehensive technical documentation that includes, among other elements, an overview of the AI system, its architecture, the data used for training and validation, conformity assessment processes, and documentation on the risk management system. As Kaminski notes citing Gianclaudio Malgieri and Giovanni Comandé, this documentation is crucial for transparency and accountability in automated systems.[48]
- *Risk Management:* This is not a one-time task but an ongoing process that must be integrated into every phase of the system's lifecycle. It involves identifying and mitigating specific risks, such as "hallucinations" in language models, as well as continuously monitoring the effectiveness of the measures put in place to address these challenges.
- *Human Oversight:* High-risk AI systems must be designed to enable meaningful human oversight, including features that allow operators to intervene or override system decisions when necessary. Providing adequate training and raising awareness among human operators are essential steps to ensuring effective control and maintaining accountability.
- *Post-Market Monitoring:* This process is not a one-time activity but an ongoing surveillance mechanism that involves systematically collecting and analyzing data on system performance after its market introduction. Providers are obligated to report any serious incidents or malfunctions to the relevant authorities, ensuring timely identification and resolution of potential issues.

As argued by Wachter and Mittelstadt, transparency and explainability are crucial elements for building trust in AI systems and ensuring their responsible use.[49] Through its requirements for technical documentation, traceability, risk management, and human oversight, the AI Act seeks to promote these principles.

---

[48] Margot E. Kaminski, "The Right to Explanation, Explained", *Berkeley Technology Law Journal* 34, no. 1 (November 26, 2019): 190–218, https://doi.org/10.15779/Z38TD9N83H.

[49] Sandra Wachter and Brent Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI", *SSRN Scholarly Paper* (Rochester, NY: Social Science Research Network, 2018): 7–12, https://papers.ssrn.com/abstract=3248829.

Furthermore, the proactive approach to risk management and the post-market surveillance system reinforces safety and ethics in the development and use of AI. Collectively, these measures not only ensure adherence to regulatory requirements but also promote the responsible integration of AI technologies. They are designed to align with the highest standards of transparency, accountability, and the protection of fundamental rights, fostering trust and confidence in their use.

1.2. The Brussels Effect: The Global Influence of the EU's AI Act on Artificial Intelligence Regulation

The regulation of artificial intelligence (AI) worldwide is undergoing a phenomenon that could be termed the "Brussels Effect." This concept refers to the growing influence of the European Union's (EU) regulatory approach, as embodied in its Artificial Intelligence Act (AI Act), on other jurisdictions around the globe. While each region adapts its regulations according to its political, cultural, and economic priorities, the EU's risk-based approach, centered on protecting user rights and freedoms, is becoming a global benchmark. This phenomenon underscores the importance of addressing AI challenges comprehensively while balancing innovation with citizen protection.

1.2.1. The Risk-Based Approach as a Global Standard

The EU's AI Act is distinguished by its risk-based approach, which classifies AI systems according to their potential impact on fundamental rights and safety. This model has inspired other countries and regions to adopt similar regulatory frameworks, albeit with local variations. For instance, Canada's Artificial Intelligence and Data Act (AIDA) incorporate principles such as transparency, accountability, and risk assessment, reflecting the influence of the EU's approach.[50]

The risk-based approach is especially significant in contexts where AI is deployed in critical applications, such as automated decision-making, healthcare, and public safety. As Cath highlights, this framework allows regulators to prioritize resources effectively by categorizing AI systems into distinct risk levels—from minimal to unacceptable—and

[50] Canada, *Digital Charter Implementation Act*, 2022, First Reading, Bill C-27 (House of Commons), 44th Parliament, 1st Session, June 16, 2022, art. 5.

establishing requirements that are proportionate to the potential impact of each technology.[51] [52]

This method not only safeguards users but also offers developers and businesses clear guidance on regulatory expectations. The increasing delegation of complex, high-risk processes to AI systems—such as granting parole, diagnosing medical conditions, and managing financial transactions—introduces significant challenges. For instance, questions arise regarding liability for autonomous vehicles, the adequacy of current legal frameworks in addressing the disparate impacts of big data, and the prevention of algorithmic harms.

Given the far-reaching impact of AI, addressing these pressing questions requires a multidisciplinary approach to ensure comprehensive and effective solutions. Furthermore, it is essential to critically examine who is shaping AI governance and what motivations or potential benefits might drive the actions of these individuals or organizations.

1.2.2. Protecting Rights and Freedoms: A Shared Objective

A central pillar of the Brussels Effect is the prioritization of protecting user rights and freedoms. The AI Act establishes specific safeguards for high-risk AI systems, such as those used in hiring, credit scoring, or mass surveillance. These safeguards include mandatory impact assessments, algorithmic transparency, and human oversight in critical decisions.[53]

This approach has resonated with other jurisdictions, where ethical concerns about AI are gaining traction. For example, in the United States, while Executive Order 14110adopts a more flexible, innovation-focused approach, it also includes provisions to protect civil rights and prevent algorithmic discrimination.[54] However, this order was revoked by President Donald Trump on January 20, 2025, as part of his initial actions upon assuming his second term. Trump justified this revocation by arguing that Biden's

---

[51] Corinne Cath, "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018): 2–4, https://doi.org/10.1098/rsta.2018.0080.

[52] Ibid.

[53] European Union, *Artificial Intelligence Act,* Regulation (EU) 2024/1689, Official Journal of the European Union, 2024, art. 14.

[54] United States, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, The White House, October 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

order imposed unnecessary barriers to AI innovation and promoted a social agenda that, in his view, limited technological development and the global competitiveness of the United States.

Trump's decision reflects a significant shift in U.S. AI policy, moving from a regulatory approach that prioritized safety, transparency, and equity to one that emphasizes deregulation and the promotion of innovation to maintain global leadership in AI. This shift has raised concerns about the potential increase in risks such as algorithmic bias, misinformation, and cybersecurity vulnerabilities, given that Biden's order established requirements like security testing (red teaming) and impact assessments for high-risk AI systems.

Furthermore, the revocation of Executive Order 14110 has created a regulatory vacuum at the federal level, which could lead to fragmentation in state-level regulations and complicate compliance for companies operating in multiple jurisdictions. In response to mounting concerns over AI-generated harms—particularly the proliferation of malicious deepfakes and non-consensual intimate imagery—the U.S. Senate approved the TAKE IT DOWN Act (whose full name is "Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act") on May 19, 2025,[55] introducing a targeted legal framework to address one of the most egregious manifestations of AI misuse.

The U.S. TAKE IT DOWN Act, is not a general AI regulation. Instead, it focuses on a specific, harmful outcome that can be produced or significantly exacerbated by generative AI capabilities, including the generation of deepfakes: the publication of non-consensual intimate visual depictions. This law addresses a manifestation of potential AI hallucinations (or malicious uses of AI generation) in the form of harmful content, rather than regulating the AI system that created it.

Its mechanism for dealing with this specific problem is reactive and content-focused:

h) **Direct Legal Mandate for Platforms:** The Act places a direct legal obligation on "covered platforms" to establish a process for individuals to notify the platform about the non-consensual intimate visual depiction and request its removal.

---

[55] United States, Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (TAKE IT DOWN ACT), 119th Congress, S.146 § (2025), https://www.congress.gov/bill/119th-congress/senate-bill/146/text.

i) **Expedited Removal Requirement:** Upon receiving a valid notification and request, the covered platform is legally required to remove the depiction and make reasonable efforts to remove identical copies "as soon as possible, but not later than 48 hours."

Therefore, while the EU AI Act tackles the problem of AI hallucinations broadly by imposing preventative requirements on the systems themselves to minimize their occurrence across various contexts, the TAKE IT DOWN Act provides a specific, targeted, and rapid legal tool for the removal of a particular type of harmful content that can be a direct or indirect result of AI generative capabilities, including deepfakes that may appear as hallucinations (false depictions).

It is a mechanism focused on remediation and content control post-publication for a defined harm, rather than a system-level approach to improve overall AI reliability and prevent hallucinations in the first place. This difference highlights a fragmented approach in the U.S. compared to the EU's comprehensive strategy, but provides a specific, legally backed mechanism for addressing a particularly egregious form of harmful AI-generated content.

However, this approach contrasts sharply with the European Union's Artificial Intelligence Act (AI Act), which adopts a comprehensive, risk-based strategy to preemptively mitigate AI hallucinations and systemic risks through systemic design, data governance, and human oversight requirements. Together, these developments underscore a growing divergence in global AI governance: the EU's proactive, rights-centered model versus the U.S.'s fragmented, reactive measures, raising critical questions about balancing innovation, accountability, and societal trust in an era of rapidly evolving AI capabilities.

1.2.3. Toward Global AI Governance

The "Brussels Effect" highlights the need for global AI governance that combines robust regulatory frameworks with international collaboration. For this approach to be effective, it is essential to:[56]

a) *Harmonize standards:* Establish common criteria for risk assessment and AI system classification.

---

[56] Corinne Cath, "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018): 1–8, https://doi.org/10.1098/rsta.2018.0080.

b) *Promote training:* Educate professionals in ethics, regulation, and technical evaluation of AI systems.
c) *Ensure transparency:* Guarantee that algorithms are explainable and that users understand how decisions affecting them are made.
d) *Strengthen international cooperation:* Create mechanisms for sharing best practices and coordinating responses to global challenges, such as AI-generated disinformation or algorithmic bias.

The "Brussels Effect" reflects the growing influence of the EU's regulatory approach on global AI governance. By prioritizing risk management and the protection of user rights and freedoms, this model provides a robust framework for addressing the ethical and legal challenges of AI. However, its long-term success will depend on the ability of countries to adapt it to their specific contexts and work together to build a future where AI benefits society.

However, it is important to critically assess the Brussels Effect, particularly in the context of technological innovation. While the EU AI Act sets a strong ethical and legal foundation, adopting it wholesale in jurisdictions with different socio-economic, technological, or industrial conditions may be counterproductive. For countries still in the process of digital transformation or where innovation ecosystems are emerging, overly rigid or burdensome regulatory frameworks could stifle innovation rather than foster safe development.

In fact, growing concerns are being raised within the EU itself about the AI Act's potential to hinder competitiveness, especially for startups and small enterprises lacking the resources to comply with complex regulatory demands. As such, using the EU as a global regulatory model must be approached with nuance, recognizing that a one-size-fits-all approach may not accommodate local contexts or development priorities. In this defense, while the EU can serve as a useful reference, its regulatory framework should not be uncritically adopted in jurisdictions where flexibility and innovation are essential for technological growth.

2. Role of control authorities and regulated entities in the protection of rights

The phenomenon of hallucinations in AI represents a significant challenge in the integration of these technologies across various sectors. This phenomenon poses considerable risks, especially in areas where precision, reliability, and safety are critical.

In the realm of public administration, this issue takes on a critical dimension due to its ethical, social, and legal implications, requiring a strategic and well-coordinated response from responsible institutions.

Misuraca and Van Noordt emphasize that "Consequently, it is the combination of software and hardware with human behavior that possibly leads to an impact, however measured, or to a transformative change of the previous external conditions".[57] This perspective underscores the importance of addressing hallucinations not only as a technical problem but also as a social and ethical challenge. In this regard, the EU AI Act seeks to establish clear standards for the development and implementation of AI systems, aiming to minimize these risks and ensure that their use is transparent and responsible.

The obligations of the involved actors are key to addressing this phenomenon. Developers must design robust and transparent systems capable of reducing the occurrence of hallucinations. Deployers, on the other hand, must ensure that these systems are properly integrated into administrative processes, respecting principles such as fairness and non-discrimination. Finally, end users, including public officials and citizens, must be aware of the limitations of AI and receive training to interpret and use its results critically.

Among the proposed strategies to mitigate the risks associated with AI hallucinations are the implementation of continuous auditing and monitoring mechanisms, the promotion of transparency in algorithms and decision-making, and collaboration between the public and private sectors to share best practices. Misuraca and Van Noordt stress that "Thus, in order to understand the effects of AI use in governments, an approach which takes into consideration how the technology is used will give more insights on how AI provides impact".[58]

In the field of justice, judicial decision-support systems, such as the well-known COMPAS system referenced in section 1.5 of the first chapter of this thesis, have demonstrated biases that perpetuate structural inequalities. These systems, designed to assess the risk of recidivism, have shown a tendency to favor discriminatory decisions, particularly to the detriment of minority groups. A hallucination in this context—that is, the generation of erroneous or biased information by AI—could compromise essential

---

[57] Gianluca Misuraca and Colin Van Noordt, "AI Watch Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU", *Publications Office of the European Union* (2020): 15, https://doi.org/10.2760/039619.

[58] Ibid.

guarantees such as due process and the presumption of innocence. This would not only affect the individuals directly involved but also erode the legitimacy and trust in the judicial system.

As Veale and Zuiderveen Borgesius point out, "The Draft AI Act's provisions on emotion recognition and biometric categorization seem insufficient to mitigate the risks. A recent literature review concluded that, '[i]t is not possible to confidently infer happiness from a smile, anger from a scowl, or sadness from a frown, as much of current technology tries to do when applying what are mistakenly believed to be the scientific facts".[59] This quote reflects the authors' concern about the lack of scientific basis in certain AI applications, such as emotion recognition, and how this can lead to biased or incorrect decisions.

In the context of justice, the lack of transparency in these systems hinders accountability and perpetuates injustices that could be avoided with a more critical and regulated approach. AI systems used in judicial decision-making must undergo rigorous audits to ensure they do not perpetuate discriminatory biases. Additionally, it is essential that operators of these systems understand their limitations and that affected citizens have access to mechanisms for review and appeal.

In other critical sectors such as healthcare, the implications of AI hallucinations are equally concerning. Artificial intelligence systems used for medical diagnoses or therapeutic recommendations could make errors that endanger human lives, thereby violating the right to health and personal integrity. Misuraca and Van Noordt emphasize that, while AI has the potential to revolutionize medicine, its implementation must be accompanied by rigorous oversight and effective corrective measures to avoid catastrophic consequences. In their report, they state that "the potential benefits of AI technologies are massive, but risks must also be governed while democratic values and human rights respected. For this reason, the EU in particular, aims to develop 'trusted AI' based on truly European ethical and societal values borrowed from the European Charter of Fundamental Rights".[60]

---

[59] Michael Veale and Frederik Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act", *SSRN Scholarly Paper* (Rochester, NY: Social Science Research Network, 2021): 11, https://papers.ssrn.com/abstract=3896852.

[60] Gianluca Misuraca and Colin Van Noordt, "AI Watch Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU", *Publications Office of the European Union* (2020): 9, https://doi.org/10.2760/039619.

This approach underscores the importance of developing trustworthy AI, grounded in ethical principles and respect for fundamental rights, as established in the Charter of Fundamental Rights of the European Union. Inaccurate AI-generated diagnoses or the recommendation of unsuitable treatments not only jeopardize patient health but also raise profound ethical and legal concerns regarding accountability for such errors. This scenario reinforces the need to establish clear protocols and continuous auditing mechanisms to ensure that these systems operate safely and reliably, ensuring that AI in the healthcare sector aligns with the principles of transparency, accountability, and the protection of human rights.

In the realm of fundamental rights, the opacity of machine learning algorithms poses serious risks, particularly regarding non-discrimination and transparency. As Burrell notes, "The opacity of algorithms, according to Pasquale, could be attributed to willful self-protection by corporations in the name of competitive advantage, but this could also be a cover for a new form of concealing sidestepped regulations, the manipulation of consumers, and/or patterns of discrimination".[61] This lack of transparency makes it difficult to identify biases in automated systems, which can lead to discriminatory decisions that disproportionately affect vulnerable groups. In contexts such as credit scoring, job recruitment, or content classification on social media, algorithmic opacity can perpetuate existing inequalities and undermine the right to fair and equitable treatment. Therefore, addressing this opacity is crucial to ensuring that automated systems respect the fundamental rights of individuals.

2.1. The Need for Transparency and Algorithmic Openness

Algorithmic opacity is a factor that further complicates these challenges. As Burrell notes, "Opacity seems to be at the very heart of new concerns about 'algorithms' among legal scholars and social scientists. The algorithms in question operate on data. Using this data as input, they produce an output; specifically, a classification (i.e., whether to give an applicant a loan, or whether to tag an email as spam)".[62] This lack of transparency makes it difficult to detect errors and assign accountability, creating an environment where AI errors can go unnoticed or, worse, be ignored. This increases the

[61] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", Big Data & Society 3, n.° 1 (2016): 4, doi:10.1177/2053951715622512.

[62] Ibid.

risk of violations of fundamental rights without consequences for those who design or implement these systems.

These challenges in critical areas such as justice, healthcare, and security highlight the urgency of addressing not only algorithmic transparency but also a broader and deeper concept: algorithmic openness. While transparency refers to the ability to understand and audit AI decision-making processes, algorithmic openness goes a step further, promoting the accessibility and public scrutiny of AI systems, as well as the participation of multiple stakeholders in their design and evaluation.

2.1.1. Transparency in AI

Algorithmic transparency is a fundamental requirement to ensure that AI systems are fair, equitable, and accountable. As the Ada Lovelace Institute points out, it is essential that algorithmic systems "are doing the 'right thing': that they behave as we expect, that they are fair and do not unlawfully discriminate, that they are consistent with regulation, and that they are furthering, not hindering, societal good".[63]

Additionally, bias audits allow for the evaluation of algorithmic systems without needing access to their internal code, as "they generally don't look at the code of the system. Instead, they compare the data that goes into the system with the results that come out".[64] This is especially crucial in high-risk contexts, such as judicial or medical decision-making, where errors can have devastating consequences. To address these risks, algorithmic risk assessments aim to be holistic, evaluating not only the data or the model itself but also "how it will be used in practice and how users and the wider public will interact with or be affected by it".[65] Finally, regulatory oversight is key to ensuring that systems comply with regulations, as "a regulatory inspection could be used to assess whether an algorithmic system complied with data protection law, equalities legislation, or insurance industry requirements, for instance".[66]

However, transparency alone is not enough. As Burrell warns, "the claim that algorithms will classify more 'objectively' (thus solving previous inadequacies or injustices in classification) cannot simply be taken at face value given the degree of

---

[63] Ada Lovelace Institute, "Examining the Black Box", *Ada Lovelace Institute*, 2020, 6, https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

[64] Ibid, 9.

[65] Ibid, 15.

[66] Ibid, 12.

human judgment still involved in designing the algorithms, choices which become built-in".[67] This underscores the need to avoid blindly trusting the objectivity of algorithms, especially when they are designed with inherent biases. Even when algorithms are technically transparent, their mathematical complexity and the scale of the data they process can make their functioning remain incomprehensible to most people.

2.1.2. Algorithmic Openness

Algorithmic openness goes beyond transparency in the functioning of artificial intelligence (AI) systems; it also involves the active participation of multiple stakeholders in their design, implementation, and evaluation. This approach recognizes that responsibility for AI systems cannot rest solely on developers but must involve regulators, ethicists, civil society representatives, and, ultimately, the citizens affected by these technologies.

The EU AI Act reflects this principle by establishing strict requirements for transparency and accountability in high-risk AI systems. For example, the law mandates that providers of high-risk AI systems design and develop their technologies in a way that allows individuals to oversee their operation and ensure they are used as intended.[68] Additionally, it promotes the creation of independent oversight mechanisms, such as the AI Board and a scientific panel of independent experts, which play a crucial role in risk assessment and issuing qualified alerts.[69]

The AI Act actively promotes the involvement of diverse stakeholders —such as industry players, small and medium-sized enterprises (SMEs), civil society organizations, and academic institutions— in shaping the governance of AI systems. This is evidenced by the establishment of a consultative forum that provides technical input and recommendations to the European Commission and the AI Board. This collaborative approach guarantees that the oversight of AI systems is not limited to developers alone but actively engages a diverse range of stakeholders in the decision-making process.

Furthermore, the law states that deployers of high-risk AI systems, particularly those operating in the public sector or delivering public services, conduct thorough

---

[67] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society* 3, n.° 1 (2016): 3, doi:10.1177/2053951715622512.

[68] European Union, Artificial Intelligence Act, Regulation (EU) 2024/1689, Official Journal of the European Union, 2024, recital 48.

[69] Ibid, recital 76.

fundamental rights impact assessments prior to implementation.[70] This requirement reinforces algorithmic openness by ensuring that risks associated with AI are evaluated transparently and with the participation of independent experts, enabling more robust and responsible oversight.

Moreover, algorithmic openness not only seeks to ensure transparency and accountability in AI systems but also fosters a culture of participation and shared responsibility among all stakeholders involved. As Wachter and Mittelstadt note, "the underlying problem goes much deeper and relates to the tension of whether individuals have rights, control, and recourse concerning how they are seen by others".[71]

This approach underscores the importance of empowering citizens to understand and challenge automated inferences while promoting the creation of fairer and more equitable AI systems. However, for these principles to be effectively realized, the active role of public administration is indispensable as both a guarantor of fundamental rights and a facilitator of a regulatory framework that ensures the responsible implementation of these technologies.

As a mediator between developers, regulators, and society, public administration bears the responsibility of ensuring that AI systems not only meet technical and ethical standards but also safeguard citizens' rights in an era of rapid technological advancement.

---

[70] Ibid, recital 58.

[71] Sandra Wachter and Brent Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI", *Columbia Business Law Review* 2019, no. 2 (October 5, 2018): 1–130, https://papers.ssrn.com/abstract=3248829.

2.2. Role of Public Administration as a Guarantor

Once regulations are implemented, public administrations must ensure continuous oversight and monitoring of AI systems to guarantee they operate in accordance with ethical and legal principles. This is particularly important in the context of hallucinations, as errors in AI systems can have serious consequences for individuals' fundamental rights.

The governance of AI should be viewed as an evolution of existing regulatory tools, ensuring that ethical and societal implications are effectively addressed through comprehensive and robust frameworks. This approach of continuous oversight is essential to identify and mitigate risks associated with hallucinations, especially in critical systems such as those used in justice, healthcare, and security.

Oversight must be adaptive and proactive, meaning that public institutions must be prepared to address emerging challenges in a timely manner. For example, if a pattern of hallucinations is detected in an AI system used in the justice sector, authorities must act quickly to correct the issue and prevent further harm. This may include suspending the system until the necessary corrections are made. In this regard, Misuraca and Van Noordt emphasize that public sector organizations must adopt a proactive approach to AI governance. This includes implementing robust monitoring mechanisms to identify and address potential risks before they escalate into significant challenges.[72] This proactive approach not only minimizes risks associated with hallucinations but also strengthens public trust in AI systems.

Additionally, continuous oversight should involve the participation of multiple stakeholders, both within and outside public administration. This includes collaboration with AI experts, civil society organizations, and other stakeholders to ensure that AI systems are evaluated from diverse perspectives. The active involvement of diverse stakeholders in the monitoring process is essential to ensure that AI systems align with societal values and avoid perpetuating biases or errors. This multidisciplinary collaboration becomes particularly critical when addressing issues like hallucinations, as it enables the detection and correction of errors that might otherwise remain undetected in more narrowly focused evaluations.

---

[72] Gianluca Misuraca and Colin Van Noordt, "AI Watch Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU," *Publications Office of the European Union* (2020): 51, https://doi.org/10.2760/039619.

Finally, continuous oversight and monitoring must be supported by mechanisms of transparency and accountability. Public administrations are responsible for ensuring that the outcomes of AI system evaluations are transparent and publicly accessible, as well as for implementing corrective measures whenever necessary. Transparency and accountability are essential pillars of AI governance, ensuring that citizens can trust the systems that impact their lives. This approach not only reinforces the legitimacy of AI systems but also promotes a culture of responsibility in the use of these technologies.

2.3. Promotion of Intersectoral Collaboration

AI has proven to be a transformative tool across various fields, from healthcare to public administration. However, hallucinations pose significant risks, especially in contexts where precision and reliability are critical. To address this issue, intersectoral collaboration among public, private, and civil society actors emerges as an indispensable mechanism, as it not only enables the sharing of knowledge and resources but also facilitates the design of inclusive and sustainable solutions to mitigate associated risks.

The technical and ethical complexity of AI hallucinations requires a multidisciplinary approach involving diverse stakeholders. Public administrations alone lack the technical capacity to fully understand and regulate these systems. Therefore, collaboration with the private sector, which possesses the necessary technical expertise and resources, and with civil society, which can provide a human rights and equity-focused perspective, is essential.

One of the main benefits of intersectoral collaboration is the exchange of knowledge and best practices. Public administrations can learn from the private sector's experiences in developing and implementing AI systems, while civil society organizations can offer critical perspectives on the social and ethical impacts of these technologies.

Citizen participation is a key component of AI governance. The end users of these systems are on the front lines of detecting and reporting hallucinations or errors. Therefore, it is essential for public administrations to establish accessible and effective feedback mechanisms that allow citizens to report their concerns.

To facilitate intersectoral collaboration, the creation of specialized bodies that act as connection points among different stakeholders is recommended. These bodies could be composed of experts in technology, law, ethics, and social sciences, and would be

responsible for overseeing the development and implementation of AI systems in the public sphere.

An inspiring example is the proposed Artificial Intelligence Observatory by the European Union, which aims to foster cooperation among member states, technology companies, and civil society organizations. Such initiatives not only facilitate the exchange of knowledge but also promote the adoption of common standards and best practices at an international level.

Intersectoral collaboration is a fundamental pillar for addressing the challenges posed by AI hallucinations. By fostering collaboration among public administrations, the private sector, and civil society, it becomes possible to design inclusive and sustainable solutions that maximize the benefits of AI while mitigating its risks. This collaborative approach not only strengthens the ability of public administrations to oversee and regulate these systems but also promotes transparency, accountability, and public trust in AI technologies.

In this regard, the creation of feedback mechanisms, the exchange of knowledge and best practices, and the formation of specialized bodies are key strategies for promoting effective AI governance. Only through collaborative and multidisciplinary efforts can we ensure that this technology serves the common good and contributes to the development of fairer and more equitable societies.

2.4. Accountability and Sanctions

In the context of the phenomenon of hallucinations in AI systems, the EU AI Act establishes a regulatory framework that assigns clear responsibilities to its various stakeholders: developers, operators, and users of AI systems, with the aim of mitigating associated risks and ensuring the protection of fundamental rights and freedoms.

2.4.1. Developer Accountability

Developers of AI systems have the primary obligation to ensure that their models are trained with high-quality data and undergo rigorous testing before implementation. According to Article 9 of the AI Act, developers must ensure that their models comply with safety and ethical standards, especially in critical contexts such as medical diagnosis or judicial decision-making. This includes the need to train models with high-quality data and subject them to rigorous testing before deployment. High-risk AI systems, such as those used in medical or judicial applications, must meet strict requirements, including

conformity assessments, human oversight, and transparency in their functioning, as explained in this work.[73]

The EU AI Act underscores the importance of developing high-risk AI systems using training, validation, and testing datasets that are relevant, representative, and error-free, thereby ensuring alignment with ethical and safety standards. These datasets must also adhere to rigorous data governance practices, incorporating measures to identify and address biases that could potentially impact fundamental rights or result in discrimination. For instance, the Act requires that data preparation processes, such as annotation and labeling, be carefully managed to avoid introducing biases that could compromise the system's integrity.

Furthermore, the Act requires providers of high-risk AI systems to establish a comprehensive risk management system that operates throughout the entire lifecycle of the system. As stated in Article 9 of the EU AI Act, "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems." This system must incorporate iterative testing and validation processes to ensure continuous compliance with regulatory requirements. Providers are also required to establish a quality management system that documents design choices, data collection processes, and procedures for addressing data gaps or shortcomings.

In addition to the technical requirements, the EU AI Act establishes transparency obligations for AI systems that interact directly with humans. For instance, users must be explicitly informed when they are engaging with an AI system, and any AI-generated content must be clearly identified as such. This is particularly important in contexts where AI systems influence decision-making processes, such as in healthcare or law enforcement.

The Act also establishes a governance framework to ensure consistent enforcement across EU member states. This includes the creation of an AI Office responsible for overseeing general-purpose AI models and developing guidelines for compliance. National authorities are tasked with enforcing the Act's provisions, with the support of the European AI Board, which facilitates coordination and information sharing among member states.

---

[73] European Union, Artificial Intelligence Act, Regulation (EU) 2024/1689, Official Journal of the European Union, 2024, Art. 9.

In summary, the EU AI Act sets a high standard for developer accountability, particularly for high-risk AI systems. As noted by Kempf & Rauer, "Providers and deployers of so-called 'high-risk' AI systems will be subject to significant regulatory obligations when the EU AI Act takes effect, with enhanced thresholds of diligence, initial risk assessment, and transparency".[74] By requiring rigorous data governance, risk management, and transparency, the Act aims to ensure that AI systems are safe, ethical, and aligned with fundamental rights. The European Parliament further clarifies that "AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories".[75] These measures are critical for fostering public trust in AI technologies and mitigating the risks associated with their deployment.

Article 15 of the AI Act sets forth mandatory technical standards to ensure that high-risk AI systems maintain sufficient levels of accuracy, robustness, cybersecurity, and resilience throughout their entire lifecycle.[76] To this end, AI systems must be designed to function consistently in dynamic environments. This involves avoiding errors resulting from unforeseen interactions with users or other systems. For example, a medical diagnostic model must maintain its accuracy even when data input varies due to regional differences in symptomatology or the quality of medical images. The requirement for technical robustness is linked to the concept of algorithmic robustness as a pillar of AI ethics.

The European Commission, in collaboration with technical authorities, will promote standards for measuring accuracy and robustness. This approach recognizes the complexity of evaluating AI systems, where traditional metrics (e.g., overall accuracy) may hide biases in subgroups. The standardization of inclusive metrics, as proposed in Article 15(2), would prevent these systemic failures.

The requirement to declare verifiable accuracy levels in the user instructions of high-risk AI systems is based on two pillars: functional transparency and the protection of fundamental rights. The Court of Justice of the European Union (CJEU) established in Schrems II (C-311/18) that technical opacity in automated systems can constitute a

---

[74] Anna-Lena Kempf and Nils Rauer, "A Guide to High-Risk AI Systems under the EU AI Act", Pinsent Masons, February 13, 2024, https://www.pinsentmasons.com/out-law/guides/guide-to-high-risk-ai-systems-under-the-eu-ai-act.

[75] "EU AI Act: First Regulation on Artificial Intelligence", Topics | European Parliament, June 8, 2023, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[76] European Union, Artificial Intelligence Act, Regulation (EU) 2024/1689, Official Journal of the European Union, 2024, Art. 15.

violation of the right to data protection and non-discrimination, as it prevents users from understanding how decisions affecting them are made.[77, 78] This ruling underscore the lack of clarity in the performance of AI systems, especially in sensitive contexts such as employment, can perpetuate structural biases, as demonstrated by the 2018 Amazon case, where a recruitment algorithm discriminated against female candidates due to historical patterns in training data.[79]

The technical resilience required by Article 15(4) of the AI Act involves two dimensions: operational redundancy and bias control in continuous learning cycles. The case of Microsoft Tay in 2016 illustrates the risks of not implementing these mechanisms. Tay, a chatbot designed to learn from interactions on Twitter, was manipulated by malicious users who injected racist and misogynistic messages, creating a negative feedback loop that distorted its behavior in less than 24 hours.[80] This episode highlighted the need for "dynamic ethical filters" and real-time audits, as proposed by UNESCO in its recommendations for AI ethics.[81]

The AI Act requires developers to implement measures such as:

a) *Technical Redundancy:* Backup systems to mitigate critical failures, like aviation safety protocols.
b) *Mitigation of Feedback Loops:* Mechanisms to prevent biases in model outputs from contaminating new training data. For example, in credit scoring systems, an algorithm that underestimates the income of ethnic minorities could perpetuate exclusions if its impact on future cycles is not monitored.

This approach advocates for integrating ethical responsibilities into the technical architecture of AI.

In the context of cybersecurity, the requirements of Article 15(5) of the AI Act address emerging threats such as data poisoning and adversarial examples, which

---

[77] Christopher Kuner, "The Schrems II Judgment of the Court of Justice and the Future of Data Transfer Regulation", *European Law Blog* (2020): para. 1, https://doi.org/10.21428/9885764c.aed20daf.

[78] Maria Helen Murphy, "Assessing the implications of Schrems II for EU–US data flow", *International & Comparative Law Quarterly* 71, n.° 1 (2022): 245–62, https://doi.org/10.1017/S0020589321000348.

[79] Avi Perera, "AI Ethics Case Studies: Lessons Learned from Real-World Failures – Avi Perera", October 30, 2024, https://aviperera.com/ai-ethics-case-studies-lessons-learned-from-real-world-failures/.

[80] Oscar Schwartz, "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation - IEEE Spectrum", January 4, 2024, https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[81] UNESCO, *Recommendation on the Ethics of Artificial Intelligence,* 2021, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

compromise the integrity of AI systems. A study on medical diagnosis using an AI model demonstrated that minimal alterations (0.1% of pixels) in medical images can deceive breast cancer diagnostic models, leading them to classify benign tumors as malignant in 69.1% of cases.[82] This type of attack not only endangers human lives but also erodes trust in AI-based medical tools.

The AI Act mandates specific technical measures, such as:

a) *Detection of Model Poisoning:* Implementation of algorithms to identify corrupted data during training.
b) *Defenses Against Adversarial Examples:* Techniques such as adversarial training, where models are exposed to manipulated data during development to improve their robustness.

An applied example is the use of redundant neural networks in medical imaging systems, which compare multiple models to detect inconsistencies caused by attacks. These practices align with the GDPR's "security by design" principle, which prioritizes risk prevention over post-hoc remediation.[83]

The example of a medical diagnostic system generating hallucinations due to incomplete data illustrates how non-compliance with Article 15 can escalate to serious violations under Article 99. Suppose the system was primarily trained on data from European patients, underrepresenting populations with distinct genetic profiles (e.g., sub-Saharan Africans or East Asians). This could lead to misdiagnoses of diseases such as breast cancer, where breast density varies significantly among ethnic groups.

The penalty of 3% of global turnover (Article 99(4)) would apply if the error is due to negligence in data selection. However, if the same system is used in prohibited contexts, such as the cognitive manipulation of minors through false diagnoses that alter their perception of reality, Article 99(3) would be triggered, with fines of up to 7%. This gradation reflects the principle of proportionality in penalties, aligned with the jurisprudence of the Court of Justice of the EU, which distinguishes between unintentional harm and malicious uses.

As Misuraca and Van Noordt note:

---

[82] Qianwei Zhou, Margarita Zuley, Yuan Guo, Lu Yang, Bronwyn Nair, Adrienne Vargo, Suzanne Ghannam, Dooman Arefan, and Shandong Wu, "A Machine and Human Reader Study on AI Diagnosis Model Safety under Attacks of Adversarial Images," *Nature Communications* 12, no. 1 (2021): 7281, https://doi.org/10.1038/s41467-021-27577-x.

[83] European Union, Regulation (EU) 2016/679, Official Journal of the European Union, 2016.

> The difficulty of operationalizing such high-level definitions is evident, though, even more so when dealing with AI use in the public services. In fact, machine learning techniques or predictive models do not interact per se with the world around them but only as embedded in existing software or hardware. Studying the development and use of algorithmic models in the public sector is worthwhile but only shows a narrow view of the algorithms themselves and not how they are embedded into existing infrastructure and work practices.[84]

The AI Act achieves a balance between fostering innovation and ensuring ethical safeguards by recognizing that developers, as the architects of complex systems, must take on both technical and societal responsibility.

2.4.2. Responsibility of Users and Deployers

Institutions deploying high-risk AI systems (such as hospitals, courts, or government agencies) assume legal and ethical co-responsibility under the framework of the AI Act. Their role is not limited to the technical use of technology but also includes ensuring that its implementation respects fundamental rights and avoids social harm. This approach aligns with the OECD principle that" AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art".[85] This recommendation emphasizes that deployers, as end users, must mitigate operational and ethical risks arising from the technology through verifiable accountability mechanisms.

Article 26 of the AI Act establishes critical obligations for deployers of high-risk AI systems, aiming to mitigate risks arising from the gap between testing environments and operational reality, including AI "hallucinations". This article requires continuous monitoring of system performance and immediate reporting of incidents to the relevant authorities, recognizing that AI systems may degrade or produce erroneous results when interacting with changing data or unpredictable contexts. These provisions seek to ensure that deployers not only deploy the technology but also take active responsibility for its post-market evolution, adapting to emerging social, regulatory, and technical dynamics.

The AI Act requires deployers to ensure that personnel understand the limitations of AI and maintain human control over critical decisions. This obligation is linked to the

---

[84] Gianluca Misuraca and Colin Van Noordt, "AI Watch Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU", *Publications Office of the European Union* (2020): 14, https://doi.org/10.2760/039619.

[85] OECD, *Recommendation of the Council on Artificial Intelligence,* OECD Legal Instruments, 2025, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

concept of "meaningful oversight" developed by UNESCO in its Recommendation on the Ethics of AI (2021), which prohibits the complete delegation of ethical or legal judgments to automated systems.

A paradigmatic case is the use of AI in courts to assess recidivism risks. In 2016, the COMPAS system, used in the U.S., demonstrated racial biases by incorrectly predicting recidivism rates among African Americans, as explained in section 1.5 of this thesis. Under the AI Act, a European court that ignores similar bias alerts would be considered negligent, as Article 26 requires training judges to critically interpret AI outputs and contrast them with human evidence.

Deployers are required to notify individuals whenever a decision impacting them is made or supported by AI, ensuring they can contest it. This right is rooted in Article 22 of the GDPR, which prohibits decisions based solely on automated processing without meaningful human intervention.[86] The AI Act expands this protection by requiring accessible, non-technical explanations of how the system operates.

Deployers must monitor the performance of AI systems in real-world environments and report incidents to the relevant authorities. This requirement is based on the need to address the gap between testing environments and operational reality, a phenomenon exemplified by the ruling of the District Court of The Hague in NJCM v. State of the Netherlands.[87] In this case, the SyRI system —used to detect social fraud— was declared illegal for violating the right to privacy under the European Convention on Human Rights (Article 8), as it operated opaquely and without effective human oversight mechanisms in real-world environments. When an AI hallucination affects fundamental rights, deployers may be held co-responsible if negligence in oversight is proven.

To mitigate this, the AI Act adopts a proportional approach: only high-risk systems (e.g., judicial, medical) are subject to strict obligations. Additionally, Recital 37 emphasizes the need for cooperation between developers and deployers to share technical knowledge. This collaborative governance model not only reduces institutional burdens but also establishes a framework of shared responsibility that avoids unilateral blame-shifting in cases of AI hallucinations. By linking proportionality with operational

---

[86] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[87] Christiaan van Veen, "Landmark Judgment from the Netherlands on Digital Welfare States and Human Rights", *OpenGlobalRights*, March 19, 2020, https://www.openglobalrights.org/landmark-judgment-from-netherlands-on-digital-welfare-states/?lang=English.

transparency —as required by Article 26— the AI Act ensures that critical systems are auditable and adaptable to changing contexts, aligning with the OECD's accountability principle and UNESCO's ethical standards. Thus, the European regulatory framework not only prevents harm but also fosters technological innovation aligned with human rights.

2.4.3. Analysis of Shared Responsibility Among Stakeholders

Article 99 of the AI Act establishes a differentiated and proportional sanctioning regime based on two pillars: the objective severity of the infringement, linked to the harm caused to fundamental rights (Art. 99(7)(a)), and the subjective degree of diligence, which evaluates the technical and organizational measures implemented by each operator (Art. 99(7)(g)). This dual approach rejects absolute strict liability, instead adopting a negligence-based model (Art. 99(7)(i)). As stated in Recital 145 of the AI Act, "sanctions should reflect whether the operator acted intentionally, with gross negligence, or with slight negligence". This framework allows for a fairer and more context-specific attribution of responsibility, avoiding disproportionate sanctions that could discourage innovation.

The attribution of responsibility requires demonstrating a causal link between the hallucination and the operator's failure to fulfill their duties. For developers, this involves proving that they violated safe design obligations (Art. 16), such as training models with unrepresentative data (Art. 10), leading to systemic biases. For deployers, it means demonstrating that they neglected post-deployment controls (Art. 26), such as monitoring outcomes or training end users. Article 99(7)(j) prioritizes harm remediation: if an operator proves they implemented immediate corrective measures (e.g., updating the model after detecting biases), they could reduce their fine by up to 40% according to the Commission's guidelines.[88]

Additionally, the AI Act applies the principle of "those who can do more, must do more," which translates into a graduated sanctioning system based on each operator's capacity for control. Developers of high-risk AI systems (Annex III) face fines of up to 7% of their global turnover for violating Article 5 (Art. 99(3)), due to their greater technical expertise. In contrast, deployers (e.g., hospitals or employers) are fined up to

[88] European Union, Artificial Intelligence Act, Regulation (EU) 2024/1689, Official Journal of the European Union, 2024, art. 99.

3% (Art. 99(4)(e)), based on their duty of contextual oversight. An illustrative case would be algorithmic discrimination in bank credit systems: the developer could receive a 5% fine for failing to audit biases in historical data (Art. 10), while the bank would face a 2% fine for not adjusting risk thresholds for vulnerable groups (Art. 26).

Article 99(7) establishes nine criteria for modulating sanctions, three of which are particularly relevant in practice. First, cooperation with authorities (subsection f) can reduce the fine by up to 30% if the operator shares key information, such as training logs. Second, the impact on fundamental rights (subsection a) determines the severity of the sanction: systems affecting human dignity (Art. 5(1)(d)) receive fines 50% higher than those affecting only intellectual property. Finally, the size of the operator (subsection d) influences the fine amount; for example, an SME with a turnover below €50 million could pay 1% instead of 3% for the same infringement (Art. 99(6)).

Although the AI Act unifies substantive criteria, its enforcement depends on national authorities, creating significant risks. One major risk is fragmentation in sanction quantification: Article 99(9) allows national courts to set fines, which could lead to forum shopping (companies relocating to countries with more lenient sanctions). Additionally, the burden of proof poses a challenge: demonstrating the "degree of responsibility" (Art. 99(7)(g)) requires costly technical expertise, which may be inaccessible to many SMEs. These obstacles could undermine the effectiveness of the sanctioning regime, especially in cross-border cases where coordination between authorities is limited.

The AI Act's sanctioning model surpasses traditional approaches in three key aspects. First, it promotes dynamic prevention through tiered fines (Art. 99(3)-(5)), incentivizing self-regulation and proactive compliance. Second, it prioritizes comprehensive harm remediation: Article 99(7)(j) states that corrective measures and victim compensation must be considered before imposing financial sanctions. Finally, it fosters transparency by requiring member states to publish annual reports on imposed sanctions (Art. 99(11)), deterring opaque practices and promoting accountability.

The AI Act constructs a hybrid sanctioning paradigm that combines quasi-strict liability for developers (with maximum fines of 7%) and negligence-based liability for deployers (capped at 3%). This model reflects a balance between the need to protect fundamental rights and the promotion of technological innovation. However, its effectiveness faces two critical challenges: the asymmetry between the technical resources of national authorities and the complexity of AI systems, and the risk of SMEs bearing disproportionate compliance costs (Art. 99(1)). The success of the sanctioning

regime will depend on its ability to adapt to technical advancements without sacrificing legal certainty.

2.4.4. Sanctions and Legal Consequences

The AI Act establishes a sanctioning system designed to ensure compliance with its provisions and deter practices that jeopardize fundamental rights. This regime is based on principles of proportionality, effectiveness, and deterrence, as outlined in Article 99. Sanctions vary depending on the severity of the infringement, the type of operator, and their level of responsibility.

- *Civil Implications:* Administrative sanctions do not preclude civil actions for damages. Individuals affected by AI hallucinations may file lawsuits against developers, providers, or deployers to seek compensation for the harm suffered.

The AI Act does not establish a specific civil liability regime, but the provisions of Article 99(10) allow victims to file civil claims based on national legislation. For example, if an AI system used by a bank discriminates against certain groups on racial grounds, the affected individuals could sue the bank and the provider for negligence, using the violation of Article 50 (transparency) as evidence of fault.

In many cases, civil liability may be shared among the various actors in the supply chain. For instance, if a distributor markets an AI system without verifying its compliance with AI Act standards, they could be held jointly liable if the system causes harm due to a hallucination.

- *Criminal Implications:* Although the AI Act primarily focuses on administrative sanctions, it does not exclude the possibility of criminal actions in severe cases. Member States may classify malicious uses of AI as offenses under their national criminal codes.

In cases where AI hallucinations result in significant harm to fundamental rights, developers and users could face criminal penalties. For example, if an AI system is used to commit large-scale fraud or manipulate election results, those responsible could be charged with offenses such as computer fraud or data manipulation.

The AI Act does not establish direct criminal sanctions, but Member States must ensure that their national laws are compatible with the regulation's provisions. This includes classifying serious AI-related misconduct and imposing significant fines or prison sentences.

- *Administrative Implications:* The administrative sanctioning regime established in the AI Act is one of the key pillars for ensuring compliance with AI regulations. These sanctions aim not only to punish offenders but also to deter practices that may endanger fundamental rights.

Article 99(7) sets out a series of factors that competent authorities must consider when determining the number of fines. These factors include the nature and severity of the infringement, the number of affected individuals, the extent of harm suffered, and the level of cooperation from the operator with the authorities.

For example, in a case where an AI system used in a hospital generates erroneous diagnoses that harm patients, the sanction could vary depending on whether the developer acted intentionally or negligently. If it is proven that the developer concealed recurring errors in the system, the fine could reach 7% of the annual global turnover under Article 99(3).

In addition to financial penalties, the AI Act allows for non-pecuniary measures, such as withdrawing the product from the market or prohibiting the development or use of AI systems in the future. These measures are particularly relevant in cases where AI hallucinations pose a significant risk to public safety or fundamental rights.

Article 99 of the AI Act establishes a three-tier sanctioning system that links the number of administrative fines to the level of risk posed by the infringements. This proportional approach seeks to balance the deterrence of unlawful conduct with the need to avoid disproportionate economic burdens, especially for small and medium-sized enterprises (SMEs). The three sanction levels are structured as follows:

- *Severe Infringements (Article 99(3)):* The most serious violations, such as deploying prohibited AI systems (e.g., real-time biometric identification without judicial authorization), carry fines of up to €35 million or 7% of the annual global turnover, whichever is higher. This sanction level reflects the gravity of practices that directly undermine fundamental rights such as privacy and individual freedom. For example, the unauthorized use of facial recognition systems in public spaces could incur maximum penalties due to their potential to erode citizen autonomy and security.
- *Non-Compliance with Technical Obligations (Article 99(4)):* Failures to meet technical and ethical requirements, such as inadequate conformity assessments or omissions in AI system transparency (Article 50), are penalized with fines of up to €15 million or 3% of the annual turnover. A

paradigmatic case would be a provider failing to document the limitations of a medical diagnostic model, leading to erroneous diagnoses that violate the right to health.

- *Misleading Information or Omissions (Article 99(5):* Providing false or incomplete information to competent authorities is penalized with fines of up to €7.5 million or 1% of the annual turnover. This provision aims to ensure regulatory oversight integrity, such as when a company conceals recurring incidents of hallucinations in an AI system used in financial services.

  The gradation of fines is not limited to predefined categories but incorporates dynamic criteria to adapt to the specific circumstances of each case. Article 99(7) lists key factors that authorities must consider:

- *Severity of Harm to Fundamental Rights:* The quantitative and qualitative impact of the infringement is assessed, considering the number of affected individuals and the nature of the harm (e.g., health damage, restrictions on freedom, or systemic discrimination). For example, an AI hallucination causing erroneous diagnoses for hundreds of patients would justify a higher penalty than an isolated error in a low-risk system.
- *Intentional or Negligent Conduct by the Operator:* Deliberate or negligent actions in committing the infringement aggravate or mitigate the sanction. A developer ignoring technical alerts about recurring system failures would act with gross negligence, while a company implementing immediate correction protocols after detecting an error could see a reduced fine.
- *Mitigation Measures Implemented:* Post-incident corrective actions, such as compensating victims or updating the system to prevent future hallucinations, are considered mitigating factors. For example, a hospital collaborating with authorities to rectify errors in a medical AI model would demonstrate good faith, positively influencing the sanction decision.

The sanctioning system under Article 99 of the AI Act not only sets clear financial thresholds but also integrates a contextualized assessment of each infringement. This proportional and flexible structure ensures that fines are both deterrent and fair, adapting to the severity of harm, the operator's intent, and the corrective measures taken. Thus, the

European sanctioning regime strengthens its effectiveness by linking legal responsibility to ethical and technical principles, setting a global precedent in AI governance.

3. Legal and organizational strategies for integrated risk management

Hallucinations generated by artificial intelligence (AI) systems represent an overly complex challenge in legal and organizational risk management. This phenomenon transcends the technical realm, directly impacting public trust in these technologies and exposing organizations to significant legal and ethical risks, especially in sensitive sectors such as healthcare, justice, and public safety. The multifaceted nature of this problem demands an interdisciplinary approach that integrates international regulations, advanced technical frameworks, and robust organizational strategies.

3.1. Management Systems and Standards: ISO 42001

The ISO 42001 standard, published in 2023 by the International Organization for Standardization, provides a detailed framework for managing risks associated with AI, emphasizing ethical governance and transparency.[89] According to this standard, organizations must develop iterative processes to identify, mitigate, and manage risks inherent to AI systems. ISO 42001 emphasizes the need to ensure that decisions generated by AI models are explainable, auditable, and justifiable—essential conditions for minimizing the occurrence of hallucinations.

The residual risk of hallucinations in AI systems stems from the inherent complexity and limitations of machine learning models. While significant advancements in AI have led to the creation of highly sophisticated systems, these models remain heavily reliant on the data they are trained on and the algorithms that guide their behavior. Even when equipped with high-quality data and transparent algorithms, AI systems can still encounter unforeseen or ambiguous scenarios that fall outside the scope of their training data.

Additionally, the probabilistic nature of many AI models means there is always a possibility that the system will make decisions based on statistical patterns that do not fully align with reality. These factors contribute to a residual risk that cannot be eliminated but can be effectively managed.

---

[89] International Organization for Standardization (ISO), *ISO/IEC 42001:2023 - Information Technology —Artificial Intelligence— Management System* (ISO, December 2023), https://www.iso.org/standard/81230.html.

One of the pillars of the ISO 42001 standard is its structured approach to risk identification and mitigation, applicable to both technical and legal aspects. The standard acknowledges that, despite efforts to minimize risks, a residual risk will always exist due to the inherent limitations of AI systems. To address this challenge, the standard requires organizations to implement specific measures:

a) *Rigorous Data Assessments:* To ensure the quality, representativeness, and absence of biases in datasets used to train models, a critical requirement for avoiding hallucinations based on flawed information.
b) *Algorithmic Transparency:* The standard mandates that AI decision-making processes be auditable and explainable, facilitating the identification and correction of errors before they lead to legal consequences.
c) *Dynamic Monitoring:* Through periodic audits and performance metrics, the adaptability of systems in changing environments—such as legislative updates or shifts in social patterns evaluated, which is key to preventing failures in sensitive contexts like healthcare or public safety.

These controls not only meet technical requirements but also establish a robust evidentiary framework for legal disputes. For example, in cases of civil liability for damages caused by AI, the documentation required by the standards such as decision logs and data traceability—would serve as evidence to determine the organization's due diligence.

The ISO/IEC 42001:2023 standard structures its management approach using the Plan-Do-Check-Act (PDCA) cycle, an iterative model that integrates risk management into all stages of AI system development.[90] This cycle not only enhances technical processes but also fosters a proactive framework for compliance with emerging regulations, such as the European Union's Artificial Intelligence Act (AI Act). The AI Act categorizes AI systems into risk levels—unacceptable, high, limited, and minimal—while imposing stringent transparency, safety, and human oversight requirements for high-impact systems.

[90] Tobias Faiss, "Introducing ISO 42001: Setting the Standard for AI Management Systems", *Cybernavigator,* November 22, 2024, para. 3, https://www.cybernavigator.org/p/introducing-iso-42001-setting-the.

a) *Planning Phase (Plan):* In this stage, organizations must define specific policies and objectives to mitigate legal risks associated with AI hallucinations. This includes:

- Establishing clear protocols for assigning legal responsibilities in case of failures (e.g., determining whether an error is attributed to the model provider, developer, or end user).
- Identifying sector-specific regulatory requirements (such as the General Data Protection Regulation [GDPR] in the EU) and aligning them with the technical design of systems.
- Designing strategies to address algorithmic biases and ensure procedural fairness, a critical principle in areas like justice or credit scoring, where an erroneous AI decision could lead to discrimination lawsuits.

b) *Implementation Phase (Do):* This phase involves applying technical and operational controls to implement the defined policies. The standard requires:

- Bias testing and model validation, using quantifiable metrics (e.g., statistical disparity in outcomes across demographic groups) to detect and correct distortions in data or algorithms.
- Detailed documentation of AI training and decision-making processes, essential for demonstrating compliance with Article 13 of the AI Act, which requires informing users about the use of high-risk systems.
- Integration of technical explainability mechanisms to enable lawyers and regulators to understand how automated decisions are generated, facilitating legal audits.

c) *Verification Phase (Check):* This phase involves systematic reviews to evaluate the effectiveness of implemented measures. ISO/IEC 42001 highlights:

- Independent audits conducted by specialized third parties, examining both the technical quality of systems and their alignment with legal frameworks. For example, in the healthcare sector, an audit could verify whether a diagnostic model complies with the EU Medical Devices Directive (2017/745).
- Legal impact assessments, analyzing hypothetical failure scenarios (e.g., a hallucination in an automated hiring system excluding candidates based on gender) to measure the organization's legal exposure.

d) Action Phase (Act): The cycle culminates with the incorporation of improvements based on findings from the previous phase. This includes:

- Updating models and policies to adapt to legislative changes (e.g., modifications in civil liability laws for AI) or relevant case law (such as rulings setting precedents on algorithmic harm attribution).
- Continuous feedback mechanisms, such as incident reporting systems for users and employees, enabling real-time detection of failures and adjustments to controls before they escalate into legal conflicts.

The PDCA cycle acts as a bridge between the technical requirements of ISO/IEC 42001 and the legal demands of the AI Act. For example, AI systems classified as high-risk under the AI Act (e.g., those used in critical infrastructure or law enforcement) must undergo conformity assessments before commercialization. The PDCA cycle ensures that these assessments are not mere formalities but are integrated into corporate governance, where the verification phase includes compliance testing with European standards, and the action phase incorporates updates in response to new EU regulatory guidelines.

By institutionalizing continuous improvement and the transversal integration of legal risks, the PDCA cycle not only reduces the likelihood of AI hallucinations but also transforms organizations into proactive entities in a dynamic legal landscape. This approach positions ISO/IEC 42001 as an essential tool for balancing the demands of technological innovation with the need for legal security.

In this context, the standard reinforces the proactive responsibility of organizations by requiring traceability systems that document each stage of AI development and deployment. This includes:

a) Detailed logs of algorithmic decisions.

b) Protocols for incident reporting to stakeholders and authorities.

c) Mechanisms for remediation in cases of harm caused by hallucinations.

In the legal field, this documentation not only mitigates litigation risks but also sets a technical-legal precedent. For example, in cases where an AI system in the healthcare sector misdiagnoses a patient, the traceability required by the standard would allow for disaggregating the causes of the error (technical failure, data bias, or human negligence), assigning responsibilities with precision. This level of clarity goes beyond

reactive measures: it creates an evidentiary ecosystem that informs future regulations and standardizes due diligence criteria for AI systems.

In this way, ISO/IEC 42001:2023 transcends its technical nature to position itself as a bridge between technological innovation and legal compliance. By institutionalizing practices such as transparency, algorithmic openness, and impact assessments — articulated through the PDCA cycle— the standard not only reduces the incidence of AI hallucinations but also redefines corporate governance in the digital age. Its adoption becomes a strategic imperative, equipping organizations to anticipate risks in a fragmented regulatory landscape while building social legitimacy through alignment with principles such as restorative justice and the prevention of systemic harm.

3.2. Practical Risk Management: The NIST Artificial Intelligence Risk Management Framework (AI RMF)

The Artificial Intelligence Risk Management Framework (AI RMF), developed by the National Institute of Standards and Technology (NIST) in 2024, emerges as an essential tool for addressing the inherent risks of artificial intelligence (AI) systems. This framework not only complements standards such as ISO 42001 but also delves deeper into identifying technical vulnerabilities and implementing specific controls to mitigate legal impacts arising from failures in generative systems. Its iterative and multidisciplinary approach positions it as a critical tool for ensuring transparency, accountability, and regulatory compliance in high-risk AI applications.

The AI RMF structures its methodology around four core functions—Govern, Map, Measure, and Manage—to address risks such as confabulation, defined as: "the production of confidently stated but erroneous or false content [...] by which users may be misled or deceived".[91] This phenomenon, also referred to as hallucinations or fabrications, stems from the statistical nature of generative models, which predict token sequences based on training data distributions without guaranteeing factual accuracy.[92]

Below is a detailed explanation of how each function of the AI RMF mitigates this risk, supported by specific actions outlined in the framework:

---

[91] National Institute of Standards and Technology (US), "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," (Gaithersburg, MD: National Institute of Standards and Technology (U.S.), 2024): 10-8, https://doi.org/10.6028/NIST.AI.600-1.

[92] Ibid.

a) *Govern: Establishment of transparency policies and operational limits:* In this stage, the AI RMF emphasizes the need to document the origin of data and the system's capabilities. For example, action GV-1.2-001 recommends: "Establish transparency policies and processes for documenting the origin and history of training data and generated data [...] to advance digital content transparency".[93]

This policy is crucial in medical or legal contexts, where confabulation could lead to incorrect diagnoses or fabricated legal citations, exposing developers to negligence lawsuits. Additionally, the framework requires defining minimum performance thresholds ("minimum thresholds for performance or assurance criteria") prior to deployment, linking their fulfillment to implementation approvals.

b) *Map: Identification of risks in specific contexts:* The Map phase requires assessing scenarios where confabulation could cause critical harm. The document states: "Risks from confabulations may arise when users believe false content [...] such as in healthcare, where a confabulated summary of patient information reports could cause doctors to make incorrect diagnoses".[94]

To address this, action MP-1.1-003 proposes documenting risk measurement plans that include "known past GAI system incidents and failure modes," while MP-2.3-001 requires validating outputs by "comparing [outputs] to a set of known ground truth data," ensuring consistency with verifiable evidence.

c) *Measure: Implementation of explainability and validation techniques:* In this stage, the framework prioritizes quantitative and qualitative methods to detect and correct confabulations. Action MS-2.9-001 details: "Apply and document ML explanation results such as: Analysis of embeddings, Counterfactual prompts, Gradient-based attributions [...]".[95]

These techniques allow for the disaggregation of the model's internal logic, identifying patterns of false content generation. Additionally, MS-2.5-003 mandates "review and verify sources and citations in GAI system outputs during pre-deployment

[93] Ibid.
[94] Ibid.
[95] Ibid., 35.

[...] activities" (NIST 2024, 31), mitigating risks of fraudulent citations in legal or academic applications.

d) *Manage: Adaptive responses and continuous updates:* The final phase integrates mechanisms to correct failures and update controls. For example, MG-4.1-004 recommends "implementing active learning techniques to identify instances where the model fails or produces unexpected outputs", while MG-2.3-001 requires updating incident response plans to include newly encountered risks. [96] This ensures that systems evolve to address new forms of confabulation, such as adversarial prompting, where malicious users manipulate inputs to generate misinformation.

By systematically addressing risks through these four functions, the AI RMF provides a robust structure for managing the complexities and challenges associated with generative AI systems, ensuring their safe and responsible deployment.

The AI RMF aligns these functions with global legal standards. For example, the transparency requirement in GV-1.2-001 reflects Article 13 of the GDPR, which mandates explainability in automated decisions. Similarly, red-teaming tests (MP-5.1-005) align with Article 14 of the EU AI Act (2024), which calls for risk assessments in high-impact systems. By structuring an iterative governance cycle, the framework not only reduces the likelihood of harm from confabulation but also establishes a precedent for proactive due diligence, which is critical in civil liability or intellectual property litigation.

In the realm of civil liability, a model that generates incorrect medical diagnoses could expose developers and operators to lawsuits for damages, particularly if the implementation of adequate technical safeguards cannot be demonstrated. The AI RMF addresses this risk by prioritizing continuous safety evaluations (MS-2.6-006), such as adversarial testing (red-teaming) to identify dangerous responses before deployment (MP-5.1-005). Additionally, the framework requires documenting the system's "generalization limits" (MS-2.5-001), a requirement that aligns with legal principles of due diligence in the European Union under the Artificial Intelligence Act.

Hallucinations can also lead to intellectual property rights violations. For instance, if an AI model reproduces copyrighted content due to memorization of training data,

[96] Ibid., 45.

developers could face plagiarism claims. The AI RMF mitigates this risk by implementing filters to detect patented or sensitive information in outputs (MP-4.1-009) and reviewing data sources to ensure their legality (GV-6.1-004).

One of the most significant contributions of the AI RMF is its emphasis on interdisciplinary collaboration. The framework underscores that risk management teams should include not only engineers and data scientists but also ethicists, lawyers, and social science experts (MP-1.2-001). This diversity is crucial for anticipating scenarios where hallucinations might exacerbate discriminatory biases or violate fundamental rights. For example, in legal applications, a model that confabulates responses based on gender or racial stereotypes could perpetuate systemic injustices, violating principles of equality before the law. To prevent this, the framework recommends fairness assessments disaggregated by demographic groups (MS-2.11-002), aligning with jurisprudence such as that of the U.S. Supreme Court in Wisconsin v. Loomis (2016), which questioned the use of biased algorithms in judicial sentencing. [97]

Furthermore, the AI RMF introduces a risk prioritization system based on potential impact. For instance, in sectors such as finance or healthcare, where hallucinations can cause irreversible harm, the framework demands stricter controls, such as real-time human oversight (GV-3.2-003) and the creation of shutdown protocols in the event of critical failures (GV-1.7-001).

Finally, the AI RMF adopts a cyclical and iterative approach to risk management, recognizing that AI systems are constantly evolving. This involves periodically updating safety assessments (GV-1.5-002), incorporating feedback from affected users (MG-4.3-001), and adjusting technical controls in response to regulatory changes. For example, if new case law restricts the use of AI in electronic contracts, the framework would facilitate adaptation through governance reviews (GV-1.3-006). This flexibility not only strengthens operational resilience but also ensures continuous compliance with ever-evolving legal standards.

In summary, the NIST AI RMF provides a robust framework for addressing the legal impact of hallucinations in AI, integrating technical rigor, interdisciplinary diversity, and regulatory adaptability. By aligning its functions with key legal principles—such as transparency, non-discrimination, and proactive accountability—this

---

[97] United States, State v. Loomis, No. 2015AP157-CR (Supreme Court of Wisconsin, July 13, 2016), https://law.justia.com/cases/wisconsin/supreme-court/2016/2015ap000157-cr.html.

framework not only mitigates immediate risks but also lays the groundwork for the ethical and sustainable development of generative technologies in the global legal landscape.

3.3. Fundamental Rights Impact Assessment (FRIA)

The proliferation of AI systems in critical areas such as justice, healthcare, and public services has intensified the debate on how to manage inherent risks, such as hallucinations. These failures not only compromise the technical efficacy of AI but also threaten fundamental rights enshrined in international instruments, including privacy, access to accurate information, and the right to a fair trial.

The Fundamental Rights Impact Assessment (FRIA) is part of an ex-ante regulatory paradigm that prioritizes prevention over remediation. According to Mantelero, its design is inspired by pre-existing methodologies, such as Human Rights Impact Assessments (HRIAs), but adapted to the unique characteristics of AI: globalized systems with transversal applications and diffuse impacts on multiple rights.[98] Unlike generic risk assessments, the FRIA requires a specific analysis of each affected rights such as non-discrimination or freedom of expression—considering both the likelihood of harm and its potential severity.

This approach is critical in addressing hallucinations. For example, in the judicial domain, a language model (LLM) that generates false legal citations could distort decision-making processes, violating the right to a fair trial (Article 47 of the EU Charter of Fundamental Rights). The FRIA would require developers and deployers to identify such scenarios during the design phase, assessing not only the system's technical accuracy but also its interaction with specific sociotechnical contexts—such as stress in humanitarian emergency situations.

The effectiveness of the FRIA lies in its iterative structure, which combines three key stages:

a) *Planning and Scoping:* The scope of the AI system is defined, identifying potentially affected rights and vulnerable groups. For example, in a judicial chatbot, its impact on populations with limited access to legal advice would be analyzed.

---

[98] Alessandro Mantelero, "The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template", *Computer Law & Security Review* 54 (March 30, 2024): 1–18, https://doi.org/10.2139/ssrn.4782126.

b) *Risk Analysis:* The risk analysis combines two approaches: qualitative and quantitative methods to assess potential harm.

- Qualitatively, matrices evaluate the likelihood and severity of risks, focusing on factors such as the nature of the harm (e.g., misinformation in a judicial chatbot) and the vulnerability of affected groups (e.g., users with limited legal knowledge).
- Quantitatively, metrics measure exposure (e.g., number of users affected by errors) and error reversibility (e.g., time and cost to correct false information). For example, in a judicial chatbot, the probability of hallucinations could be estimated using error rates from pilot testing, while user feedback data could quantify the actual impact on decision-making. This dual approach ensures a robust and actionable risk assessment, aligning with the AI Act and GDPR requirements.

c) *Risk Management:* Technical and organizational measures are implemented, such as continuous audits or human feedback mechanisms, prioritizing solutions that eliminate risks at their source (by design). This process not only aligns with the AI Act —which requires conformity assessments for high-risk systems (Article 9)— but also complements the Data Protection Impact Assessment (DPIA) under the General Data Protection Regulation (GDPR), integrating a broader human rights perspective.

To complement the theoretical framework of FRIA, let's delve into a practical analysis using the FAIR model customization in biased ranking algorithms. The FAIR model provides a structured approach to quantify and mitigate biases in AI systems, ensuring that fairness risk controls are used as resistance strength.

Threat: Violation of the Right to Non-Discrimination

Biased ranking algorithms can violate the right to non-discrimination by unfairly ranking individuals based on protected attributes such as race, gender, or socioeconomic status. The FAIR model customization for Algorithm Impact Assessments from a physical person's perspective involves several key components: [99]

[99] Luis Enríquez Álvarez, "Personal Data Breaches: Towards a Deep Integration between Information Security Risks and GDPR Compliance Risks" (PhD thesis, Université de Lille, 2024), 451, https://theses.hal.science/tel-04723327.

1. *LEF (Likelihood of Violation):* The probable violation in a given time-frame of the right to non-discrimination by biased algorithms.
2. *LM (Likelihood of Harm):* The physical persons' probable magnitude of harm on their right to non-discrimination.
3. *TER (Threat Event Rate):* The probable frequency, within a given time-frame, that the right to non-discrimination is threatened by biased ranking algorithms.
4. *Vuln (Vulnerability):* The probability of physical persons to be discriminated against due to the biased ranking algorithm capacity and a poor implementation of fairness metrics.
5. *PL (Primary Loss):* The physical person's direct harms for being discriminated against.
6. *SL (Secondary Loss):* The physical persons' secondary harms due to a secondary stakeholder's reactions to the primary harmful event.
7. *SLEF (Secondary Likelihood of Violation):* The probability of a secondary stakeholder's reaction to the harmful event that may violate the physical persons' rights and freedoms that have been violated due to secondary stakeholders' reactions.
8. *GLM (General Loss Magnitude):* The magnitude of the physical persons' rights and freedoms that have been violated due to secondary stakeholders' reactions.

To minimize bias in ranking algorithms, fairness metrics such as demographic parity, equal opportunity, and equalized odds can be applied. These metrics are designed to ensure equitable treatment across different groups, effectively reducing the risk of discrimination:[100]

1. *RS (Fairness Metric Implementation):* The fairness metric implemented to reduce bias in ranking algorithms.
2. *TCAP (Technical Capacity):* The biased ranking algorithm's capacity to process discriminatory decisions on physical persons.

---

[100] Luís Enríquez Álvarez, "Using FAIR as a 'Swiss Army Knife' on Privacy Quantification for GDPR", *FAIR INSTITUTE,* December 3, 2024, para. 13, https://www.fairinstitute.org/blog/fair-model-privacy-uncertainty-quantification-gdpr.

3. *POA (Probability of Algorithm):* The probability that the ranking algorithms are biased.
4. *CR (Contact Rate):* The probable frequency, within a given time-frame, that data is in contact with ranking algorithms.

Another approach to analyze this risk, in the context of judicial chatbots, AI hallucinations can lead to the generation of false legal citations, distorting decision-making processes and violating the right to a fair trial. The FAIR model can be applied to assess and mitigate these risks.

1. *LEF:* The likelihood that the chatbot will generate false legal citations within a year.
2. *LM:* The potential harm to individuals who receive incorrect legal advice.
3. *TER:* The frequency at which false legal citations are generated.
4. *Vuln:* The probability that individuals with limited legal knowledge will be harmed by false information.
5. *PL:* The direct harm to individuals, such as incorrect legal decisions.
6. *SL*: Secondary harms, such as reputational damage to the judicial system or emotional distress to individuals.
7. *SLEF:* The likelihood that stakeholders (e.g., legal advocacy groups) will react to the false information, potentially leading to legal actions or public backlash.
8. *GLM:* The overall impact on the rights and freedoms of affected individuals.

The mere adoption of the FRIA is insufficient unless accompanied by an institutional culture centered on transparency and accountability. Organizations must:

1. *Establish robust internal policies:* Include protocols to detect and correct hallucinations, such as random reviews of AI-generated outcomes in judicial processes. These policies should be updated periodically, reflecting lessons learned from prior incidents.[101]

[101] Ibid.

2. *Promote interdisciplinary training:* Train technical and legal teams in AI ethics, using practical cases to illustrate how hallucinations in credit systems could perpetuate socioeconomic biases.[102]
3. *Create independent oversight committees:* As Mantelero suggests, the involvement of external experts —from academics to civil society ensures impartial evaluations and avoids conflicts of interest.[103]

Finally, transparency should not be limited to the design phase. As Kaminski argues, users affected by AI-based decisions have the right to receive understandable explanations, especially when errors like hallucinations impact their rights.[104] This principle, supported by Article 22 of the GDPR, reinforces the FRIA by providing an ex-post mechanism to audit and correct failures, closing the accountability loop.

Hallucinations are not mere technical errors but manifestations of systemic risks that demand proactive legal responses. By adopting a preventive and context-sensitive approach, the FRIA provides a framework to balance innovation with the protection of fundamental rights. However, its effectiveness will hinge on the commitment of both states and corporations to not only implement regulatory frameworks but also embrace organizational practices that prioritize human dignity over mere technological efficiency. As Mantelero warns, without rigorous implementation—backed by effective sanctions and citizen participation, even the most sophisticated tools will remain dead letter.

3.4. Responsibility by Design of AI

As AI systems increasingly permeate domains where fundamental rights are at stake—such as justice, healthcare, education, and public security, the phenomenon of AI hallucinations, understood as factually incorrect or fabricated outputs, poses not only epistemic risks but also direct legal consequences. Moving beyond reactive or remedial strategies, the principle of Responsibility by Design (also known as "AI by Design,") emerges as an essential governance paradigm. This approach insists on embedding legal, ethical, and accountability safeguards at every stage of the AI lifecycle—beginning at the ideation and system architecture levels—rather than relegating such measures to post-deployment correction or damage control.

---

[102] Ibid.
[103] Ibid.
[104] Margot E. Kaminski, "The Right to Explanation, Explained", *Berkeley Technology Law Journal* 34, no. 1 (November 26, 2019): 190–218, https://doi.org/10.15779/Z38TD9N83H.

Within the context of hallucinations, which often result from probabilistic extrapolations based on incomplete or biased training data, Responsibility by Design requires that developers explicitly address these risks from the outset. This includes implementing mechanisms for output traceability, adversarial robustness testing, and ethical scenario simulation. For example, generative language models intended for use in legal, journalistic, or academic settings must be equipped with verifiability constraints—such as fact-checking subsystems or citation validation—to prevent the automated dissemination of fabricated legal cases or false accusations that may infringe rights to honor, due process, or professional reputation.

Responsibility by Design is not a mere aspirational principle, it is increasingly codified in binding instruments. Under the EU Artificial Intelligence Act, Articles 9 and 14 mandate comprehensive risk management procedures and demand that high-risk AI systems be designed with transparency, verifiability, and robust human oversight mechanisms. This requirement aligns closely with standards such as ISO/IEC 42001, which operationalizes AI management systems through formalized risk assessment cycles and compliance auditing. Complementary frameworks like the NIST AI Risk Management Framework reinforce this proactive approach, emphasizing governance functions that are anticipatory rather than reactive. Moreover, instruments like the Fundamental Rights Impact Assessment (FRIA), as proposed in EU digital governance practices, further ground Responsibility by Design in ex ante fundamental rights protection, ensuring that AI models are subjected to rigorous rights-based evaluations prior to deployment.

Crucially, Responsibility by Design also functions as a liability mitigation strategy. In legal contexts where hallucinations may cause material or reputational harm—as illustrated in the Walters v. OpenAI case already referred to in section 2.2 of chapter one of this thesis, developers and deployers may invoke demonstrable adherence as part of their due diligence defense. If it can be shown that exhaustive testing, ethical auditing, and oversight mechanisms were integrated throughout development, this may reduce exposure to negligence claims or regulatory sanctions. Thus, embedding ethical safeguards and traceability into the system's design is not only a compliance measure but a risk-avoidance mechanism grounded in legal rationality.

Responsibility by Design further enhances AI trustworthiness, a concept that transcends technical safety and enters the domain of democratic legitimacy. In the absence of verifiable safeguards, hallucinations risk eroding public confidence in AI

applications, particularly when such outputs affect judicial outcomes, medical diagnoses, or reputational assessments. By contrast, systems built on transparent, rights-aligned architectures can reinforce legitimacy and help prevent what the thesis earlier identified as "automated violations" of fundamental rights.

This must be understood as an ethical-technical-legal imperative for AI governance. It provides a blueprint for integrating constitutional and human rights standards into technological infrastructure, thereby operationalizing the principle that legal compliance and ethical integrity are not external constraints on innovation but core design features of trustworthy and lawful AI. Within the specific context of hallucinations, this approach offers not only a preventive strategy but a normative standard for AI development that upholds human dignity and legal certainty.

4. Technical measures for detection and mitigation of hallucinations

AI hallucinations represent a critical challenge from both a technical and regulatory perspective. These hallucinations, stemming from biases in training data, algorithmic limitations, or malicious attacks, pose significant risks in areas such as privacy, security, and fundamental rights. The need to implement advanced technical measures, including predictive methods, becomes indispensable to detect and mitigate these phenomena, reducing associated risks and ensuring regulatory compliance.

The complexity of AI hallucinations lies in their multifaceted nature, which goes beyond mere technical errors to reflect the interaction of systemic and contextual factors. In Noise: A Flaw in Human Judgment, Daniel Kahneman, along with Sibony and Sunstein, explores how unwanted variability in human judgments-referred to as “noise” compromises accuracy and fairness in systems such as justice, medicine, or insurance.[105] Although the book does not directly address AI, its analysis of noise as systemic inconsistency provides a relevant theoretical framework for understanding hallucinations: these could be interpreted as a manifestation of “algorithmic noise,” where unwanted variability in automated decisions arises from hidden biases, ambiguous data, or unpredictable interactions within the model. These parallel underscores the need for proactive approaches to manage both human noise and AI technical dysfunctions.

---

[105] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (London: William Collins, 2021).

Following Kahneman's logic on reducing "noise" in human judgments—through structured protocols, systematic audits, and simplified algorithms, the mitigation of AI hallucinations demands analogous measures: standardization in model training (e.g., balanced datasets and uniform evaluation criteria), continuous monitoring of outputs using anomaly detection tools, and rigorous controls based on ethical and legal principles. This interdisciplinary approach, which integrates technical advancements (such as fine-tuning with human feedback) with clear regulatory frameworks, not only reduces unwanted variability but also aligns AI systems with legally enforceable standards of transparency and accountability.

In addition to reactive measures, predictive methods emerge as key tools to anticipate and neutralize hallucinations. Techniques such as probabilistic uncertainty analysis (to identify responses with low statistical confidence), the use of adversarial verification models (which detect inconsistencies before system deployment), and the implementation of real-time context sensors (which adjust outputs based on the interpreted environment) allow the problem to be addressed at its root. These strategies, inspired by Kahneman's notion of "decision hygiene" which prioritizes prevention over correction—reinforce the need to develop AI that is not only intelligent but also predictable and auditable, where the anticipation of technical risks is combined with legal guarantees of safety and equity.

4.1. Predictive Methods in the Detection and Mitigation of Hallucinations

Proactive management of hallucinations in AI systems cannot be limited to ex post facto corrections; it requires predictive mechanisms that anticipate and neutralize "algorithmic noise" before it materializes into legal or social harm. Inspired by Kahneman's analogy of "decision hygiene" which prioritizes structuring processes to minimize unwanted variability—predictive methods emerge as key tools to transform the inherent uncertainty of models into quantifiable and controllable risks. This approach not only addresses technical failures but also responds to the legal imperatives of transparency, security, and proactive accountability required within the framework of AI risk management.

The methods to be analyzed in this section —ranging from rigorous statistical frameworks to iterative consensus techniques— share a common goal: reducing the gap between algorithmic confidence and real-world reliability. Each operates at a different layer of the problem, forming a mitigation ecosystem that transcends the technical: they serve as bridges between model engineering and the legal principles of prevention, precaution, and transparency. Their adoption not only minimizes hallucinations but also builds a verifiable framework to assign accountability —whether personal or institutional— when systems fail.

4.1.1. Conformal Prediction

Conformal prediction is a statistical technique that quantifies uncertainty in the predictions of artificial intelligence (AI) models, providing confidence intervals that help identify potentially hallucinated responses. This methodology generates predictions accompanied by a predefined confidence level, making it a robust tool for detecting inconsistencies in model outputs. According to Angelopoulos and Bates,[106] conformal prediction is a versatile and easy-to-implement technique applicable to a wide range of problems in fields such as natural language processing and deep learning, where uncertainty and hallucinations are common challenges. Additionally, its ability to generate reliable prediction sets makes it particularly useful in environments where reliability is critical.

---

[106] Anastasios N. Angelopoulos and Stephen Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification", *arXiv* (2022): 4, https://doi.org/10.48550/arXiv.2107.07511.

Conformal prediction offers a structured way to quantify the uncertainty of AI-generated outputs, helping distinguish between reliable predictions and potential hallucinations. By assessing whether a model's output falls within a predefined confidence interval, stakeholders can determine the extent to which they should trust a given prediction.

AI hallucinations are particularly problematic in fields that rely on precise and factual data, such as law, healthcare, and finance. In legal settings, for instance, AI-generated hallucinations can result in fabricated case citations or misleading legal arguments, potentially affecting judicial decisions. In healthcare, an AI model that hallucinates symptoms or treatment recommendations could lead to incorrect diagnoses, endangering patient safety. Similarly, in finance, hallucinated predictions about market trends or credit risk assessments could result in substantial financial losses. Because these hallucinations often appear plausible, they are difficult to detect without robust uncertainty quantification methods such as conformal prediction.

One of the most concerning aspects of AI hallucinations is their persistence even when confidence intervals suggest high certainty. This means that an AI model may generate an incorrect response while still placing it within a seemingly reliable prediction set. Conformal prediction mitigates this risk by allowing stakeholders to adjust confidence thresholds dynamically and integrate additional validation layers, such as external auditing or cross-referencing against verified datasets. However, confidence intervals alone are insufficient to guarantee the correctness of a prediction, making the combination of statistical validation and domain expertise essential.

In high-impact applications, such as medical diagnosis, conformal prediction has proven to be an invaluable tool. This technique allows for the evaluation of the accuracy of AI systems in sensitive tasks, identifying diagnoses that fall outside expected statistical boundaries and thereby reducing associated risks. This anomaly detection capability is also relevant in other areas, such as cybersecurity and finance, where it can prevent systemic failures and protect user interests.

Furthermore, conformal prediction is particularly useful for real-time monitoring of critical systems, as it enables the early identification of unusual or unexpected outputs. As Angelopoulos and Bates note, this technique can be adapted to detect outliers, which is crucial for identifying hallucinated responses in AI models.[107] This approach not only

[107] Ibid., 19.

reinforces the operational safety of systems but also adds an additional layer of confidence in high-uncertainty scenarios.

One of the fundamental principles of conformal prediction is its validity under the assumption of exchangeability, which ensures that the confidence intervals maintain their statistical reliability across multiple predictions.[108] This means that, as long as the input data follows the same distribution, the method guarantees that the predicted interval contains the true value with a specified probability. Additionally, the errors in conformal prediction are probabilistically independent, which allows for a direct interpretation of the confidence levels.[109]

Despite its advantages, conformal prediction has limitations that must be considered in a legal context. First, its effectiveness depends on the quality and representativeness of the calibration data. If these data contain historical biases, for example, discrimination in bank loans, the confidence intervals will reflect and perpetuate those inequities. This poses a significant legal challenge, as AI systems must comply with principles of fairness and non-discrimination, as required by regulations such as the General Data Protection Regulation (GDPR) in the European Union.

While ensemble and Bayesian techniques are powerful methods for uncertainty quantification, conformal prediction often outperforms them in specific scenarios due to its simplicity, flexibility, and robustness. Conformal prediction does not rely on strong assumptions about the underlying data distribution, unlike Bayesian methods, which typically require specifying a priority and assume that the model is well-calibrated. This makes conformal prediction more adaptable to real-world datasets, which are often messy and non-conforming to idealized statistical assumptions.

Moreover, conformal prediction differs from Bayesian approaches in that it does not require large datasets to achieve reliable confidence intervals. Unlike Bayesian methods, which depend on having a well-defined priority, conformal prediction maintains its reliability even when data availability is limited, making it a valuable tool in regulatory and legal applications.[110] Additionally, while ensemble techniques such as bagging and boosting can enhance model reliability, they are often computationally expensive.

---

[108] Glenn Shafer and Vladimir Vovk, "A Tutorial on Conformal Prediction", *Journal of Machine Learning Research* 9 (2008): 372, https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf.

[109] Ibid., 375.

[110] Ibid., 376.

Conformal prediction, in contrast, provides a straightforward framework for generating prediction sets with guaranteed coverage, regardless of the base model used.[111]

Second, a prediction may fall within the confidence interval and still be a hallucination. For instance, a language model that generates false but statistically "reliable" legal citations could go unnoticed if reliance is placed solely on confidence intervals. This underscores the need to combine conformal prediction with complementary approaches, such as explainability (XAI), which will be discussed in this section, or external audits, to ensure that predictions are not only statistically reliable but also semantically correct.

Furthermore, the application of conformal prediction in high-stakes decisions, such as finance or healthcare, must account for the fact that confidence intervals reflect statistical uncertainty but do not inherently validate the correctness of a prediction. This limitation emphasizes the need for external validation mechanisms, including domain-specific heuristics and human oversight.[112]

Conformal prediction has profound implications in the legal and regulatory realm. By providing objective metrics of uncertainty, this method enables regulators and lawyers to assess whether an AI developer implemented "sufficient technical measures" to mitigate harm, as required by the European AI Act (Article 9). For example, in the context of automated decisions affecting individual rights—such as loan denials or candidate selection for employment, the confidence intervals generated by conformal prediction can be used as evidence that the system operates within acceptable risk limits.

Moreover, in cases of liability for damages, confidence intervals can serve as a clear criterion for determining whether a developer acted with due diligence. If an AI system generates a prediction with an extremely wide confidence interval —indicating high uncertainty— and this prediction results in harm, the developer could be deemed negligent for failing to implement mechanisms to identify and mitigate this risk.

Conformal prediction does not eliminate AI hallucinations, but it transforms them into a manageable risk through objective metrics. By quantifying uncertainty, this method provides lawyers and regulators with a clear criterion to evaluate whether a developer implemented "sufficient technical measures" to mitigate harm, as required by regulations such as the GDPR and the European AI Act. This bridge between statistics and law is

---

[111] Ibid., 381.
[112] Ibid., 382.

essential: it turns an abstract technical phenomenon—hallucinations—into an actionable legal parameter—enforceable confidence levels.

While conformal prediction provides a mathematically sound approach to quantifying uncertainty, it should not be relied upon in isolation. The combination of statistical validation, legal oversight, and domain expertise is essential to ensure that AI-generated predictions are not only statistically plausible but also ethically and legally justifiable.

4.1.2. The Delphi Method

The Delphi method, originally conceived by the RAND Corporation during the Cold War to address strategic problems under uncertainty, has emerged as a key mechanism for optimizing artificial intelligence (AI) systems in domains where precision and transparency are imperative.[113] This approach, based on the structured iteration of perspectives among multiple experts or algorithmic models, not only mitigates biases and "hallucinations" but also bridges the gap between technological innovation and legal and ethical frameworks. Its application in regulatory and public policy contexts reflects a necessary convergence between technical agility and social responsibility.

In the legal realm, the integration of the Delphi method becomes relevant considering regulatory requirements such as the European Union's General Data Protection Regulation (GDPR). Article 22 of the GDPR stipulates that automated decisions must be "subject to human review," implying the need for traceability and procedural justification. Here, the Delphi method operates as a collaborative filter: by iterating predictions among independent models or consulting interdisciplinary panels (legal experts, engineers, ethicists), it generates layered explanations that break down the logic behind each decision. This process not only satisfies the "right to explanation" but also mitigates litigation risks by documenting how conflicting norms or ethical principles were weighed during algorithmic design.

It is worth noting that the effectiveness of the method depends on its institutional configuration. Poorly structured participatory processes can reproduce power asymmetries. Therefore, in projects such as the regulation of algorithms in public services, it is essential to establish clear protocols for expert selection, opinion weighting,

---

[113] Harold A. Linstone and Murray Turoff, *The Delphi Method: Techniques and Applications* (United States: Addison-Wesley Publishing Company, Advanced Book Program, 2002), https://www.foresight.pl/assets/downloads/publications/Turoff_Linstone.pdf.

and external validation of results. In this regard, Vestri proposes that public organizations should "create a permanent position that could be called the Algorithmic Protection Officer (APO) [...] responsible for monitoring the use of algorithmic and AI tools, reminding the entity's leaders of the importance of complying with current regulations".[114] This role, inspired by the Data Protection Officer model, would ensure continuous technical and ethical oversight, ensuring that diverse perspectives translate into robust and legitimized solutions.

Technically, the Delphi method can be implemented through collaborative AI architectures, where multiple models —each specialized in a domain or trained on heterogeneous data— exchange predictions and critiques in iterative cycles. For example, in medical diagnostic systems, one model analyzes radiological images, another reviews clinical histories, and a third evaluates treatment protocols; through cross-feedback, a consensus diagnosis is reached that minimizes false positives. This mechanism is particularly useful in areas where the cost of error is high, such as justice or healthcare.

However, its implementation faces computational challenges. The need to synchronize multiple agents and process large volumes of feedback can increase system latency. To address this, techniques such as federated learning or distributed consensus optimization allow the method to scale without sacrificing efficiency. Thus, the balance between precision and operability is maintained.

The Delphi method represents an innovative synthesis of technical rigor and social responsibility in AI development. By structuring interdisciplinary collaboration and critical iteration, it not only enhances algorithmic accuracy but also ensures that automated systems operate within acceptable ethical and legal boundaries. Its adoption in sensitive sectors, such as law and public policy, is not an option but an imperative for societies aspiring to inclusive and legitimate technological governance.

4.2.3. Stress Testing and Adversarial Testing

Stress testing and adversarial testing represent a foundational methodological pillar for ensuring the reliability of artificial intelligence (AI) systems, particularly in addressing the phenomenon of hallucinations—understood as the generation of erroneous or factually unfounded content. These techniques not only evaluate the technical

---

[114] Gabriele Vestri, "The Transformative Fusion Between the Public Sector and Artificial Intelligence (AI): The 'Impact Assessment Test' as a Priority", *International Journal of Digital Law – IJDL* 4, no. 3 (2024): 17, https://doi.org/10.47975/digital.law.vol.4.n.3.

robustness of models but also establish a framework of proactive accountability for legal and ethical risks. In this sense, their systematic implementation emerges as a mechanism of algorithmic governance, aligned with the principles of transparency and accountability required by emerging regulatory frameworks, such as the European Union's General Data Protection Regulation (GDPR).

From a technical perspective, AI hallucinations often arise when models encounter operational environments not anticipated during training, revealing intrinsic limitations in their generalization capabilities. Deep learning-based medical diagnostic systems are susceptible to adversarial attacks, where minimal perturbations in radiological images —imperceptible to the human eye— can induce catastrophic classification errors. This phenomenon is not limited to the biomedical field: in language models like GPT-4, subtle changes in input (e.g., semantic ambiguities or contextual biases) can trigger contradictory or fabricated responses.

To address these vulnerabilities, adversarial testing employs strategies such as generating adversarial examples, designed to explore the limits of a model's inferential capacity. Additionally, stress testing subjects' systems to extreme operational loads (e.g., massive data volumes or highly ambiguous contexts), revealing patterns of predictive degradation that could lead to hallucinations.

The integration of these tests into AI development cycles gains legal relevance by linking them to obligations of algorithmic due diligence. As Enríquez Álvarez notes, hybrid risk assessment methods, which combine quantitative and qualitative approaches, include stress testing as "a wide range of techniques, starting with substituting a simple number by a worser one ending in a full stochastic simulation environment." These approaches aim to: "capture and synthesize diverse opinions and concerns, to better handle hard-to-predict risks, discover vulnerabilities of the organization, and improve the transparency of inefficient activities and make them visible to the management body".[115]

In regulated sectors such as finance, stress testing has emerged as an alternative to Value at Risk (VaR), albeit with the limitation of depending "on the judgment and experience of the people applying it".[116] Automated auditing mechanisms, complemented by these tests, enable the detection of anomalies associated with hallucinations before

[115] Luis Enríquez Álvarez, "Personal Data Breaches: Towards a Deep Integration between Information Security Risks and GDPR Compliance Risks" (PhD thesis, Université de Lille, 2024), 276, https://theses.hal.science/tel-04723327.

[116] Ibid.

they escalate into violations of fundamental rights. For example, in data protection, an AI model generating false inferences about personal information could violate the GDPR's accuracy principle (Art. 5.1.d), exposing organizations to fines of up to 4% of global revenue.

An illustrative case is the use of real-time drift detection, a technique that monitors the statistical consistency between training data and operational inputs. When a critical discrepancy is identified —such as unanticipated biases— the system activates contingency protocols, ranging from automated notifications to the temporary disabling of the model. This oversight reinforces the principle of privacy by design and mitigates legal risks arising from decisions based on hallucinations.

The legal ramifications of AI hallucinations extend beyond GDPR compliance and data protection concerns. In sectors such as healthcare, misdiagnoses resulting from AI-generated errors could lead to malpractice claims, raising liability issues for both developers and end-users. Similarly, in financial services, incorrect risk assessments produced by AI-driven decision-making could violate consumer protection laws, exposing financial institutions to regulatory sanctions. Moreover, in criminal justice, AI-based profiling errors could result in wrongful arrests or discriminatory outcomes, potentially breaching fundamental rights under international human rights frameworks.

Additionally, AI hallucinations pose significant risks in contractual and corporate liability contexts. Businesses relying on AI for automated decision-making —such as hiring, credit approval, or legal analysis— may face lawsuits if incorrect outputs lead to wrongful terminations, financial losses, or contractual breaches. This underscores the necessity of integrating stress testing and adversarial testing into AI governance frameworks, ensuring that organizations can anticipate and mitigate legal exposure before AI errors result in tangible harm.

The need for robust legal frameworks is further highlighted by emerging AI regulations, such as the EU AI Act, which categorizes high-risk AI applications and mandates stricter compliance requirements, including transparency obligations, risk assessments, and human oversight. AI hallucinations, if left unaddressed, could violate these mandates, leading to fines, operational restrictions, or even bans on AI deployment in critical sectors. Therefore, legal accountability must be embedded into the AI lifecycle, aligning with stress testing and adversarial testing methodologies to prevent regulatory breaches and safeguard individual rights.

The use of stress testing and adversarial testing as tools for assessing the robustness of AI systems aligns with established methodologies in operational risk management, as detailed in actuarial literature. In this regard, the Actuarial Association of Europe (AAE) emphasizes that these tests allow for capturing and synthesizing diverse opinions and concerns, particularly in operational risks where purely quantitative methods may be insufficient.[117]

From a methodological perspective, stress testing in AI shares characteristics with its application in sectors such as finance and insurance. The AAE highlights that these tests can identify organizational vulnerabilities and improve the transparency of inefficient activities, key elements for regulatory compliance and risk management in AI models.[118] This reinforces the need for algorithmic governance that integrates both technical oversight and legal regulation, in line with principles such as privacy by design.

Furthermore, the hybrid risk assessment approaches described in the original text find a clear parallel to actuarial literature. The AAE underscores that stress testing can address hard-to-predict risks, known as black swans, and facilitate decision-making through alternative scenarios.[119] This coincides with the importance of designing tests that reflect both technical and regulatory challenges, ensuring that AI systems are not only technically robust but also transparent and auditable.

The application of stress testing and adversarial testing in AI, as in operational risk management, heavily depends on the quality of expert judgment used to select the tests and analyze the results. In this context, techniques such as real-time drift detection and robustness certification emerge as additional mechanisms to ensure model reliability in production environments.

The effectiveness of adversarial testing depends on its integration into a holistic governance framework. First, it is essential to adopt detailed documentation standards (e.g., model cards, datasheets) that record overall performance and behavior in extreme scenarios. Second, as Zhou suggests, diversifying training data, including synthetic adversarial examples—reduces the attack surface. Third, interdisciplinary collaboration

---

[117] Malcolm Kemp, Christoph Krischanitz, and Daphné De Leval, "Actuaries and Operatonal Risk Management", *Actuarial Association of Europe (AAE)*, January 2021, 34, https://actuary.eu/wp-content/uploads/2021/01/Actuaries-and-Operational-Risk-Management-FINAL.pdf.

[118] Ibid.

[119] Ibid.

among engineers, legal experts, and ethicists enables the design of tests that reflect both technical and regulatory challenges.[120]

A key advancement is the development of robustness certification frameworks, such as the one proposed by NIST, which establishes standardized metrics for evaluating resistance to adversarial attacks. These initiatives lay the groundwork for evidence-based regulation, where stress testing becomes a binding requirement for AI in high-risk contexts.

Stress testing and adversarial testing transcend their technical function to become instruments of accountability in the algorithmic age. By systematically exposing vulnerabilities, they prevent hallucinations and build social trust in technologies whose opacity challenges the principles of the rule of law. Their rigorous implementation, supported by adaptive legal frameworks, is an ethical and legal imperative to ensure that innovation does not eclipse individual rights.

[120] Qianwei Zhou, Margarita Zuley, Yuan Guo, Lu Yang, Bronwyn Nair, Adrienne Vargo, Suzanne Ghannam, Dooman Arefan, and Shandong Wu, "A Machine and Human Reader Study on AI Diagnosis Model Safety under Attacks of Adversarial Images", *Nature Communications* 12, n.°. 1 (2021): 2, https://doi.org/10.1038/s41467-021-27577-x.

Conclusions and Recommendations

Conclusions

After analyzing the phenomenon of AI hallucinations, the following conclusions can be drawn:

1. AI hallucinations pose a significant risk, as they generate unexpected and inaccurate outputs that can affect decision-making in critical contexts such as healthcare and justice.
2. A robust regulatory framework, such as the European Union's AI Act, is essential to ensure the accountability of AI developers and operators while protecting user rights.
3. AI developers must adhere to strict obligations, including the use of high-quality data and rigorous testing, to minimize the likelihood of hallucinations in their systems.
4. Transparency in AI model functioning is crucial. Users and affected parties must be able to understand how and why certain decisions are made, necessitating the development of explainability tools.
5. Fundamental Rights Impact Assessments (FRIAs) are a vital tool for identifying and mitigating risks associated with hallucinations, ensuring that emerging technologies do not compromise basic rights.
6. Integrating diverse disciplines into the design and development of AI systems will enrich the understanding of their implications and promote an ethical approach to their implementation.
7. The rapid evolution of AI technology requires constant adaptation of the legal framework to efficiently address new challenges and protect individual rights.
8. Raising public awareness about AI hallucinations and their potential effects, as well as promoting education on the responsible use of these technologies, is essential.
9. Given the global nature of AI, fostering international collaboration in developing standards and regulations is necessary to effectively address the challenges posed by hallucinations in a global legal context.

10. Advanced mitigation techniques, such as Conformal Prediction, Adversarial Testing, and Stress Testing, should be integrated into AI model validation to enhance reliability and robustness.

Recommendations

1. Develop and implement specific protocols for detecting, evaluating, and mitigating AI hallucinations, ensuring proactive risk management.
2. Regulatory authorities should establish clear legal responsibilities for AI developers and operators in cases of hallucinations, ensuring effective accountability mechanisms.
3. Implement training programs for AI developers, users, and public officials on the ethical and legal risks of AI hallucinations, promoting responsible decision-making.
4. Conduct regular audits of AI systems, complemented by real-time monitoring tools, to assess performance and promptly identify anomalies.
5. Promote transparency in AI decision-making by ensuring that evaluation criteria, training data sources, and decision-making processes are accessible and understandable.
6. Foster collaboration between public, private, and academic sectors to develop best practices and design more effective mitigation strategies for AI hallucinations.
7. Encourage the integration of advanced validation techniques, such as Conformal Prediction, Adversarial Testing, and Stress Testing, into AI development to improve accuracy and reliability.
8. Strengthen the implementation of Fundamental Rights Impact Assessments (FRIAs) before deploying AI systems in high-risk areas, ensuring that AI-driven decisions do not infringe on human rights.
9. Promote international coordination in the establishment of AI regulations, recognizing the global nature of AI-related risks and fostering legal harmonization across jurisdictions.
10. Enhance public engagement by launching awareness campaigns and educational initiatives on AI risks and responsible usage, ensuring that society is well-informed about the implications of AI hallucinations.

Bibliography

119 Congress of the United States. Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (TAKE IT DOWN ACT), S.146 TAKE IT DOWN Act § (2025). https://www.congress.gov/bill/119th-congress/senate-bill/146/text.

Ada Lovelace Institute. "Examining the Black Box." *Ada Lovelace Institute*, 2020. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

Angelopoulos, Anastasios N, and Stephen Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." *Cornell University* 6 (December 7, 2022): 1–51. doi:10.48550/arXiv.2107.07511.

Ankit. "What Are AI Hallucinations? The Complete Guide." *GeeksforGeeks*, January 24, 2025. https://www.geeksforgeeks.org/what-is-ai-hallucination/.

AWS. "What is LLM (Large Language Model)?" *AWS*. Accessed February 11, 2025. https://aws.amazon.com/es/what-is/large-language-model/.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (June 1, 2016). doi:10.1177/2053951715622512.

Cath, Corinne. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (October 15, 2018): 20180080. doi:10.1098/rsta.2018.0080.

Council of the European Union. "Artificial Intelligence Act." *https://artificialintelligenceact.eu/wp-content/uploads/2024/02/AIA-Trilogue-Coreper.pdf*, January 26, 2024. https://artificialintelligenceact.eu/wp-content/uploads/2024/02/AIA-Trilogue-Coreper.pdf.

Echeverría Muñoz, Darío. "Derecho al honor e Inteligencia Artificial." *Revista Digital EDI No. 35*, August 31, 2020. https://issuu.com/elderechoinformatico.com/docs/revista_35.

———. "The Right to Honor, Image and Good Reputation: Historical Background and Constitutional Regulation in Ecuador." *Ius Humani. Law Journal* 9, no. 1 (February 13, 2020): 209–30. doi:10.31207/ih.v9i1.228.

Elfman, Liz. “What Are AI Hallucinations? Examples & Mitigation Techniques.” *Data.World*, September 10, 2024. https://data.world/blog/ai-hallucination/.

Enríquez Álvarez, Luis. “Personal Data Breaches: Towards a Deep Integration between Information Security Risks and GDPR Compliance Risks.” Phdthesis, Université de Lille, 2024. https://theses.hal.science/tel-04723327.

Enríquez Álvarez, Luís. “Using FAIR as a ‘Swiss Army Knife’ on Privacy Quantification for GDPR.” *FAIR INSTITUTE*, December 3, 2024. https://www.fairinstitute.org/blog/fair-model-privacy-uncertainty-quantification-gdpr.

Escolano Ruíz, Francisco, Patricia Compañ Rosique, Ramón Rizo Aldeguer, and Miguel Ángel Cazorla Quevedo. *Fundamentos de inteligencia artificial*. Alicante: Publicaciones de la Universidad de Alicante, 1999. https://www.digitaliapublishing.com/a/661/fundamentos-de-inteligencia-artificial.

European Commission. “Approval of the Content of the Draft Communication from the Commission - Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act),” February 4, 2025. https://ec.europa.eu/newsroom/dae/redirection/document/112367.

———. “Approval of the Content of the Draft Communication from the Commission - Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act),” February 4, 2025. https://ec.europa.eu/newsroom/dae/redirection/document/112366.

———. “Approval of the Content of the Draft Communication from the Commission - Commission Guidelines on the Definition of an Artificial Intelligence System Established by Regulation (EU) 2024/1689 (AI Act),” February 6, 2025. https://ec.europa.eu/newsroom/dae/redirection/document/112455.

European Parliament. “EU AI Act: First Regulation on Artificial Intelligence.” *European Parliament*, June 8, 2023. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, Regulation (EU) 2016/679 General Data Protection

Regulation § (2016). https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:32016R0679.

———. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, Regulation (EU) 2024/1689 Artificial Intelligence Act § (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Faiss, Tobias. "Introducing ISO 42001: Setting the Standard for AI Management Systems." *Cyber Navigator*, November 22, 2024. https://www.cybernavigator.org/p/introducing-iso-42001-setting-the.

GeeksforGeeks. "What Is a Large Language Model (LLM)." *GeeksforGeeks*, January 22, 2025. https://www.geeksforgeeks.org/large-language-model-llm/.

IBM. "What Are AI Hallucinations?" *IBM*, September 1, 2023. https://www.ibm.com/think/topics/ai-hallucinations.

———. "What Are Large Language Models (LLMs)?" *IBM*, November 2, 2023. https://www.ibm.com/think/topics/large-language-models.

International, Amnesty. "Report: Israeli Authorities Are Using Facial Recognition Technology to Entrench Apartheid." *Amnesty International Australia*, May 1, 2023. https://www.amnesty.org.au/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/.

International Organization for Standardization (ISO). "ISO/IEC 42001:2023 - Information Technology — Artificial Intelligence — Management System." *ISO*, December 2023. https://www.iso.org/standard/81230.html.

Jorqui Azofra, María. *Responsabilidad por los daños causados por productos y sistemas de inteligencia artificial*, 2023. https://www.digitaliapublishing.com/a/131264/responsabilidad-por-los-danos-causados-por-productos-y-sistemas-de-inteligencia-artificial.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. *Noise: A Flaw in Human Judgment*. London: William Collins, 2021.

Kemp, Malcolm, Christoph Krischanitz, and Daphné De Leval. "Actuaries and Operatonal Risk Management." Actuarial Association of Europe (AAE), January

2021. https://actuary.eu/wp-content/uploads/2021/01/Actuaries-and-Operational-Risk-Management-FINAL.pdf.

Kempf, Anna-Lena, and Nils Rauer. “A Guide to High-Risk AI Systems under the EU AI Act.” *Pinsent Masons*, February 13, 2024. https://www.pinsentmasons.com/out-law/guides/guide-to-high-risk-ai-systems-under-the-eu-ai-act.

Koetsier, Teun. “A Note on Adrienne Mayor’s Gods and Robots.” In *Advances in Mechanism and Machine Science*, 73:1187–96. Cham: Springer International Publishing, 2019. doi:10.1007/978-3-030-20131-9_117.

Krisher, Tom. “EEUU investiga sistema de conducción autónoma de Tesla tras muerte de peatón arrollado.” *Los Angeles Times en Español*, October 18, 2024. https://www.latimes.com/espanol/eeuu/articulo/2024-10-18/eeuu-investiga-sistema-de-conduccion-autonoma-de-tesla-tras-muerte-de-peaton-arrollado.

Kuner, Christopher. “The Schrems II Judgment of the Court of Justice and the Future of Data Transfer Regulation.” *European Law Blog*, July 17, 2020. doi:10.21428/9885764c.aed20daf.

Larson, Jeff, Julia Angwin, Lauren Kirchner, and Surya Mattu. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*, May 23, 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lateef, Zulaikha. “Types of AI: Understanding Different Types of Artificial Intelligence in 2024.” *Edureka*, June 18, 2019. https://www.edureka.co/blog/types-of-artificial-intelligence/.

Linstone, Harold A., and Murray Turoff. *The Delphi Method: Techniques and Applications*. United States: Addison-Wesley Publishing Company, Advanced Book Program, 2002. https://www.foresight.pl/assets/downloads/publications/Turoff_Linstone.pdf.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. “Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.” In *Dartmouth Conference*. Hanover, New Hampshire, 1955. doi:https://doi.org/10.1609/aimag.v27i4.1904.

Misuraca, Gianluca, and Colin Van Noordt. “AI Watch Artificial Intelligence in Public Services: Overview of the Use and Impact of AI in Public Services in the EU.” *Publications Office of the European Union*, July 1, 2020, 1–96. doi:10.2760/039619.

Murphy, Maria Helen. “Assessing the Implications of Schrems II for EU–US Data Flow.” *International & Comparative Law Quarterly* 71, no. 1 (January 2022): 245–62. doi:10.1017/S0020589321000348.

National Institute of Standards and Technology (US). “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.” Gaithersburg, MD: National Institute of Standards and Technology (U.S.), July 26, 2024. doi:10.6028/NIST.AI.600-1.

Naughtin, Claire, and Sarah Vivienne Bentley. “Both Humans and AI Hallucinate — but Not in the Same Way.” *The Conversation*, June 16, 2023. http://theconversation.com/both-humans-and-ai-hallucinate-but-not-in-the-same-way-205754.

OECD. “Recommendation of the Council on Artificial Intelligence.” *OECD Legal Instruments*, 2025. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Parliament of Canada. “Government Bill (House of Commons) C-27 (44-1) - First Reading - Digital Charter Implementation Act, 2022 - Parliament of Canada.” *Parliament of Canada*, June 16, 2022. https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading.

Perera, Avi. “AI Ethics Case Studies: Lessons Learned from Real-World Failures – Avi Perera.” *Aviperera*, October 30, 2024. https://aviperera.com/ai-ethics-case-studies-lessons-learned-from-real-world-failures/.

Prego, Carlos. “Un Abogado Usó ChatGPT En Un Juicio. Ahora Es Él Quien Debe Dar Explicaciones a Un Juez Por Incluir Citas Falsas.” *Xataka*, May 29, 2023. https://www.xataka.com/legislacion-y-derechos/abogado-uso-chatgpt-juicio-ahora-quien-debe-dar-explicaciones-a-juez-incluir-citas-falsas.

Rebollo Delgado, Lucrecio. *Inteligencia artificial y Derechos fundamentales*. Madrid: Dykinson, S.L., 2023. https://www.digitaliapublishing.com/a/128997/inteligencia-artificial-y-derechos-fundamentales.

Ríos, Por Juan. “El gran engaño que vivió Scarlett Johansson con la IA y el uso de su voz con ChatGPT.” *Infobae*, August 28, 2024. https://www.infobae.com/tecno/2024/08/28/el-gran-engano-que-vivio-scarlett-johansson-con-la-ia-y-el-uso-de-su-voz-con-chatgpt/.

Schwartz, Oscar. “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation.” *IEEE Spectrum*, January 4, 2024. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

Shafer, Glenn, and Vladimir Vovk. “A Tutorial on Conformal Prediction.” *Journal of Machine Learning Research* 9 (March 2008): 371–421. doi:https://jmlr.org/papers/v9/shafer08a.html.

The White House. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” *The White House*, October 30, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

UNESCO. “Recommendation on the Ethics of Artificial Intelligence,” 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

United Nations. “Universal declaration of human rights.” United Nations, December 10, 1948. https://www.un.org/es/about-us/universal-declaration-of-human-rights.

United States. DoNotPay, Inc vs. FTC, No. 232-3042 (Federal Trade Commission January 14, 2025).

———. State v. Loomis, No. 2015AP157-CR (Supreme Court of Wisconsin July 13, 2016).

———. Walters v. OpenAI, L.L.C., 1:23-cv-03122, No. 1:23-cv-03122 (District Court, N.D. Georgia December 31, 2024).

Veale, Michael, and Frederik Zuiderveen Borgesius. “Demystifying the Draft EU Artificial Intelligence Act.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, July 31, 2021. https://papers.ssrn.com/abstract=3896852.

Veen, Christiaan van. “Landmark Judgment from the Netherlands on Digital Welfare States and Human Rights.” *OpenGlobalRights*, March 19, 2020. https://www.openglobalrights.org/landmark-judgment-from-netherlands-on-digital-welfare-states/?lang=English.

Vestri, Gabriele. “La fusión transformadora entre el sector público y la Inteligencia Artificial (IA): el ‘test de evaluación de impacto’ como prioridad.” *International Journal of Digital Law – IJDL* 4, no. 3 (March 9, 2024): 43–64. doi:10.47975/digital.law.vol.4.n.3.

Wachter, Sandra, and Brent Mittelstadt. “A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI.” *Columbia Business Law Review* 2019, no. 2 (October 5, 2018): 1–130. https://papers.ssrn.com/abstract=3248829.

Zhou, Qianwei, Margarita Zuley, Yuan Guo, Lu Yang, Bronwyn Nair, Adrienne Vargo, Suzanne Ghannam, Dooman Arefan, and Shandong Wu. “A Machine and Human Reader Study on AI Diagnosis Model Safety under Attacks of Adversarial Images.” *Nature Communications* 12, no. 1 (December 14, 2021): 7281. doi:10.1038/s41467-021-27577-x.