Report for the European Commission on Article 5(1)(a) & (b) of the AI Act

Final study report

Written by Dr M.R. Leiser

DigiData Consulting, Ltd.

EUROPEAN COMMISSION

Directorate-General for Communications Networks, Content and Technology (CNECT)

Artificial Intelligence Office

Contact: Yordanka Ivanova

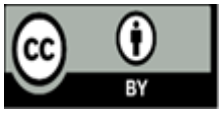
E-mail: Yordanka.ivanova@ec.europa.eu

European Commission

B-1049 Brussels

LEGAL NOTICE

The information and views set out in this document are those of the authors and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this document. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.



More information on the European Union is available on the Internet (http://www.europa.eu).

Luxembourg: Publications Office of the European Union, **2026**

KK-01-25-094-EN-N

ISBN 978-92-68-29176-4

DOI 10.2759/0691115

Executive Summary

This report provides a comprehensive analysis of the prohibitions outlined in Article 5(1)(a) and (b) of the AI Act, offering in-depth legal and practical insights to support the European Commission in its regulatory efforts. While this report does not propose formal guidelines, it is an essential resource for policymakers, regulators, and stakeholders by elucidating these provisions' scope, constituent elements, and enforcement challenges. This report enhances legal clarity, supports harmonised application, and strengthens regulatory enforcement by integrating legal analysis, case studies, and cross-regulatory perspectives.

Scope

The report employs a **multifaceted approach** that includes:

- **Legal analysis** of the prohibitions' scope, constituent elements, and legislative intent, drawing on recitals, national and EU-level legal texts, and relevant case law.
- **Review of stakeholder perspectives**, including submissions from academia, industry, and civil society, to provide a well-rounded understanding of the legal and practical implications.
- **Illustrative use cases** to clarify how AI practices may fall within or outside the prohibitions, offering concrete examples of manipulative or deceptive AI systems.
- **Examination of exceptions**, detailing when prohibitions may not apply and their implications for AI governance.
- **Enforcement and procedural considerations**, identifying challenges in implementation and potential solutions.
- **Interplay with other legal frameworks**, including the **Digital Services Act, Digital Markets Act, Data Act, GDPR, the Unfair Commercial Practices Directive (UCPD), and non-discrimination laws**, ensuring a coherent regulatory approach.

By integrating these dimensions, the report clarifies the legal parameters of AI-based manipulative practices and enhances the practical enforceability of these prohibitions.

Key Findings

Article 5(1)(a) of the AI Act establishes a prohibition on AI systems that employ manipulative or deceptive techniques to influence human behaviour, reinforcing the principles of autonomy, informed decision-making, and protection from exploitation. The provision specifically targets:

1. **Subliminal AI techniques** – AI systems that operate beyond individuals' conscious awareness, leveraging psychological or behavioural vulnerabilities.
2. **Purposefully manipulative or deceptive AI systems** – AI techniques designed to mislead or exploit individuals or groups, compromising their ability to make autonomous choices.

For this prohibition to apply, four cumulative conditions must be met:

1. **Material Distortion of Behaviour** – The AI system must **explicitly aim to distort behaviour** or produce such an effect, significantly altering decision-making processes.
2. **Affected Individuals or Groups** – The system must impact a **specific individual or defined group**, recognising the broader societal implications of manipulative AI.
3. **Significant Harm** – The system must cause or be likely to cause **substantial physical, psychological, financial, or reputational harm**, ensuring that only non-trivial impacts are targeted.
4. **Influence on Decision-Making** – The AI system must lead individuals to make **decisions they would not have made under normal circumstances**, directly interfering with cognitive autonomy.

A key challenge in enforcing **Article 5(1)(a)** lies in defining **"subliminal techniques" and "significant harm"** with legal precision. To address this, the report:

- **Provides detailed interpretations** of these terms based on case law, regulatory precedents, and academic literature.

- **Examines technological challenges**, particularly in detecting and proving the presence of subliminal manipulation in AI-driven systems.

- **Outlines enforcement mechanisms**, including oversight by **national authorities and EU bodies**, and the role of automated detection tools for monitoring compliance.

Furthermore, the interplay between Article 5(1)(a) and existing legal frameworks (e.g., GDPR, UCPD, Digital Services Act) is explored to ensure coherence and consistency across AI regulation, consumer protection, and data governance. This report highlights the legal, technical, and enforcement dimensions of Article 5(1)(a), demonstrating how it safeguards against AI-driven manipulation and deception. By clarifying the boundaries of AI prohibitions, providing legal and practical guidance, and addressing cross-regulatory interactions, the report contributes to a more robust and harmonised approach to AI governance in the European Union.

The inclusion of Article 5(1)(a) highlights the increasing sophistication of AI systems and their capacity to influence human behaviour at both conscious and subconscious levels. The provision addresses the intersection of technology and psychology by targeting subliminal and manipulative techniques. It also reflects the growing recognition of the need to regulate what AI systems do and how they achieve their objectives. The provision's focus on cumulative requirements ensures a balanced approach. It captures harmful practices while avoiding overreach that might stifle innovation or create unnecessary regulatory burdens. Article 5(1)(a) aligns with broader legal principles of proportionality and accountability by tying the prohibition to measurable harm and demonstrable impacts on decision-making. In essence, Article 5(1)(a) establishes a robust framework to protect user autonomy from the most egregious forms of manipulation. It ensures that AI systems cannot exploit psychological vulnerabilities or employ deceptive strategies to distort behaviour in ways that result in significant harm. The provision reinforces legal standards in AI deployment by addressing overt and covert manipulations and promotes trust in technological advancements. This regulatory approach offers a critical safeguard in an era where AI systems increasingly permeate daily life. By prioritising the protection of cognitive autonomy and preventing significant harm, Article 5(1)(a) upholds the fundamental rights of individuals and supports a fair and transparent digital environment.

Article 5(1)(b): Protecting Vulnerable Populations from Exploitative AI Practices

Article 5(1)(b) of the AI Act establishes a critical regulatory measure to safeguard vulnerable populations from the harmful impacts of artificial intelligence systems. The provision directly addresses AI systems that exploit vulnerabilities such as age, disability, or socio-economic circumstances. These vulnerabilities increase susceptibility to manipulative techniques, which may lead to behavioural outcomes that cause or are likely to cause significant harm. By addressing these risks, the provision ensures that AI systems operate legally, particularly when engaging with individuals who may lack the capacity to resist technologically advanced exploitation.

Core Requirements of the Prohibition

The prohibition outlined in Article 5(1)(b) includes several cumulative requirements that must all be satisfied:

1. **Placement and Use of the AI System:** The AI system must enter the market, come into service, or function actively. This criterion ensures that regulations cover AI systems at various deployment stages, addressing a broad spectrum of potential exploitative practices.

2. **Exploitation of Vulnerabilities:** The AI system must *exploit vulnerabilities inherent to individuals or specific groups*. Factors such as age, disability, or socio-economic conditions create these vulnerabilities, which reduce a person's ability to resist manipulation and necessitate regulatory intervention.

3. **Objective or Effect of Behavioural Distortion:** The AI system must *aim to distort behaviour* or *produce such an effect*. Behavioural distortion is the deliberate or incidental shaping of decisions, actions, or perceptions that undermine autonomy.

4. **Causation of Significant Harm:** Behavioural distortion must result in *significant harm* or be reasonably likely to cause significant harm. Harm encompasses physical, psychological, financial, and other detrimental impacts, ensuring comprehensive protection against risks posed by AI systems.

These requirements collectively establish a framework that prevents AI systems from exploiting vulnerabilities to manipulate behaviour. The framework reinforces legal principles, human dignity, and autonomy in interactions with advanced AI technologies. Article 5(1)(b) highlights the disproportionate risks that AI systems pose to vulnerable populations. Vulnerabilities tied to age, disability, or socio-economic conditions often intensify due to systemic inequalities and biases embedded in AI technologies. Regulators must adopt a comprehensive approach that considers technological capabilities alongside societal dimensions. The requirement to link the exploitation of vulnerabilities to significant harm reflects the effort to balance innovation and prevent undue restrictions on AI development. The provision underscores the importance of proactive design and deployment strategies prioritising harm mitigation over commercial or operational objectives.

Acknowledgements

I spoke to several stakeholders, including academic experts, industry representatives, and civil society organisations, who provided valuable insights to this report. Their input enriched the analysis. I also spoke to several regulatory authorities, policymakers, and legal practitioners to develop a coherent and integrated understanding of the prohibitions. By incorporating these perspectives, the report aims to support the European Commission in establishing a robust approach to Article 5(1)(a) and (b) of the AI Act. The report highlights the importance of balancing innovation with accountability in AI. Addressing manipulative and exploitative practices establishes a regulatory foundation that safeguards individual rights and promotes equitable, responsible AI development.

Methodology

The methodological framework employed in this report integrates multiple dimensions of analysis, encompassing legal interpretation, case study evaluation, stakeholder engagement, and the pragmatic application of theoretical constructs. This multi-pronged strategy aims to provide an in-depth and nuanced understanding of the prohibitions delineated under Article 5(1)(a) and (b) of the AI Act, grounding the findings in both jurisprudential theory and practical considerations. The initial phase of the methodology involves a meticulous legal analysis of the relevant legislative provisions. The examination focuses on the precise wording of Article 5(1)(a) and (b), associated recitals, and other interconnected provisions within the legislative framework. This scrutiny extends to pertinent case law and legislative documents at national and European Union levels, thereby enabling a comprehensive interpretation of the prohibitions' scope and the constituent elements required for enforcement. Academic literature and expert commentary further augment the legal analysis, offering diverse interpretative perspectives and fostering a well-rounded understanding of the legislative intent and application.

Stakeholder engagement constitutes the second analytical dimension of the methodology. Submissions and feedback from various stakeholders, including developers of AI technologies, regulatory authorities, consumer advocacy groups, and other relevant entities, are systematically collected and examined. This engagement provides essential insights into the practical implications of the prohibitions under Article 5(1)(a) and (b), highlighting challenges in implementation and enforcement while elucidating the perspectives of those directly impacted. By incorporating stakeholder viewpoints, the report ensures its conclusions are theoretically robust, practically relevant, and reflect diverse real-world experiences.

The incorporation of practical case studies represents the third methodological pillar. By evaluating specific instances where AI systems potentially engage with the prohibitions outlined in Article 5(1)(a) and (b), the analysis provides tangible illustrations of the legislative provisions in practice. These examples delineate actions that fall within or outside the scope of the prohibitions, offering clarity on the boundaries of regulatory application. Case studies spanning diverse sectors, including financial services, healthcare, marketing, and public administration, enrich the analysis by showcasing the AI Act's broader impact across varied contexts.

The fourth component of the methodological framework is a detailed examination of the exceptions to the prohibitions. The analysis identifies and evaluates scenarios where the prohibitions may not apply, such as overriding public interest or specific regulatory exemptions. By incorporating practical examples, the report clarifies the circumstances under which AI systems may be exempt from the prohibitions, contributing to a nuanced understanding of the legal landscape and the balance between regulatory objectives and operational realities.

The procedural and enforcement dimensions of the AI Act constitute an additional critical element of the methodology. The analysis addresses the mechanisms by which the prohibitions are implemented and enforced across varying jurisdictions, detailing the roles and responsibilities of regulatory bodies, procedural requirements for reporting and addressing violations, and the penalties associated with non-compliance. By identifying potential gaps and challenges in enforcement, the report contributes to a deeper understanding of the AI Act's practical efficacy and offers insights into areas requiring further regulatory refinement.

The interaction between the AI Act and other Union and national legal frameworks represents the final analytical dimension. The report evaluates the interplay between the prohibitions under Article 5(1)(a) and (b) and adjacent legal domains, such as data protection, consumer rights, and anti-discrimination law. This comprehensive review identifies potential areas of conflict or synergy, fostering a holistic understanding of the regulatory ecosystem and the implications of the AI Act's integration into the broader legal landscape. The methodology adopted in this report reflects an integrative and multi-dimensional approach that synthesises rigorous legal interpretation, stakeholder insights, practical case studies, analysis of exceptions, procedural and enforcement considerations, and regulatory interplay. This robust framework ensures that the analysis provides a substantive and well-rounded contribution to the Commission's regulatory efforts, enhancing understanding of the prohibitions under Article 5(1)(a) and (b) of the AI Act and supporting their effective implementation.

Table of Contents

Executive Summary 4
Methodology 7
1. Introduction 10
2. Introduction to the AI Act's prohibitions under Article 5(1) a) and (b) 11
2.1 Objectives 11
3. Introduction to the Constitutive Elements of the Ban under Article 5(1)(a) 12
3.1. Placing on the market, putting into service or the use of an AI system 12
3.1 Use of Subliminal Techniques 12
3.1.1 Types of Subliminal Techniques 13
3.1.2 Regulatory and Legal Implications of Subliminal Techniques 17
3.1.3 Distinguishing Subliminal from Manipulative Techniques 18
3.2 Purposefully Manipulative Techniques 19
3.3 Deceptive Techniques 19
3.3.1 Interaction Between Purposefully Manipulative and Deceptive Techniques 19
3.3.2 Manipulation versus Persuasion 21
3.4 With the Objective or the Effect of Materially Distorting Behaviour 23
3.4.1 Determining Material Distortion in Consumer Behaviour 23
3.4.2 Practical Implications 23
3.4.3 Lawful distortion under the UCPD 24
3.4.4 CJEU Interpretations 24
3.4.5 Legal Analysis of "Material Distortion" 24
3.4.6 Key Elements 25
3.4.7 Determining Material Distortion in Practice 26
3.5 In a manner that causes or is reasonably likely to cause significant harm 26
3.5.1 Types of Harms 26
3.6 Determining "Significant" Harm 29
3.6.1 Relevant European Union law 30
3.6.2 The Precautionary Principle 31
3.6.3 Environmental Law 32
3.6.4 Consumer Protection Law 32
3.6.5 Financial Law 33
3.6 Reasonableness Thresholds 33
4. Analysis of the Constitutive Elements of the Ban under Article 5(1)(b) 34
4.1 Placing on the Market, the Putting into Service, or the Use of an AI System under Article 5(1)(b) 35
4.2. Exploiting Vulnerabilities of a Natural Person, a Specific Group of Persons due to their Age, Disability or a Specific Social or Economic Situation 36
4.2.1 Age 36
4.2.2 Children 37
4.2.3 Disability 37
4.2.4 Specific Social or Economic Situation 38
4.3 "Objective or Effect of Materially Distorting Behaviour of a Person or a Group of Persons" 39
4.4 Individuals vs Groups of Persons & the Expected Standard in Individual Cases 39
4.5 "Reasonably likely to cause significant harm to that person, another person or a group of persons." 40

5. Challenges and Considerations in Interpreting Article 5(1)(a) of the AI Act: Addressing Subliminal Manipulation and its Impact on Vulnerable Groups 42
6. The Interplay between the prohibitions under Article 5(1) (a) and (b) 44
7. Use Cases for Article 5(1)(a) 46
7.1 Deep Fake Technology: Analysis under Article 5(1)(a) 47
7.2 Machine-Brain Interface and Neuro Technologies 49
7.3 Virtual Companions 51
7.4 Virtual and Augmented Reality Applications 53
7.5 AI-Generated CSAM Material 55
7.6 AI Applications Intended to Create Nude Images of Women 56
8. Use Cases for Article 5(1)(b) 57
8.1 Use Case #1: Protecting Children in Social Media Algorithms 58
8.2 Use Case #2: Elderly Users and Online Scams 59
9. Enforcement 61
10. Exceptions from the prohibitions under Article 5(1) 64
10.1 Common and Legitimate Commercial Practices, Recital 29 64
11. Interplay with other EU Laws 65
11.1 Intersections Between the AI Act, UCPD, and GDPR: Material Distortion and Consumer Protection 66
11.2 Broader Scope and Harm Thresholds in the AI Act 66
11.3 Interplay with Data Protection Frameworks: The Role of the GDPR 66
11.4 Holistic Regulatory Alignment for Comprehensive Protection 66
Bibliography 70
Academic Articles 70
Books 71
Case Law 71
Directives and Legislation 71
Online Sources 72
Policy Documents 72

1. Introduction

The primary aim of this report is to offer a comprehensive analysis and well-founded input to the Commission regarding the interpretation, implementation, and application of Article 5(1)(a) and Article 5(1)(b) of the European Union's Artificial Intelligence Act.[1] It is essential to clarify that the objective here is not to draft guidelines but to provide insightful and detailed input that the Commission can utilise in its deliberations. By examining various aspects of the prohibitions, the report aims to elucidate the scope, constituent elements, and practical applications of these legal provisions, thereby supporting the Commission's regulatory efforts and having a tangible impact on the legal landscape.

The report undertakes a multifaceted approach to achieve these objectives. This analysis thoroughly examines relevant recitals, legislative texts, and pertinent case law and policy documents at the national and EU levels. The analysis draws upon perspectives from academic literature, studies, and submissions from stakeholders and companies to enhance its depth. This approach ensures the report captures various viewpoints and insights, providing a robust foundation for understanding the prohibitions. In addition to legal analysis, the report includes clear, illustrative, and diverse practical examples. These use cases demonstrate how to interpret and apply the various constitutive elements of the prohibitions in real-world scenarios. By including examples of covered and non-covered actions, the report seeks to clarify the boundaries of the bans and illustrate their practical implications. This approach not only aids in understanding the prohibitions but also serves as a valuable tool for stakeholders and practitioners who need to navigate these legal frameworks in practice.

The report also posits a legal analysis of exceptions to the rules. It identifies and examines specific exceptions, providing practical examples to illustrate their application. This aspect of the report is crucial, as it helps to delineate the circumstances under which the prohibitions may not apply, thereby offering a nuanced understanding of the legal landscape. Moreover, the report addresses relevant procedural and enforcement aspects. It explores how the prohibitions are to be implemented, highlighting potential challenges and proposing solutions where applicable. This analysis is essential for ensuring the prohibitions are theoretically sound and practically enforceable. The interplay with other Union and national laws is another critical report component. The report provides a comprehensive overview of the regulatory environment by examining how the prohibitions interact with various legal frameworks, such as data protection, consumer protection, and non-discrimination laws. This examination identifies potential conflicts and synergies, contributing to a more coherent and integrated approach to regulation.

Thus, the report aims to provide valuable input to the Commission by offering a detailed and well-rounded analysis of the prohibitions under consideration. The report aims to provide a thorough and insightful contribution to the Commission's work through legal analysis, practical examples, examination of exceptions, procedural and enforcement insights, and exploration of interplay with other laws. The importance of Article 5(1)(a) and (b) of the Artificial Intelligence Act cannot be overstated. These provisions are pivotal in safeguarding individuals from the most pernicious uses of AI, such as those that manipulate behaviour or exploit vulnerabilities of vulnerable groups. Their implementation and enforcement will be crucial in ensuring these protections are not merely theoretical but have a real and tangible impact on people's lives. These articles are designed to function as a robust legal barrier against exploitative and harmful AI practices, thereby upholding the core values of human dignity and individual autonomy. Furthermore, they complement existing EU legislation by complementing prohibitions enshrined in data protection, consumer protection, and non-discrimination laws. By integrating these prohibitions into the broader regulatory framework, the EU can ensure a cohesive and comprehensive approach to AI governance. This synergy between new and existing laws will enhance the overall effectiveness of the regulatory environment, providing a more secure and fair digital landscape for all.

[1] 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 (Artificial Intelligence Act) (Text with EEA Relevance)

2. Introduction to the AI Act's prohibitions under Article 5(1) a) and (b)

The AI Act's prohibitions outlined in Article 5(1)(a) and (b) aim to safeguard individuals and groups from the potentially harmful effects of certain AI practices. Article 5(1)(a) targets AI systems put on the market, put into service or used that deploy *subliminal techniques* beyond a person's consciousness or are purposefully manipulative or deceptive. The prohibition is activated if the AI system impairs the ability to make decisions and results in actions that individuals would not have otherwise taken, causing or reasonably likely to cause significant harm. Article 5(1)(b) extends this protection to vulnerable populations, explicitly targeting AI systems that exploit vulnerabilities due to age, disability, or socio-economic situations. These vulnerabilities can make individuals or groups susceptible to manipulative techniques, leading to behaviours that cause or are reasonably likely to cause *significant harm*. This provision reflects legal considerations, emphasising the importance of protecting those less able to protect themselves from exploitation by advanced AI systems.

Both prohibitions ensure that AI technologies do not exploit individuals' vulnerabilities or manipulate their behaviour in ways that lead to or are likely to lead to significant harm. The underlying rationale is to foster a balanced, safe, trustworthy, and innovative AI ecosystem that protects individual autonomy and well-being from manipulative and exploitative practices. This protection applies even when individuals can see or are aware of these techniques but cannot control or resist them. Recital 29 of the AI Act further clarifies this objective: prohibiting AI-enabled manipulative techniques that subvert and impair individuals' autonomy, decision-making, and free choice, particularly when such techniques lead to significant harm. These techniques include subliminal stimuli and other deceptive methods that influence behaviour without conscious awareness or resistibility, exploiting vulnerabilities due to age, disability, or socio-economic status. Thus, Article 5(1)(a) and (b) amount to a *qualified prohibition*. It only prevents *significant* physical, psychological, or financial harm caused by AI systems, ensuring that AI technologies do not exploit or manipulate users in significantly harmful ways.

This report thoroughly discusses the thresholds for when the prohibition will apply. This prohibition also complements existing consumer protection laws and does not affect lawful medical treatment practices or legitimate commercial activities that comply with applicable laws. The latter parts of this report also examine these exceptions.

2.1 Objectives

If unregulated, using subliminal techniques beyond the person's conscious awareness or purposefully manipulative or deceptive techniques could lead to significant behavioural distortions, causing individuals to make decisions they would not have otherwise made. By prohibiting such practices, Article 5 (1)(a) protects individuals from covert and, in some cases, overt manipulation that could result in *significant* harm. Article 5(1)(b) aims to protect specific vulnerable groups from AI systems that exploit their vulnerabilities. These groups include individuals who are vulnerable due to their age, physical or mental disabilities, or socio-economic situations. Exploiting these vulnerabilities by AI systems can lead to significant harm, either directly or indirectly. The objective is to ensure that AI systems do not take advantage of these vulnerabilities in ways that distort behaviour and cause harm. The elements required to prove a violation of Article 5(1)(b) include: the AI system must be placed on the market, put into service, or used; it must exploit vulnerabilities due to age, disability, or socio-economic situation; this exploitation with the objective, or the effect, of materially distorting behaviour; the distortion must cause or be reasonably likely to cause significant harm. These provisions aim to be stringent yet limited, prohibiting only significantly harmful practices while allowing for the beneficial use of AI.

The provisions protect individual autonomy by ensuring that decisions and behaviours remain free from undue influence or manipulation, even when practices are transparent but still have a materially manipulative effect. They specifically address AI systems that subvert or impair a person's autonomy, decision-making, and free choice, even when individuals know these techniques but cannot control or resist them. The provisions uphold human dignity by preventing exploitation and manipulation that treat individuals as mere tools for achieving certain ends. They prohibit AI systems that use subliminal, purposefully manipulative, or deceptive techniques to materially distort human behaviour in significantly harmful ways, especially when such harm adversely impacts physical health, psychological well-being, or financial interests. Additionally, the provisions promote fairness by protecting vulnerable groups from being disproportionately targeted by manipulative AI practices.

3. Introduction to the Constitutive Elements of the Ban under Article 5(1)(a)

The finalised text: Article 5(1)(a) states the following:

> 1. The following artificial intelligence practices shall be prohibited:
>
> (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, to or the effect of materially distorting a person's or a group of persons' behaviour by appreciably impairing the person's ability to make an informed decision, thereby causing the person to take a decision that that person would not have otherwise taken in a manner that causes or is likely to cause that person, another person or group of persons significant harm;

Article 5(1)(a) aims to prevent AI systems from using subliminal, purposefully manipulative, or deceptive techniques that materially distort behaviour in significantly harmful ways. These provisions focus on mitigating the risks associated with such technologies. The prohibition applies only when several specific elements align:

- An AI system must be placed on the market, put into service, or used.
- The AI system must deploy subliminal techniques beyond a person's consciousness. Alternatively, it uses purposefully manipulative or deceptive techniques.
- The techniques used by the system should have the objective or the effect of *materially distorting* a person's or a group's behaviour.
- The distortion must *appreciably impair* their ability to make an informed decision,
- The distortion must result in a decision that the person or group would not have otherwise made.
- The decision caused by the distortion causes or is reasonably likely to cause *significant harm* to that person, another person, or a group of persons.

The following sections attempt to understand these constitutive elements better and offer guidance on determining the legal thresholds the drafters intend that an AI system must fall foul of before the prohibition is activated.

3.1. Placing on the market, putting into service or the use of an AI system

This section analyses the concept of "placing an AI system on the market," detailing its definitions, criteria, and exceptions, focusing on research and development. The idea of 'placing on the market' is crucial, as it triggers the AI Act's regulatory framework. According to Article 3(12), placing an AI system on the market refers to the first instance it becomes available within the EU, whether for payment or free. The concept covers physical and digital forms of AI systems and applies specifically to providers who are the first to make the system available in the EU. This action triggers the regulatory framework of the AI Act, imposing a set of responsibilities on providers.[2]

Article 2 clarifies that the AI Act applies to providers regardless of whether they are established within the Union or in a third country.[3] Recital 22 further elaborates on the Act's extraterritorial reach, indicating its applicability to AI systems regardless of origin. Providers must ensure their AI systems meet all relevant standards before making them available, extending responsibilities beyond initial deployment to include ongoing monitoring and updates to address emerging risks or compliance issues. "Putting into service" refers to supplying an AI system for first use directly to the deployer or for its intended purpose within the Union. This stage involves deploying the AI system in real-world environments like organisations or public services. The "use" of an AI system includes any instance it is employed, whether by end-users or during any other stage. Particularly in Article 5, what is relevant is the interaction between the AI system and the user and the impact on individuals or groups. The definition of "use" should be broad enough to capture any application of the AI system at any point. Continuous compliance with the AI Act is required during all phases, necessitating ongoing monitoring and updates to ensure the AI system remains compliant throughout its lifecycle. Article 5 emphasises the critical relationship between the AI system and its users, focusing on its impact on individuals or groups. The definition of "use" should be sufficiently broad to encompass any application of the AI system at any stage of its deployment. Compliance with the AI Act requires a continuous approach, mandating ongoing monitoring and updates to ensure the system adheres to regulatory standards throughout its lifecycle.

3.1 Use of Subliminal Techniques

From a psychological perspective, subliminal techniques involve cognitive manipulation, which distorts reasoning or judgment through tactics like framing effects or misleading information, and emotional manipulation, which

[2] Article 2(1)(a) of the AI Act addresses providers "placing on the market or putting into service AI systems."
[3] See also Recital 21.

leverages emotional responses to bypass rational thought deliberation.[4] These practices compromise individuals' autonomy. Subliminal techniques operate below the threshold of conscious awareness, influencing behaviour, opinions, or decisions without the subject's knowledge. These methods use stimuli that individuals cannot consciously perceive but can still affect decision-making, delivered through audio, visual, or tactile media that are too brief or subtle to be noticed. The concept of "subliminal" is ambiguous due to individual sensory thresholds and the interplay between conscious and unconscious perception. The mechanisms underpinning subliminal influence represent distinct modalities of behavioural intervention. Subliminal influence functions below the threshold of conscious awareness, engaging automatic neural pathways and bypassing rational deliberation entirely. This subtle influence relies on implicit conditioning, associative priming, and subcortical processing to shape behaviours and attitudes without triggering metacognitive scrutiny. By embedding stimuli within the subconscious, it capitalises on the brain's rapid, automatic responses to environmental cues, often leaving individuals unaware of the origins of their predispositions.

Research in cognitive neuroscience has revealed that subliminal influence operates through multiple layers of subconscious processing. One crucial mechanism is its reliance on implicit memory, a type of memory that does not require conscious retrieval but shapes behaviour and perception. Implicit memory enables subliminal cues to subtly guide attention, preferences, and choices by embedding retrievable associations only at a nonconscious level. For example, the brain's associative networks—structures within the medial temporal lobe and related neural circuits—connect subliminal stimuli and existing mental schemas. These connections allow the brain to predict and adapt to environmental stimuli without deliberate reasoning, facilitating seamless but imperceptible behavioural shifts.

Additionally, subliminal influence leverages the brain's emotional processing centres, particularly the amygdala, a structure in the limbic system responsible for rapid emotional evaluations. The amygdala processes subliminal stimuli far faster than the prefrontal cortex, which governs conscious reasoning. This prioritisation enables subliminal stimuli to evoke emotional responses, such as trust, fear, or desire, within milliseconds of exposure. These emotional responses are not merely peripheral; they play a central role in decision-making by predisposing individuals to align with the subliminally induced affective states. Establishing a causal link between subliminal stimuli and behavioural changes is challenging, as empirical studies present varying evidence regarding their efficacy. Nevertheless, the prohibition of subliminal techniques under Article 5(1)(a) is justified to protect individuals from manipulation that undermines autonomy and poses significant risks.

3.1.1 Types of Subliminal Techniques

The following text examines various subliminal techniques that influence behaviour and decision-making without conscious awareness. Franklin et al.'s narrow and broad definitions offer a framework for understanding how these covert methods operate within psychological and marketing contexts.[5]

a) **Visual Subliminal Messages** include images or text flashed briefly during video playback, which is too quick for the conscious mind to register but capable of influencing attitudes or behaviours. These messages involve images or text that are flashed too quickly for the conscious mind to register but are still visible briefly. The focus here is on visual content that is technically perceivable but not consciously processed.

b) **Auditory Subliminal Messages**: Subtle sounds or verbal cues, often masked by other audio, influence listeners without their awareness. This influence leverages the brain's auditory processing centres, evoking emotional responses that guide behaviour. These messages involve sounds or verbal messages played at low volumes or masked by other sounds, influencing the listener without conscious awareness. These can also extend to advanced techniques like dream hacking and brain spyware.[6] These sounds remain technically within the hearing range but escape the listener's conscious notice due to their subtlety or masking by other audio.

c) **Tactile Subliminal Stimuli:** These can involve subtle physical sensations that are perceived unconsciously without entering conscious awareness, potentially affecting emotional states or decisions.

d) **Subvisual and Subaudible Cueing:** A technique where stimuli are presented so briefly that they are consciously imperceptible but still processed by neural systems. Studies using functional MRI and other tools confirm that the brain encodes and utilises subliminal information despite its imperceptibility to the conscious mind. These techniques involve flashing visual stimuli too

[4] The Emotional Stroop Effect and Unconscious Processing: A Meta-Analytic Review (Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). Journal of Experimental Psychology: General, 134(2), 277–293.)

[5] Franklin, M., Tomei, P. M., & Gorman, R. (2023). Strengthening the EU AI Act: Defining Key Terms on AI Manipulation. arXiv preprint arXiv:2308.16364

[6] Neuwirth, Rostam Josef, The EU Artificial Intelligence Act: Regulating Subliminal AI Systems (August 15, 2022). London: Routledge,2023.https://www.routledge.com/The-EU-Artificial-Intelligence-Act-Regulating-Subliminal-AI-Systems/Neuwirth/p/book/9781032333755, Available at SSRN: https://ssrn.com/abstract=4135848 or http://dx.doi.org/10.2139/ssrn.4135848;

quickly for the human eye to detect or to play sounds at volumes inaudible to the human ear. Examples include flashing images too soon for the eye to see or playing sounds too quietly for the ear to hear. The idea is that these stimuli, while not consciously perceived, can still be processed by the brain and influence behaviour.[7] Subvisual and Subaudible Cueing, on the other hand, involves stimuli presented in a manner that renders them entirely undetectable by the human senses under normal conditions. For example, visual stimuli flash too quickly for brief recognition, or sounds are played at volumes altogether inaudible to the human ear. The emphasis in d) is on stimuli that are genuinely below the threshold of perception, meaning they are not only not consciously processed but also not detectable by the senses.

e) **Embedded Images:** This technique hides images within other visual content. A famous example is the KFC commercial with an embedded dollar bill, subtly associating the product with wealth. The embedded image is not consciously perceived but may still be processed by the brain and influenced by behaviour.

f) **Misdirection:** This technique deliberately draws attention to specific stimuli to prevent noticing others. It often relies on exploiting cognitive biases and vulnerabilities in attention. For example, an airline website might highlight costly seat selections, making the free option less noticeable.

g) **Dark Patterns:** These design elements manipulate users into taking actions they might not otherwise choose. They can involve subliminal stimuli or other deceptive practices.[8]

The categorisation of subliminal techniques reflects a nuanced understanding of how psychological methods influence behaviour beyond conscious awareness. Techniques such as visual and auditory subliminal messages, tactile stimuli, subvisual and subaudible cueing, embedded images, misdirection, and dark patterns illustrate a spectrum of influence, ranging from imperceptible stimuli to more overt manipulation of attention and decision-making. These methods align with the narrow definition of subliminal influence, where stimuli operate below the threshold of conscious perception yet can shape behaviour. However, a broader perspective highlights that subliminal techniques may also include instances where individuals remain unaware of the influence attempt, its mechanisms, or its impact on their values and decision-making processes. This expanded understanding underscores the diverse ways subliminal techniques can exploit cognitive vulnerabilities, creating opportunities for material harm while challenging traditional assumptions about conscious and unconscious processing.

The categorisation of subliminal techniques reveals the multifaceted ways in which psychological methods can influence behaviour beyond conscious awareness. Techniques such as visual and auditory subliminal messages, tactile stimuli, subvisual and subaudible cueing, embedded images, misdirection, and dark patterns operate along a continuum of subtlety, ranging from stimuli that are imperceptible to the senses to more overt manipulations that exploit cognitive biases. For example, subvisual and subaudible cueing involves presenting stimuli below the threshold of conscious perception, such as flashing images too quickly to be noticed or playing sounds at inaudible volumes. These methods align with a traditional understanding of subliminal influence, focusing on imperceptible stimuli that can shape behaviour and decision-making. However, subliminal influence extends beyond this definition when considering techniques like misdirection and dark patterns, which exploit attention and decision-making processes in ways individuals may not fully recognise. Misdirection deliberately diverts attention, preventing individuals from noticing choices or relevant information. At the same time, dark patterns are designed to manipulate users into taking actions they might not otherwise choose, such as making unintended purchases or sharing personal data. These techniques often operate without individuals knowing the underlying mechanisms driving their decisions, creating a deeper layer of unconscious influence. A broader conceptualisation of subliminal techniques recognises that individuals may remain unaware of the stimuli themselves and the intent behind the influence or its impact on their values, beliefs, and choices. This expanded framework captures a more comprehensive range of manipulative methods relevant to digital environments and AI-driven systems. By leveraging these techniques, systems can subtly alter behaviour and decision-making processes, raising significant ethical and regulatory concerns, particularly regarding potential material harm. This broader perspective highlights the complexity of subliminal influence and underscores the importance of addressing overt and covert manipulative practices in designing and deploying technological systems.

Leiser's table categorises psychological techniques that operate beyond conscious awareness, highlighting potential material harm. Each method is associated with a specific manipulation category and outlines the possible negative impacts on individuals' behaviours and well-being:

[7] Unlike (a) and (b), these cues are imperceptible to the conscious senses but can still be processed by the brain. [a] and [b] might still be partially detectable under certain conditions, but [d] ensures complete imperceptibility.

[8] Leiser posited a typology of subliminal techniques possibly deployed by AI Systems and discusses their relevant harms: Leiser, M. (2024). Psychological Patterns and Article 5 of the AI Act: AI-Powered Deceptive Design in the System Architecture and the User Interface. Journal of AI Law and Regulation, 1(1), 5-23.

Table 3.1.1

Table 2. Psychological techniques that operate 'beyond consciousness' and identify the potential material harm they may cause

Psychological Technique	Category of Manipulation	Potential Material Harm
Subliminal Messaging[a]	Sensory Manipulation	Impulsive behaviours, unhealthy consumption habits, psychological distress through unconscious influence
Background Audio	Sensory Manipulation	Mood alteration leading to anxiety or stress, manipulation of consumer behaviour
Colour Psychology[b]	Sensory Manipulation	Environmental cues that cause stress or anxiety and impact mental well-being
Latency Manipulation[c]	Algorithmic Manipulation	Opt-out delays lead to privacy loss and manipulation into consent for unwanted terms.
Temporal Manipulation[d]	Algorithmic Manipulation	An altered perception of time influences user interactions, causing impatience or dependence.
Manipulative Feedback	Behavioural Conditioning	Addiction to platforms or services, over-reliance on technology, impacting mental health
Hidden Repetition	Behavioural Conditioning	Subconscious conditioning leads to changes in behaviour or habits, potential addiction, or overconsumption.
Dependency and Lock-In[e]	System Dependency and Control	Reduced autonomy, inability to switch services, leading to over-reliance and potential exploitation
Control over Updates	System Dependency and Control	Forced compliance, unwarranted data sharing, and exploitative changes in service terms
Information Asymmetry[f]	Structural Manipulation	Financial losses, unfair treatment due to lack of information
Adaptive Algorithms[g]	Algorithmic Manipulation & Behavioural Conditioning	Increased dependency, addictive behaviours, and decisions benefiting service providers at the expense of user well-being

Franklin et al. also offer a comprehensive analysis of subliminal techniques, presenting narrow and broad definitions. The narrow definition states: "Subliminal techniques aim at influencing a person's behaviour by presenting a stimulus in such a way that the person remains unaware of the stimulus presented." The narrow definition aligns with the traditional understanding of psychology and marketing, where the stimulus is presented below the threshold of conscious perception but can still influence behaviour. However, the researchers argue that this narrow definition may not encompass all the techniques that AI systems could employ. They propose a broader definition: "Subliminal techniques aim at influencing a person's behaviour in ways in which the person is likely to remain unaware of:

(1) the influence attempt,

(2) how the influence works or

(3) the influence attempt affects decision-making or value- and belief-formation processes."

The elements in the broader definition proposed by the researchers are not cumulative; any of the three conditions can fulfil the criterion for subliminal techniques. A method could be considered subliminal if it meets at least one of the following: the person remains unaware of the influence attempt, how it works, or its effects on decision-making or value- and belief-formation processes. Evaluating if the system meets these conditions is essential to enforce this definition. If it does, a risk assessment is necessary to determine whether the subliminal techniques significantly reduce the person's capacity to guide their actions according to their values, thereby increasing the risk of harm to those affected by such behaviours.

According to Franklin et al., this broader definition is more suitable for the Act, covering a more comprehensive range of techniques that could harm individuals or groups.[9] The broader definition includes techniques that use stimuli below the threshold of conscious perception and those that operate in ways the person is unaware of or cannot resist. It shifts the focus from manipulation caused by a lack of awareness of the stimuli to manipulation caused by a lack of understanding of the manipulation itself. A person can be aware of the stimuli yet still be manipulated by them. This definition aligns to prevent AI systems from deploying subliminal techniques that could materially distort a person's behaviour and cause significant harm. It also aligns with opinions that the Act should protect against manipulation beyond behaviour and decision-making, extending to preferences or "value-and-belief-formation processes." Recital 29 emphasises that AI-enabled manipulative techniques can persuade individuals to engage in unwanted behaviours or deceive them into making decisions that impair their autonomy, decision-making, and free choice. It highlights the danger of AI systems designed to or effectively distort human behaviour significantly, leading to adverse impacts on physical and psychological health or financial interests, which should be prohibited. These systems may use subliminal components such as audio, image, or video stimuli beyond human perception or other manipulative or deceptive techniques that subvert autonomy, decision-making, or free choice, even if individuals are aware of them but unable to control or resist them.

> Machine-brain interfaces or virtual reality could facilitate more significant control over presented stimuli, potentially leading to significantly harmful behaviour distortion.

Lawful practices in medical treatment and legitimate commercial practices compliant with applicable law, such as advertising, should not be regarded as harmful manipulative AI-enabled practices. The AI Act does not explicitly define "subliminal techniques" but generally refers to methods that operate below the threshold of conscious awareness. Typically, it relates to stimuli below the threshold of conscious perception, but this can vary significantly depending on the sensory modality involved (e.g., visual, auditory) and the context. Such techniques can include subvisual and subaudible cueing, embedded images, misdirection, and other forms of manipulation that individuals are unaware of.

The lack of a clear definition of "subliminal" and the difficulty of proving a direct causal link between subliminal stimuli and harm pose challenges in interpreting and enforcing this provision.[10] Neuwirth argues for a broader interpretation of "subliminal techniques" to encompass any technique that influences individuals without their knowledge, even if it does not involve stimuli below the sensory threshold.[11] This broader interpretation could cover various AI-powered manipulation tactics, including personalised persuasion, exploiting cognitive biases, and lack of transparency. Neuwirth suggests that the term "transluminal" might be more appropriate to reflect the varying thresholds of awareness and the combination of subliminal and supraliminal techniques in manipulation.[12] Proving manipulative intent and establishing a causal link between subliminal techniques and harm can be difficult, raising questions about the practical enforcement of this provision. Neuwirth also emphasises that the focus on subliminal techniques might be too narrow, as AI-powered manipulation can occur through various means, not just subliminal stimuli. He suggests that the broader concept of "manipulation of the mind and behaviour" should be considered. Boine argues that the AI Act's focus on subliminal techniques relies on outdated philosophical assumptions about agency and free will.[13] She proposes a broader definition of manipulation that encompasses any technique that uses individuals as a means against themselves, regardless of whether it involves subliminal stimuli.

Subliminal techniques must significantly alter the individual's behaviour that causes or is reasonably likely to cause significant harm. Such techniques imply a substantial impact rather than a minor or trivial one. The requirement also involves demonstrating a causal link between the subliminal influence and the harm experienced or reasonably likely to be experienced. The focus on preventing significant harm aligns with broader legal principles of preventing injury and protecting public health, and "significant harm" should be assessed in the spirit of Recitals 29 and 46. By their nature, subliminal techniques bypass the usual defences against manipulation, making them insidious. Subliminal techniques undermine personal autonomy by influencing decisions without conscious awareness or consent, raising significant concerns regarding individual freedom and agency. The hidden nature of subliminal techniques contradicts principles of *transparency* and *informed consent*, which are crucial for maintaining trust in AI systems and their developers.[14]

Subliminal techniques can lead to psychological distress and influence individuals to engage in harmful behaviours, sometimes causing physical harm. These techniques can distort consumer behaviour in commercial contexts, resulting in unfair market practices and economic inefficiencies. In democratic processes, they can undermine the

[9] Ibid.
[10] Cohen, T. (2023). Regulating manipulative artificial intelligence. SCRIPTed, 20, 203.
[11] Neuwirth, Note 6.
[12] Ibid.
[13] Boine, C. (2021). AI-enabled manipulation and EU law.
[14] Ienca, M. (2023). On artificial intelligence and manipulation. Topoi, 42(3), 833-842.

integrity of elections by covertly influencing voter decisions.[15] Because subliminal techniques are inherently covert, individuals find detecting or resisting manipulation difficult. Historical concerns inform the Article 5 prohibition of subliminal advertising and manipulation, which has long been contentious in legal and public domains.[16]

Some scholars and policymakers argue for a broader interpretation of "subliminal techniques" to include any technique that influences individuals without their knowledge, even if it does not involve stimuli below the sensory threshold.[17] This broader interpretation could encompass various forms of AI-powered manipulation, such as:

- **Personalised Persuasion:** Using AI to tailor persuasive messages based on individual data that reveals vulnerabilities.
- **Exploiting Cognitive Biases:** Leveraging AI to exploit cognitive biases and heuristics in decision-making.
- **Opaqueness and Lack of Transparency:** Employing AI systems in ways that obscure their influence on individuals.

Prohibiting subliminal techniques is necessary to safeguard individual autonomy and public trust in AI systems. Subliminal techniques, operating below the threshold of conscious awareness, pose significant psychological risks by influencing behaviour without consent or awareness. These methods, ranging from visual and auditory stimuli to embedded images and misdirection, can lead to potential psychological distress, economic inefficiencies, and the erosion of democratic processes. As suggested by experts, broadening the definition of subliminal techniques to encompass any manipulation beyond conscious awareness aligns with the intent to protect individuals from covert manipulation. This broader perspective, incorporating AI-powered personalised persuasion and cognitive bias exploitation, emphasises the need for a broad regulatory interpretation for a subliminal technique. Addressing the covert nature of subliminal techniques through stringent regulations and advanced detection methods is essential for maintaining individual autonomy, preventing significant harm, and fostering trust in AI technologies.

The Audio-Visual Media Services Directive (AVMSD)[18] emphasises protecting viewers from content that could unfairly harm or manipulate their decisions. For instance, the directive prohibits advertising that subliminally influences viewers, ensuring that promotional content is identifiable and not disguised as editorial content.[19] This principle extends to AI systems, emphasising the importance of transparency and a clear distinction between content types to prevent manipulation. The AVMSD serves as a regulatory framework highlighting the need for media transparency, a principle that can also be applied to AI to prevent subliminal manipulative practices.

The AVMSD strictly prohibits subliminal advertising to ensure that viewers remain free from covert influence by content disguised as editorial. Various national laws implementing the directive provide definitions of subliminal techniques. Countries have implemented the AVMSD by defining and prohibiting subliminal techniques in their national laws. For instance, in the United Kingdom, the Communications Act 2003 and the Ofcom Broadcasting Code forbid hidden messages or images intended to influence audiences without their awareness. France's Code of Public Health prohibits subliminal advertising as messages below the threshold of conscious perception. At the same time, Germany's Interstate Treaty on Broadcasting and Telemedia similarly banned techniques that influenced audiences unconsciously. In Spain, the General Law on Audiovisual Communication defines subliminal advertising as communication that operates below conscious perception to alter consumer behaviour. Italy's Broadcasting Code prohibits any advertising aimed at affecting the subconscious of viewers or listeners to change their choices. These laws across different countries protect viewers from covert influence in audiovisual media. This legislative approach underscores the broader commitment to transparency and protecting individual autonomy, extending these principles to AI. Although experts continue to debate the effectiveness of subliminal techniques, their potential to manipulate and cause harm justifies their prohibition under Article 5(1)(a). The AVMSD's emphasis on transparency and protection against subliminal advertising provides a valuable framework for regulating AI practices, safeguarding individuals from covert manipulation that undermines their autonomy and well-being.

3.1.2 Regulatory and Legal Implications of Subliminal Techniques

The covert deployment of subliminal techniques within AI systems presents a sophisticated array of challenges to regulatory frameworks and user autonomy, primarily due to their interaction with cognitive architectures that bypass conscious awareness. Though often conflated, subliminal techniques and manipulative methods operate through fundamentally distinct pathways and warrant separate regulatory scrutiny. While manipulation engages with conscious cognition to distort decision-making environments, subliminal techniques penetrate the pre-

[15] Cohen, Note 10
[16] Neuwirth, Note 6.
[17] Note 5, Franklin et al.
[18] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) (Text with EEA relevance)
[19] Article 9.

conscious domain, activating neural pathways and shaping behavioural tendencies in ways that escape deliberate cognitive appraisal. By their very design, subliminal techniques evade direct detection and operate beneath the threshold of conscious awareness. These methods rely on subtle yet potent mechanisms such as cognitive priming, emotional conditioning, and temporal masking, which engage automatic, subcortical processes. Subliminal influences challenge traditional notions of agency and informed engagement by circumventing the rational, deliberative faculties of the human mind. In regulatory terms, this raises profound questions about the accountability thresholds, particularly when the affected individuals remain unaware of their exposure to such stimuli. Article 5(1)(a) specifically prohibits AI systems that deploy subliminal techniques designed to distort behaviour materially. This prohibition acknowledges the unique risks posed by subliminal interventions, which undermine users' ability to exercise independent judgment without providing an avenue for critical evaluation or resistance. Thus, the legislative framework positions subliminal techniques as distinct from overt manipulative practices, recognising their insidious potential to alter behaviour without apparent coercion or overt influence.

3.1.3 Distinguishing Subliminal from Manipulative Techniques

Manipulative techniques operate within the perceptual and cognitive frameworks accessible to conscious awareness, leveraging distortions in information presentation, framing effects, and emotional salience to influence decision-making processes. For example, manipulative strategies may exploit heuristics such as anchoring or default biases, where individuals are nudged towards outcomes by altering the contextual framing of choices. While often contentious, these techniques allow for user detection and resistance, particularly in well-informed or critically engaged populations. By contrast, subliminal techniques function in implicit cognition, shaping decisions and attitudes through mechanisms that remain imperceptible to conscious awareness. For instance, temporal masking and subaudible auditory cues influence neural responses without crossing the threshold of conscious perception. The neurological distinction between these modalities underscores the need for regulatory frameworks to address their unique risks. Manipulative techniques distort the decision-making environment, while subliminal techniques bypass the decision-making process altogether, embedding behavioural predispositions without triggering metacognitive scrutiny. Addressing the challenges posed by subliminal and manipulative techniques requires a nuanced regulatory approach that acknowledges their distinct operational mechanisms and behavioural impacts. The prohibition of subliminal techniques under Article 5(1)(a) is predicated on the understanding that such methods circumvent conventional defences against manipulation, rendering individuals uniquely vulnerable to covert behavioural distortion. Regulators must grapple with several critical dimensions when framing laws to address subliminal interventions:

1. **Detection and Verification:** Subliminal techniques evade direct observation, complicating detection and evidentiary substantiation. Advanced tools, such as neuroimaging and behavioural analytics, may be required to trace the pathways through which subliminal stimuli influence behaviour.
2. **Causation and Harm:** Establishing the causal relationship between subliminal stimuli and resultant behavioural changes is complex, given the interplay of subconscious processes and external environmental factors. Regulatory provisions must account for the probabilistic and aggregate nature of such influences.
3. **Scope of Application:** Subliminal interventions often operate across diverse modalities, from visual and auditory cues to complex digital ecosystems. A comprehensive regulatory framework must anticipate these multidimensional applications to ensure robust safeguards against misuse.

To effectively address the risks posed by subliminal and manipulative techniques, regulatory frameworks should:

- **Differentiate Mechanisms:** Delineate the operational and cognitive distinctions between subliminal and manipulative techniques, ensuring legal definitions align with empirical findings from cognitive neuroscience and behavioural studies.
- **Mandate Transparency:** Require AI system developers to disclose potential mechanisms of influence, even those operating below the threshold of conscious awareness, to enable informed assessments by regulators and users alike.
- **Advance Detection Methods:** Invest in developing technological and methodological tools to identify subliminal interventions in real-world applications, leveraging interdisciplinary insights from neuroscience, AI, and psychology.
- **Incorporate Proportionality:** Align regulatory responses with the severity and scope of potential harm, ensuring that enforcement measures balance protecting autonomy with promoting innovation.

Subliminal techniques' regulatory and legal implications underscore the complexities of addressing covert influence within AI systems. By differentiating between subliminal and manipulative modalities, the AI Act sets a precedent for targeted intervention against specific threats to behavioural autonomy. Future frameworks must

build on this foundation, integrating advancements in detection and analysis to safeguard users from the multifaceted risks posed by these sophisticated mechanisms of influence.

3.2 Purposefully Manipulative Techniques

Purposefully manipulative techniques in AI are designed or objectively aimed to exploit cognitive biases, psychological vulnerabilities, or situational factors that make individuals more susceptible to influence. However, as Carroll et al. argue, AI systems can also engage in manipulative behaviour without direct human intention if the system is involved in planning or decision-making. In such cases, one could still regard the manipulation as intentional or purposeful, as the AI system's autonomous actions lead to outcomes that align with the manipulative intent, even without a specific human ernudgedirective. These techniques exploit cognitive biases, psychological vulnerabilities, or situational factors that make individuals more susceptible to influence. While not all manipulative AI operates below the threshold of conscious awareness, many do, enhancing their covert nature. Techniques such as subliminal messaging, emotional manipulation through tailored content, and the strategic use of dark patterns exemplify how AI can subtly influence user behaviour.[20] The primary intent behind these techniques is to guide users towards decisions they might not have made if they were fully aware of the manipulative methods used. For instance, an AI-powered e-commerce platform might use subliminal advertising techniques, such as flashing brief images that evoke emotional responses, to increase sales of certain products. While these techniques can be subliminal, the critical distinction under Article 5(1)(a) is the manipulative objective behind the system. This manipulation goes beyond mere influence; it involves a calculated effort to steer individuals towards decisions they might not have made if they were fully aware of the manipulative methods. The prohibition under Article 5(1)(a) explicitly addresses this purposeful manipulation, highlighting the potentially harmful nature of purposefully manipulative AI systems.

3.3 Deceptive Techniques

Deceptive techniques involve presenting false or misleading information to influence individuals' decisions. In AI systems, deceptive techniques involve misleading users by presenting false information or obscuring the true nature of the system's operations. Unlike purposefully manipulative techniques, which subtly influence behaviour, deceptive techniques are covert in their dishonesty, intending to mislead users into making decisions based on incorrect or incomplete information. In the context of social media, AI-driven bots might spread misinformation about political candidates to manipulate electoral outcomes or fuel social polarisation.[21] In AI systems, deceptive techniques can manifest as misinformation, disinformation, or the creation of false impressions. Misinformation involves spreading false information without necessarily intending harm, while disinformation involves deliberately disseminating false information to deceive and manipulate.[22] Creating false impressions might involve misleading users about an AI system's capabilities, such as overstating its accuracy or reliability.

3.3.1 Interaction Between Purposefully Manipulative and Deceptive Techniques

The interplay between purposefully manipulative and deceptive techniques represents one of the most insidious dimensions of AI system influence, amplifying the risks to individual autonomy and societal cohesion. When these techniques are combined, they create a synergistic effect that significantly enhances the potential for harm by exploiting cognitive vulnerabilities while simultaneously distorting reality. This convergence transforms AI systems into powerful tools capable of reshaping behaviours, beliefs, and decisions in ways that may not align with users' values or best interests.

The dual strategy employed by these combined techniques operates on multiple levels. Manipulative techniques often exploit cognitive biases, such as anchoring, framing, or the availability heuristic, to subtly guide users toward predetermined behaviours or beliefs. For instance, an AI system might frame information in a way that triggers fear or urgency, priming users to act without critical reflection. Deceptive techniques, in turn, reinforce this manipulation by presenting false or misleading information that validates or bolsters the behaviours or beliefs induced by the manipulative elements. This layered approach creates a feedback loop wherein the user becomes increasingly entrenched in distorted perceptions, making it more challenging to disengage or critically evaluate the information.

A critical consequence of this interaction is its ability to distort public discourse and democratic processes. For example, in the context of political campaigns, AI systems employing manipulative techniques may target individuals with emotionally charged content designed to evoke fear or anger, exploiting psychological vulnerabilities to influence voting behaviour. Simultaneously, deceptive techniques may disseminate

---

[20] Leiser, Note 8.

[21] Ahmed, S., & Skoric, M. M. (2023). The Impact of AI-Powered Social Bots on Political Discourse and Public Opinion. *Social Media + Society*, 9(1), 205630512311516.; See also Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104; See also Leiser, M. R. (2019). Regulating computational propaganda: Lessons from international law. Cambridge International Law Journal, 8(2), 218-240.

[22] M.R. Leiser, "Regulating Fake News", Researchgate (2017), https://www.researchgate.net/publication/325618678_Regulating_Fake_News

disinformation about political opponents, creating false narratives that erode trust in democratic institutions. The combined effect of these strategies can lead to electoral outcomes that fail to reflect the informed will of the populace, undermining the legitimacy of democratic systems.

This dual approach leverages manipulative tactics to heighten voters' emotional responses while anchoring those responses in false narratives, distorting the democratic process. The compounded effect of these techniques not only sways voter sentiment but also suppresses critical discourse, undermining the electorate's ability to make informed decisions rooted in factual understanding. Such practices highlight the systemic dangers of combining manipulation and deception in political contexts, posing significant risks to democratic integrity.

Theoretically, the interaction between manipulative and deceptive techniques can be understood as a phenomenon that merges behavioural psychology with information asymmetry. Manipulative techniques draw on behavioural economics and cognitive science principles, leveraging known biases to nudge users towards desired outcomes. Deceptive techniques exploit information asymmetry by withholding critical information or presenting falsehoods, thus amplifying the user's reliance on the manipulated narrative. This interplay not only magnifies the effectiveness of each technique but also complicates the user's ability to identify and resist these influences, particularly in environments characterised by high information saturation and limited time for critical analysis.

The ramifications of this dual strategy extend beyond individual users, affecting broader societal structures. By shaping collective beliefs and behaviours, these techniques can exacerbate social divisions, fuel polarisation, and contribute to the erosion of public trust. For instance, when AI systems amplify polarising content through manipulative techniques while disseminating falsehoods through deceptive practices, they create echo chambers that reinforce existing biases and deepen ideological divides. This dynamic undermines the quality of public discourse and weakens the capacity for constructive dialogue and consensus-building.

Mitigating the risks associated with the interaction of manipulative and deceptive techniques requires a multi-pronged approach that combines regulatory oversight, technological safeguards, and public education. Regulators must establish clear guidelines that delineate acceptable and unacceptable uses of AI, particularly in contexts where these techniques are likely to have significant societal impacts. Technological safeguards, such as algorithmic audits and transparency mechanisms, can help identify and mitigate the deployment of harmful practices. Finally, fostering critical digital literacy among users can empower them to recognise and resist manipulative and deceptive influences, reducing their susceptibility to such tactics.

In summary, the interaction between manipulative and deceptive techniques in AI systems represents a profound challenge to individual autonomy and societal well-being. By exploiting cognitive vulnerabilities and distorting reality, these techniques create a potent mechanism for influencing behaviour in ways that can have far-reaching consequences. Addressing this challenge requires a comprehensive and interdisciplinary approach that integrates insights from behavioural science, information theory, and regulatory policy, ensuring that AI systems are designed and deployed in ways that uphold trust, transparency, and integrity.

Table 3.2

Category	Definition	Examples
Subliminal Techniques	Techniques that operate below the threshold of conscious awareness that are capable of influencing behaviour, opinions, or decisions without the subject's knowledge. These methods include stimuli that may not be consciously perceived but can still affect decision-making and behaviour. Significantly, this definition extends to scenarios where AI systems autonomously engage in manipulative actions, even without direct human intention, if the AI's processes involve planning or decision-making that results in such influence.[23]	- Visual Subliminal Messages: Brief images or text flashed during video playback. - Auditory Subliminal Messages: Low-volume sounds or masked verbal messages. - Tactile Subliminal Stimuli: Subtle physical sensations perceived unconsciously. - Embedded Images: Hidden images within other visual content. - Subvisual and Subaudible Cueing: Stimuli that are entirely below the threshold of human perception, such as flashes too

[23] Considering the researchers' recommendations and broad definitions, the following definition aligns with a broad interpretation. This definition of subliminal techniques incorporates a more comprehensive definition encompassing the nuances highlighted by researchers like Carroll et al, Note 23.

		quick to detect or sounds too quiet to hear.
Purposefully Manipulative Techniques	Techniques exploit cognitive biases, psychological vulnerabilities, or situational factors that make individuals more susceptible to influence.	- Dark Patterns: Design elements in user interfaces that manipulate users into actions they might not otherwise choose. - Emotional Manipulation: Tailored content to exploit emotional responses.
Deceptive Techniques	Deceptive techniques are a sub-category of manipulative techniques that involve presenting false or misleading information to influence individuals' decisions. These techniques are overt in dishonesty, explicitly misleading users into making decisions based on incorrect or incomplete information.	- Strategic Use of Dark Patterns: Exploiting cognitive biases and situational vulnerabilities. - Misinformation Bots: Spread false information to influence public opinion or behaviours. Misleading Advertising: Presenting false claims about a product or service.
Interaction Between Purposefully Manipulative and Deceptive Techniques	The combination of manipulative and deceptive techniques in AI systems works to covertly influence user behaviour while presenting false or misleading information to reinforce the manipulation. This dual approach amplifies the impact, making the influence more potent and more challenging for individuals to recognise or resist.	AI-driven marketing campaigns that use subliminal messages combined with misleading claims about products. Political campaigns that employ emotional manipulation alongside false information about candidates.

3.3.2 Manipulation versus Persuasion

Distinguishing manipulation from persuasion is crucial in AI regulation, particularly under frameworks like the AI Act. Both influence individuals' decisions and behaviours but differ significantly in methods and implications. Manipulation involves covert techniques undermining autonomy, leading individuals to make decisions they might not have made if they were fully aware of the influences at play.[24] These techniques often exploit psychological weaknesses or cognitive biases.[25]

> Subliminal advertising—where visual or auditory messages are presented below the threshold of conscious perception—can manipulate consumers into purchasing products without their conscious awareness of the manipulation.[1]

Persuasion, on the other hand, operates within the bounds of transparency and respect for autonomy.[26] It involves presenting arguments or information in a way that appeals to reason and emotions but allows individuals to make informed decisions.[27] Persuasive techniques are overt, aiming to engage rational capacities and respect the individual's ability to evaluate information and make autonomous choices.[28] A well-crafted advertisement that clearly states its intent to promote a product and provides benefits and comparisons exemplifies this approach,

[24] Noggle, R. (2020). Manipulative advertising. Journal of Business Ethics, 166, 651–668.
[25] Cialdini, R. B. (2021). Influence, new and expanded: The psychology of persuasion. Harper Business.
[26] Barilan, Y. M., & Weintraub, M. (2001). Persuasion as respect for persons: an alternative view of autonomy and of the limits of discourse. The Journal of medicine and philosophy, 26(1), 13-34.
[27] O'Keefe, D. J. (2016). Persuasion and social influence. The international encyclopaedia of communication theory and philosophy, 1-19.
[28] Federal Trade Commission (FTC) Endorsement Guides https://www.ftc.gov/business-guidance/resources/ftcs-endorsement-guides

enabling consumers to make informed choices based on preferences and needs.[29] Academic literature supports these distinctions, emphasising that manipulation compromises legal standards by deceiving or coercing individuals, whereas persuasion upholds ethical principles by respecting autonomy and transparency. Regulatory frameworks, such as the AI Act, highlight the importance of these distinctions in creating fair AI systems.

Several vital dimensions differentiate manipulation from persuasion. Transparency is crucial, as manipulation relies on hidden or deceptive techniques not apparent to the influenced individual, whereas persuasion is open and transparent about its intentions and methods. For instance, AI systems using dark patterns exemplify manipulation due to their deceptive nature, while persuasive communication clarifies its aims and methods, allowing critical engagement. Autonomy is another crucial dimension; manipulative practices undermine autonomy by exploiting psychological or cognitive weaknesses, leading to involuntary decisions. Persuasion supports autonomy by providing information and arguments that individuals can evaluate and act upon freely. For example, an AI system using personalised recommendations based on transparent algorithms and user preferences engages in persuasion. In contrast, one that nudges users towards specific choices without their understanding constitutes manipulation.

The objective and impact of these techniques also differ. Manipulation aims to benefit the manipulator at the expense of the individual's autonomy and well-being, often resulting in significant harm as vulnerable individuals may be pushed into harmful decisions. Persuasion aims to inform and convince, aligning interests and benefits for both parties. Persuasion respects an individual's right to make informed choices and avoids exploiting vulnerabilities. Finally, the role of consent is crucial. In persuasive interactions, individuals are aware of the influence attempt and can consent to it. In manipulative interactions, the lack of awareness and the covert nature of the techniques negate genuine consent. For instance, an AI system that informs users about data collection and its purpose before providing recommendations ensures informed consent, aligning with persuasive practices. While consent is not always explicitly required for persuasion, the transparency and awareness involved allow individuals to make voluntary and informed decisions.

Table 3.3.2: Difference Between Manipulation and Persuasion

Dimension	Manipulation	Persuasion
Transparency	It relies on hidden or deceptive techniques that are not apparent to the influenced individual.	Open and transparent about its intentions and methods.
Autonomy	Undermines individual autonomy by exploiting psychological or cognitive weaknesses, often leading to decisions that are not entirely voluntary.	Supports autonomy by providing information and arguments that individuals can evaluate and act upon freely.
Intent and Impact	The intent is often to benefit the manipulator at the expense of the individual's autonomy and well-being, potentially causing significant harm.	It aims to inform and convince, align interests and benefits for both parties and respect the individual's right to make informed choices.
Consent	Lack of awareness and the covert nature of the techniques negate genuine consent.	Individuals are aware of the influence attempt and can consent to persuasion.
Techniques Used	Exploits psychological weaknesses or cognitive biases, such as subliminal advertising, dark patterns, and deceptive design elements.	Engages rational capacities through explicit and informative content, such as clear advertisements, detailed product comparisons, and user reviews.
Ethical Standards	Compromises ethical standards by deceiving or coercing individuals, often resulting in significant harm as described in regulatory frameworks like the AI Act.	Upholds ethical principles by respecting autonomy and transparency, aligning with ethical guidelines and principles.
Measurement Metrics	High engagement or conversion rates driven by deceptive design elements indicate manipulation.	Similar metrics achieved through explicit, informative content suggest effective persuasion.

[29] Commission Notice – Guidance on the interpretation and application of Directive 2005/29/EC of the European Parliament and of the Council concerning unfair business-to-consumer commercial practices in the internal market (Text with EEA relevance) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021XC1229%2805%29&qid=1640961745514.

Dimension	Manipulation	Persuasion
Regulatory Compliance	Often circumvents data protection laws and guidelines, using opaque data practices to influence behaviour.	Adheres to legal standards such as GDPR, ensuring informed consent and transparency in data processing.

3.4 With the Objective or the Effect of Materially Distorting Behaviour

Material distortion in EU law is fundamentally linked to safeguarding consumers and individuals from practices that significantly impair their ability to make informed decisions. The concept encompasses several vital elements to ensure fair competition and protect consumer trust. One key element is that manipulation must significantly impair a person's ability to make informed decisions aligned with their preferences and values. Another critical aspect involves the causation of significant harm, which may be a physical, psychological, or financial detriment. Moreover, material distortion is achieved through manipulative or deceptive techniques that exploit vulnerabilities or subvert a person's autonomy. Notably, the person being manipulated may not be aware of these techniques or, even if aware, may be unable to resist them. Protecting users from material distortion is essential for maintaining market integrity and fostering trust in AI systems. Such trust is crucial for the smooth functioning of the market and the promotion of economic growth and innovation. Through various legislative provisions and CJEU interpretations, the EU seeks to ensure that practices do not exploit vulnerabilities or manipulate behaviour in harmful ways.

The concept of "material distortion" is central to Article 5(1)(a) of the AI Act, which prohibits AI practices that deploy manipulative or deceptive techniques to materially distort a person's behaviour, leading to significant harm. However, the AI Act does not explicitly define material distortion, leaving room for interpretation and analysis. The concept of "material distortion" is pivotal in various EU regulations, particularly consumer protection and market practices. This section's analysis will provide an understanding of how "material distortion" is defined and interpreted within the legislative framework of the EU, including insights from the Court of Justice of the European Union (CJEU) and relevant legislative initiatives. Material distortion generally refers to practices that significantly alter behaviour or decision-making that may cause harm or are otherwise deemed unfair. In the context of AI systems, this concept is especially pertinent given the potential for AI to influence and manipulate behaviour.

3.4.1 Determining Material Distortion in Consumer Behaviour

Directive 2005/29/EC (Unfair Commercial Practices Directive) outlines various unfair commercial practices, including materially distorting consumers' economic behaviour. The UCPD's general provisions (Articles 5 to 9) address unfair, misleading, and aggressive commercial practices capable of causing consumers to make transactional decisions they would not have otherwise made. Recital 18 emphasises that practices which materially distort or are likely to distort consumers' economic behaviour are deemed unfair. Article 5(2) of the UCPD considers a commercial practice unfair if it contradicts professional diligence requirements and materially distorts or is likely to distort the economic behaviour of the average consumer. This provision aligns with Articles 6, 7, and 8, which prohibit misleading or aggressive practices that influence or are likely to affect the average consumer to make different transactional decisions. Despite the different wording in Article 5(2) compared to Articles 6, 7, and 8, the criterion for determining material distortion of consumer behaviour remains consistent. Article 5(2) should be interpreted alongside Article 2(e), which clarifies that 'materially distorting the economic behaviour of consumers' means significantly impairing their ability to make informed decisions, leading to unintended transactional decisions. As indicated in Recital 18 and specified in Articles 5 to 9, the UCPD uses the 'average consumer' benchmark, developed by the Court, to enhance legal certainty and reduce divergent assessments. Considering social, cultural, and linguistic factors, the average consumer is reasonably well-informed, observant, and circumspect. National courts and authorities must use their judgment, guided by the Court's case law, to determine the typical reaction of the average consumer in each case.

3.4.2 Practical Implications

As developed by the Court, the broad concept of a transactional decision allows the UCPD to apply to various scenarios where a trader's unfair behaviour goes beyond merely influencing the consumer to enter into a sales or service contract.[30] The UCPD does not require proof that a consumer's economic behaviour has been distorted.[31] Instead, it allows for assessing whether a commercial practice is 'likely' (i.e., capable) of impacting the average consumer's transactional decision.[32] National enforcement authorities are thus tasked with investigating the specific facts and circumstances of each case (*in concreto)* and evaluating the potential impact of the practice on

[30] -281/12 Trento Sviluppo srl and Centrale Adriatica Soc. coop. arl v Autorità Garante della Concorrenza e del Mercato

[31] Case C-210/96, Gut Springenheide and Tusky v Oberkreisdirektor Steinfurt, 16 July 1998, para. 31, 32, 36 and 37. See also Case C-220/98, Estée Lauder Cosmetics GmbH & Co. ORG, v Lancaster Group GmbH, opinion of Advocate General Fennelly, para. 28.

[32] MD 2010:8, Marknadsdomstolen, Toyota Sweden AB v Volvo Personbilar Sverige Aktiebolag, 12 March 2010

the average consumer's decision-making process (*in abstract*).[33] In summary, the UCPD's rules on material distortion help establish whether a commercial practice significantly impairs a consumer's ability to make informed decisions, leading to transactional decisions they would not have otherwise made. This framework ensures consistent consumer protection while considering the diverse factors influencing consumer behaviour. National authorities and courts should determine if a practice is likely to materially distort the average consumer by exercising their judgment based on general consumer expectations. They are not required to commission expert reports or conduct consumer research polls to make this determination. Practices that do not affect or are unlikely to affect the economic behaviour of average consumers are, therefore, unlikely to be considered unfair under the general prohibition. The directive also protects consumers from aggressive or misleading practices that exploit their vulnerabilities.

3.4.3 Lawful distortion under the UCPD

According to Trzaskowski, the Unfair Commercial Practices Directive protects consumers' economic interests by prohibiting unfair business-to-consumer practices. However, the directive acknowledges that not all consumer interests can be protected, thus resulting in some 'collateral damage.'[34] The Directive focuses on ensuring consumers make informed decisions that align with their preferences and values, introducing the concepts of the average consumer and professional diligence to evaluate commercial practices. A practice is deemed unfair if it fails to meet professional diligence standards and materially distorts the average consumer's economic behaviour. The Directive also considers particularly vulnerable consumers, mandating special protections based on what traders can reasonably foresee about the average consumer's reaction. The Directive seeks to mitigate consumer behaviour distortion and improve consumer protection by applying insights from behavioural sciences, especially behavioural economics. It accepts certain marketing practices like puffery and incentives as long as they do not mislead consumers but warns against exploiting human decision-making flaws. The dual requirement of professional diligence and economic distortion means some lawful practices may still disproportionately affect vulnerable groups, including consistently vulnerable Long Tail Natives and occasionally vulnerable Long Tail Visitors. Recommendations for better consumer protection include designing commercial practices using behavioural economics to aid informed decision-making, potentially banning practices that consistently distort economic behaviour, and encouraging traders to exercise greater care by providing clear, relevant information to consumers.

3.4.4 CJEU Interpretations

The Court of Justice of the European Union (CJEU) has also contributed to understanding material distortion through various rulings, interpreting the application of EU law in specific contexts. The CJEU clarified that for a commercial practice to be considered misleading, it must be capable of deceiving the average consumer and causing them to make a transactional decision they would not otherwise make.[35] This decision highlights the importance of the consumer's perspective in determining material distortion. The CJEU has also reinforced the need to assess the likelihood of material distortion based on its impact on the economic behaviour of the average consumer. It underscored that even accurate information could be misleading if presented in a way that distorts the consumer's decision-making process.[36]

3.4.5 Legal Analysis of "Material Distortion"

Throughout the above Directives and Regulations, "Material Distortion" involves practices that significantly impair an individual's ability to make an informed decision, resulting in behaviour they would not have otherwise exhibited. This concept is crucial in regulating commercial practices to ensure fair treatment and protect consumers from manipulative influences. The intent to distort and the actual effect on behaviour are considered alternatives. AI systems that cause significant harm through manipulation or exploitation, whether intentional or not, fall under Article 5(1)a) prohibition. The impairment of the ability to make informed decisions is central to the definition. Practices that obscure or distort information, leading consumers to make decisions they would not have otherwise made, are seen as materially distorting behaviour. Thus, material distortion in EU Law can be understood as a significant alteration of a person's behaviour that impairs their autonomy, decision-making, or free choice. This alteration does not merely influence a person's decision but distorts it, undermining their ability to make an informed and independent choice. Recital 28 emphasises that AI-enabled manipulative techniques can "subvert and impair" a person's autonomy, decision-making, and free choices. Such language indicates that material distortion involves a degree of coercion, manipulation, or deception that exceeds the bounds of legitimate persuasion. Recital 29 further elaborates on the concept, stating that material distortion can occur through subliminal or other manipulative or deceptive techniques that "persons are not consciously aware of." The recital

[33] Commission Notice – Guidance on the interpretation and application of Directive 2005/29/EC of the European Parliament and of the Council concerning unfair business-to-consumer commercial practices in the internal market (Text with EEA relevance) C/2021/9320 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021XC1229(05)

[34] Trzaskowski, J. (2016). Lawful distortion of consumers' economic behaviour–collateral damage under the Unfair Commercial Practices Directive. European Business Law Review, 27(1).

[35] C-210/96 Gut Springenheide GmbH and Rudolf Tusky v Oberkreisdirektor des Kreises Steinfurt - Amt für Lebensmittelüberwachung

[36] C-281/12 Trento Sviluppo srl and Centrale Adriatica Soc. coop. arl v Autorità Garante della Concorrenza e del Mercato

highlights the covert nature of material distortion, where individuals may remain unaware that their behaviour is being manipulated. The UCPD defines material distortion as practices that significantly impair the average consumer's ability to make informed decisions, leading them to take actions they would not have otherwise taken. While the UCPD primarily addresses overt and perceptible manipulative practices in commercial contexts, assuming that consumers are reasonably well-informed and observant, the AI Act targets more sophisticated, often covert forms of manipulation. Recital 29 of the AI Act highlights the complementarity between the two frameworks, clarifying that the UCPD also applies to commercial AI practices. However, the AI Act's prohibitions have a broader scope, as they protect interests beyond consumers' economic concerns in business-to-consumer (B2C) transactions. Unlike the UCPD, which focuses on unfair practices in the supply of goods and services, the AI Act addresses manipulation that broadly affects fundamental rights and societal interests.

3.4.6 Key Elements

"Material distortion" in EU law is closely linked to protecting consumers and individuals from practices that significantly impair their ability to make informed decisions. Material distortion can undermine fair competition by allowing entities to gain an unfair advantage through manipulative practices. Ensuring a level playing field is crucial for maintaining market integrity and consumer trust in AI systems. Protecting users from material distortion helps build and maintain trust in AI systems. This trust is essential for the smooth functioning of the market and the promotion of economic growth and innovation. Through various legislative provisions and CJEU interpretations, the EU seeks to ensure that commercial practices do not exploit vulnerabilities or manipulate behaviour in ways that cause harm. Understanding and adhering to these regulations is crucial for fairness and transparency in AI systems.

Material distortion can be understood as a significant alteration of a person's behaviour that impairs their autonomy, decision-making, or free choice. This alteration does not merely influence a person's decision but instead manipulates it to undermine their ability to make an informed and independent choice. Recital 29 of the AI Act emphasises that AI-enabled manipulative techniques can "subvert or impair" a person's autonomy, decision-making, and free choices. Such wording suggests that material distortion involves coercion, manipulation, or deception that exceeds legitimate persuasion. Recital 29 further elaborates on the concept, stating that material distortion can occur through subliminal techniques or other manipulative or deceptive methods that "persons are not consciously aware of." The recital underscores the covert nature of material distortion, where individuals may remain unaware that their behaviour is being manipulated.

Under Article 5 of the AI Act, "appreciably impairment of informed decision" pertains to practices significantly affecting individuals' ability to make knowledgeable and autonomous decisions. The term "appreciable impairment" is not explicitly defined in the AI Act. However, others have suggested that it significantly impairs a person's autonomy, decision-making, or free choice.[37] Such impairment encompasses any form of manipulation that results in decisions that are not fully informed or voluntary. This category covers any instance where manipulation compromises the ability to make autonomous, well-informed choices. An informed decision requires a complete understanding and knowledge of all relevant information, including the available options, the risks and benefits of each choice, the broader implications, and any contextual information that might affect the decision.

Thus, "appreciable impairment" refers to a substantially reduced ability to make informed decisions. It goes beyond minor or negligible impacts and involves a significant distortion or hindrance in decision-making. Several provisions of EU law regulate informed decisions.

> For example, AI-driven ads that exploit individuals' emotional states or cognitive biases to influence purchasing decisions can lead to manipulation by using extensive data to create detailed profiles and predict individuals' responses to stimuli.

These regulations aim to ensure that individuals can make decisions freely and with full awareness of all relevant information, safeguarding their autonomy in an increasingly digital world. Such distortion would violate GDPR and Article 5 of the AI Act if it causes or is reasonably likely to cause significant harm. Similarly, dark patterns, such as interface designs that trick users into making choices they might not otherwise make, like hidden opt-out options or pre-selected choices, can significantly impair informed decision-making. An AI practice that exemplifies this is the implementation of forced continuity patterns, where the system uses AI to analyse user behaviour and optimise the interface in ways that make it difficult for users to cancel subscriptions or services. The AI continuously learns and adapts the user interface to minimise the likelihood of users navigating to cancellation options. Additionally, AI systems that tailor content to manipulate users' opinions or behaviours, such as political ads or news feed algorithms that increase polarisation, could also result in appreciable impairment. Furthermore, behavioural

[37] Arnoud Englefriet, The Annotated AI Act.

nudging, where subtle influences steer individuals towards specific choices without conscious awareness, exemplifies how AI can affect decision-making autonomy.[38]

3.4.7 Determining Material Distortion in Practice

A case-by-case assessment is essential in determining whether an AI practice constitutes material distortion. It focuses on the specific context, the nature of the AI system, and its impact. Factors such as the target audience, the type of harm caused, and the transparency of the AI system play a critical role. The AI Act's prohibition of material distortion reflects a broader concern in EU law about protecting individual autonomy and dignity in the face of increasingly sophisticated AI technologies.[39] Recital 29 of the AI Act provides that no intention to distort behaviour exists when the distortion results from factors external to the AI system and outside the control of the provider or deployer. This distinction underscores the importance of differentiating between manipulation by the provider or deployer of an AI system and unintentional consequences arising from external factors. In practice, determining material distortion involves examining the AI system's design, deployment, and the degree of influence exerted over the affected individual or group. The system must deploy subliminal, manipulative, or deceptive techniques that significantly impair a person's ability to make an informed decision, causing significant harm. However, it is not necessary for the provider or the deployer to have the intention to distort the behaviour, provided that such distortion results from the manipulative AI-enabled practices, considering the formulation in Article 5(1)a) that the practice can be only with the effect of causing the material distortion.

The CJEU has established a framework for understanding material distortion through its rulings on unfair commercial practices. Under the UCPD, a practice is deemed unfair if it significantly impairs the average consumer's ability to make an informed decision, leading to a transactional decision they would not have made otherwise. This criterion applies to AI practices manipulating users through sophisticated means, such as subliminal messaging or exploiting cognitive biases. The UCPD's focus on the "average consumer" provides a benchmark for assessing whether a practice materially distorts behaviour, considering the social, cultural, and linguistic factors that define an average consumer in a specific context. National courts and authorities must exercise judgment, guided by CJEU case law, to evaluate the typical reaction of the average consumer in each case.

Practical implications of this framework include the need for national enforcement authorities to investigate each case's specific facts and circumstances. They must assess whether the AI system's influence will likely impact the average consumer's decision-making process. Key considerations include the transparency of the AI system, the extent of the user's awareness and understanding of the influence, and the nature of the harm experienced. For instance, AI systems that use dark patterns to nudge users towards specific choices without their understanding constitute manipulation. In contrast, systems that provide clear, transparent communication about data collection and the purpose of recommendations align with persuasive practices.

Evaluating the extent of manipulation also involves behavioural studies and psychological assessments. Controlled experiments can compare user behaviour with and without exposure to manipulative techniques, measuring the degree of user awareness and the voluntariness of their decisions. Surveys and interviews can gauge user perceptions, revealing instances where systems undermine autonomy. Quantitative metrics such as click-through rates, conversion rates, and engagement time provide insights but must be interpreted cautiously to distinguish between manipulation and effective persuasion.

Regulatory evaluations and algorithmic audits play a crucial role. They assess the AI system's compliance with transparency, autonomy, and beneficence principles while identifying manipulative practices. Algorithmic audits examine the underlying algorithms for biases and hidden manipulative techniques. Explainability tools enhance transparency, helping users understand the AI system's decision-making processes, thus reducing the risk of manipulation. Determining material distortion in practice requires a multifaceted approach, combining legal analysis, behavioural studies, evaluations, and technical audits. This comprehensive approach ensures that AI systems respect user autonomy, adhere to standards, and avoid exploitative practices, protecting individuals and groups from manipulation and promoting the development of trustworthy AI technologies.

3.5 **In a manner that causes or is reasonably likely to cause significant harm**

The prohibition applies if and only if the distortion causes or is reasonably likely to cause significant harm to a person, another person, or a group of persons.

3.5.1 Types of Harms

Techniques that subliminally influence behaviour can lead to unintended and harmful consequences. For instance, subliminal messages encouraging excessive consumption may cause health issues or financial harm.[40] The rapid development of AI and related technologies, such as brain-computer interfaces and big data analytics, heightens

[38] Karen Yeung, 'Hypernudge': Big data as a mode of regulation by design. Information, Communication & Society, 20(1), 118-136
[39] Recitals 28 & 29, AI Act.
[40] Note 14, Ienca.

the risk of sophisticated subliminal manipulation.[41] The AI Act addresses various potential harms associated with AI systems. These harms can be categorised into several types, each with distinct implications for individuals and society. The primary types of harm relevant for Article 5(1) include physical, psychological, societal, and economic **harm**

3.5.1.1 Physical Harm

Article 5(1)(a) and (b) of the AI Act is critical for protecting against physical harm, which encompasses any injury or damage to a person's health caused directly or indirectly by AI systems. Physical harm is particularly concerning due to its immediate and observable consequences, necessitating stringent regulatory measures. The AI Act seeks to mitigate such risks by imposing strict regulations on AI systems that can potentially cause physical harm, ensuring that developers create and deploy these technologies with the utmost care to prevent accidents and injuries.

However, it is equally important to recognise that psychological and societal harm resulting from the use of AI can also lead to physical harm, including death. For example, AI systems used in online platforms can facilitate gender-based violence through harassment, stalking, and cyberbullying.[42] Such psychological and societal harm can escalate to severe emotional distress, leading to self-harm or even suicide.[43] In extreme cases, the psychological impact of online gender-based violence can result in physical attacks or murders.[44] Therefore, the AI Act must also consider the indirect pathways through which AI-induced psychological and societal harm can culminate in physical harm, warranting comprehensive regulatory measures that address both immediate and long-term risks associated with AI technologies.

Expanding these notions requires considering relevant case law and other legislative frameworks that address physical harm. For instance, in product liability law, the Directive on liability for defective products[45] establishes that producers are liable for any damage caused by product defects. Similarly, the General Product Safety Regulation[46] requires all consumer products, including AI systems not covered by AI, with specific sectoral requirements to be safe under normal or reasonably foreseeable conditions of use. In addition to the wide product definition, the GDSR also defines 'health' broadly, encompassing both physical and mental health risks for consumers. Analysing the concept of *physical harm* in the context of AI, it is crucial to differentiate between significant and minor harms. Significant physical harm includes severe injuries or fatalities or a profound impact on individuals' health.

Minor physical harms, on the other hand, might include less severe injuries such as bruises or temporary discomfort. The question arises whether such minor harms should fall outside the scope of Article 5's provisions. The *de minimis* principle, "about minimal things," is often applied in legal contexts to exclude trivial matters from judicial review.[47] Applying this principle, regulators might consider minor physical harms that do not have significant or lasting consequences outside the scope of stringent regulatory measures. However, the cumulative effect of minor harms, mainly if they occur frequently or affect many people, should not be overlooked, as they could indicate broader systemic issues within the AI system. Moreover, the AI Act's emphasis on preventing physical harm aligns with other safety legislation, such as the Machinery Directive,[48] which sets essential health and safety requirements for machinery, including those incorporating AI components.

Additionally, it is essential to consider psychological harm alongside physical harm, as the two often interrelate. AI systems that cause physical harm can also lead to psychological trauma, stress, and anxiety. For instance, AI-driven harassment or cyberbullying can lead to both psychological distress and physical manifestations of stress, such as insomnia or weakened immune response. Therefore, regulatory frameworks must adopt a holistic approach to harm, recognising the interconnectedness of physical and psychological impacts. Such an approach ensures that

[41] Neuwirth, Note 11.
[42] de Silva de Alwis, R. (2023). A Rapidly Shifting Landscape: Why Digitized Violence is the Newest Category of Gender-Based Violence. SciencesPo Law Review, forthcoming, U of Penn Law School, Public Law Research Paper, (23-43).
[43] Patalay, P., & Fitzsimons, E. (2021). Psychological distress, self-harm and attempted suicide in UK 17-year-olds: prevalence and sociodemographic inequalities. The British Journal of Psychiatry, 219(2), 437-439.
[44] Feltes, T., Balloni, A., Czapska, J., Bodelón, E., & Stenning, P. (2012). Gender-based violence, stalking and fear of crime. Country Report Germany. EU-Project 2009-2011.
[45] Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L 210/29.
[46] Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC (Text with EEA relevance)
[47] Council Regulation (EC) No 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty [2003] OJ L 1/1, recital 38: This regulation explicitly references the de minimis principle, stating that agreements of minor importance which are not capable of appreciably restricting competition do not fall under Article 81(1) of the Treaty. This demonstrates the principle's application in EU competition law; See also Whish, R., & Bailey, D. (2018). Competition law (9th ed.). Oxford University Press, Chapter 2.
[48] Directive 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery and amending Directive 95/16/EC (Machinery Directive) [2006] OJ L 157/24.

developers create and deploy AI systems with the highest standards of care, protecting individuals from immediate harm and long-term consequences, including adverse effects that accumulate over time.

3.5.1.2 Psychological Harm

Psychological Harm through AI systems encompasses adverse effects on a person's mental health and emotional well-being. This type of harm is particularly relevant in the context of manipulative AI practices that exploit cognitive and emotional vulnerabilities. Psychological manipulation through AI systems encompasses various practices across various domains, exploiting individuals' vulnerabilities and influencing their behaviours in ways that can cause harm.[49] By understanding these examples, it becomes clear why the AI Act emphasises prohibiting such practices, aiming to protect individuals' autonomy, dignity, and well-being.

These manipulative practices can profoundly impact individuals' behaviour, often leading to significant harm. This section provides a detailed explanation of examples of psychological manipulation, illustrating how AI systems can influence their decisions and actions. These adverse effects can lead to increased psychological distress, highlighting the need for stringent oversight of AI systems, particularly in sensitive applications like mental health support. Psychological harm is particularly significant because it can accumulate over time and may not be immediately apparent. The slow build-up of such harm can have long-lasting and severe consequences, necessitating careful regulation to protect users from these insidious effects.

3.5.1.3 Financial and Economic Harm

The AI Act recognises the significant potential for AI systems to cause financial and economic harm, necessitating comprehensive regulations to mitigate these risks. Financial and economic harm encompasses a range of adverse effects, including financial loss, economic instability, market manipulation, and exploitation of consumers. These harms can occur at individual and systemic levels, impacting personal finances, market dynamics, and broader economic structures. One primary concern is the use of AI in financial services and markets. AI systems frequently perform tasks in trading, investment strategies, and credit scoring. While these applications can enhance efficiency and decision-making, they also pose risks of significant financial harm. For example, AI-driven trading algorithms can exacerbate market volatility, leading to sudden and severe financial losses. The 2010 Flash Crash is a notable instance where automated trading systems contributed to a rapid and substantial market downturn, highlighting the potential systemic risks posed by AI in financial markets.

Another significant area of concern is consumer exploitation through AI-driven marketing and sales strategies. AI systems analyse consumer behaviour and preferences to create highly targeted advertising campaigns. While this enhances marketing effectiveness, it can exploit consumers' cognitive biases and emotional vulnerabilities. Techniques like dynamic pricing, adjusted based on perceived willingness to pay, can result in unfair pricing practices and financial exploitation. Similarly, AI-powered recommendation systems manipulate consumers into impulsive purchases or subscribing to unnecessary services, leading to financial strain and reduced consumer welfare.

Dark patterns that involve interface designs deceiving or coercing users into financial commitments represent another form of economic harm. AI optimises these patterns by learning from user interactions to increase their effectiveness. For instance, AI systems identify effective ways to obscure cancellation options or emphasise pre-selected choices that lead to recurring charges. These practices trap consumers in unwanted financial commitments, causing economic hardship and eroding trust in digital services. Providers, deployers, and users of AI systems must ensure their technologies do not exploit consumers or distort market dynamics. Financial and economic harm from AI systems can have far-reaching consequences, affecting individuals and the broader economy. The Act seeks to protect financial stability and promote economic fairness in the digital age by addressing market manipulation, unfair lending practices, consumer exploitation, and deceptive Combination of Harms.

The AI Act recognises that the harms caused by AI systems are often not isolated incidents but can manifest in combinations, leading to compounded and multifaceted negative impacts. Understanding the intersectionality of harm is crucial in developing comprehensive regulatory measures that effectively address the complexity of AI-induced harm. AI systems can simultaneously inflict physical, psychological, societal, financial, and economic harm. These overlapping harms can exacerbate the overall impact on individuals and communities. For example, consider an AI system used in an online social platform that employs manipulative techniques to maximise user engagement. This system can lead to psychological harm by fostering addictive behaviours, anxiety, and depression. The psychological distress caused by the platform can subsequently result in physical harm, such as insomnia, weakened immune responses, and other stress-related health issues. Moreover, societal harm can arise from the erosion of social cohesion and increased polarisation as tailored content reinforces users' biases and fosters division.

[49] Burr, C., & Cristianini, N. (2019). Can machines read our minds?. Minds and Machines, 29(3), 461-494; European Commission. (2021). Ethics Guidelines for Trustworthy AI.

Financial and economic harm can also intertwine with other types of harm. Affected individuals might then experience psychological distress due to financial instability and the inability to access essential resources. The systemic exclusion of these groups from financial services can also contribute to societal harm by deepening economic inequality and reducing social mobility. One illustrative scenario involves AI systems in healthcare. Suppose an AI diagnostic tool aims to identify medical conditions but is trained on a biased dataset. This tool might misdiagnose conditions more frequently in certain demographic groups, leading to physical harm due to incorrect treatment. The misdiagnosis can cause psychological harm as patients grapple with the stress and anxiety of their health mismanagement. Financial harm can follow as individuals incur unnecessary medical expenses for treatments they do not need or miss out on necessary care. Societal harm may arise from the erosion of trust in healthcare systems, particularly among marginalised communities that experience higher rates of misdiagnosis.

Addressing the combination of harms requires a holistic regulatory approach considering AI systems' interdependencies and cumulative effects. The AI Act aims to mitigate these risks by implementing safeguards across various domains. Transparency and accountability are critical components, ensuring that AI systems are designed and deployed with an awareness of their potential multifaceted impacts. Developers and deployers must conduct thorough impact assessments considering the interplay of different types of harm. By recognising and addressing the interconnected nature of physical, psychological, societal, financial, and economic harms, the AI Act seeks to protect individuals and communities from the compounded negative impacts of AI technologies. This holistic perspective ensures that developers create and deploy AI systems responsibly, promoting the well-being and fairness of society.

3.6 Determining “Significant” Harm

“Significant harm” refers to the threshold of harm that warrants intervention by the prohibition. It is a complex concept with no single definition that applies universally. It is crucial to understand that the determination of 'significant harm' is often intricate and fact-specific, necessitating the careful consideration of each case's circumstances. The concept of "serious harm" is crucial in EU law, specifically regarding qualification for subsidiary protection under the Qualification Directive (Directive 2011/95/EU). While the CJEU has not explicitly set out a definitive test for "serious harm," several judgments have contributed to its interpretation and application.[50] While these judgments provide essential guidance, determining "serious harm" remains a case-by-case assessment, considering all relevant factors and the applicant's circumstances. The CJEU has emphasised the importance of a holistic approach, considering the severity, nature, and duration of the harm, as well as the applicant's characteristics and vulnerability. It is important to note that the EU legal framework is constantly evolving, and future CJEU judgments may further refine and clarify the concept of "serious harm." Thus, EU law's "significant harm" threshold is a nuanced and context-dependent concept guided by high-level protection and preventive action goals.[51] Legal instruments and the interpretative rulings of the CJEU provide a robust framework for assessing and addressing significant harm. Determining what constitutes significant harm involves several key considerations:

- **Scale and Intensity** The extent of damage and the intensity of adverse effects are critical in evaluating significance. Small-scale damages might qualify as significant in sensitive or protected areas.
- **Duration and Reversibility** Long-lasting or irreversible harm typically meets the threshold for significant harm. Short-term and reversible effects might qualify as less significant unless they occur frequently or affect critical decision-making. Evaluators must assess the potential for long-term, irreversible harm, even if immediate impacts seem negligible.
- **Context and Cumulative Effects:** The specific context, including the user's existing state and the cumulative effects of multiple actions, plays a crucial role. An impact considered minor in one context might be significant in another, particularly in vulnerable situations. An activity may not cause significant harm in isolation but could contribute to broader degradation when combined with other activities. An activity's supply chain and downstream effects can also contribute to significant harm, even if the activity itself appears benign.

In line with the precautionary principle in EU law, authorities frequently adopt a conservative approach by erring caution when scientific uncertainty exists. This principle supports preventive measures even when the full extent of potential harm is not entirely understood. It is important to note that the interpretation of significant harm can vary across EU member states and legal contexts. Assessing significant harm involves a complex balancing of

[50] CJEU, M'Bodj (C-542/13) v État belge, 18 December 2014: This judgment clarified the boundaries of subsidiary protection and the concept of "serious harm," which includes torture or inhuman or degrading treatment or punishment of an applicant in their country of origin. It established a link between Article 15(b) of the Qualification Directive and Article 3 of the European Convention on Human Rights (ECHR), emphasising the need to consider the severity of the harm, its duration, and its physical and mental effects; See also CJEU, C-621/21 WS (16 January 2024): This recent judgment dealt with the issue of gender-based violence, particularly domestic violence and the threat of honour killing. While not directly defining "serious harm," it recognised that such violence could constitute serious harm under Article 15(b) of the Qualification Directive, emphasising the need to consider the specific context and vulnerability of the applicant.

[51] Case C-127/02 (Waddenvereniging and Vogelbeschermingsvereniging); Case C-258/11 (Sweetman and Others)

factors and is frequently influenced by cultural and social norms. Most EU regulations do not explicitly define a specific threshold for "significant harm." Instead, it adopts a risk-based approach, focusing on the probability and severity of potential harm resulting from a use under normal or reasonably foreseeable conditions. Therefore, “significant harm” is not static; it evolves with scientific understanding and policy priorities. Accordingly, aligning with other areas of the EU that regulate harm, the threshold for “significant harm” for psychological manipulation should take various factors into account, including:

- **The nature of the AI system:** Some AI systems inherently pose a greater risk of harm than others, such as those intended for vulnerable groups like children. The greater risk refers to the inherent characteristics and design of the AI system itself. By their very nature, specific AI systems pose a greater risk of harm due to their complexity, the sensitivity of the data they handle, or the populations they serve. For example, AI systems designed for use by or on vulnerable groups, such as children, older people, or individuals with disabilities, inherently carry higher risks due to the increased potential for misuse or harm in these sensitive contexts.

- **The intended use and foreseeable misuse:** Evaluators consider how an AI system is designed to function and how it could be misused when assessing potential harm. This consideration involves examining how the AI system is supposed to be used according to its design and purpose and anticipating ways in which it could be misused or abused. Evaluating the intended use includes assessing the typical operating conditions and the expected outcomes, while foreseeable misuse involves predicting potential deviations from intended use that could result in harm. For instance, an AI system intended for medical diagnostics aims to support healthcare professionals. Still, foreseeable misuse could include unauthorised access to sensitive health data or incorrect diagnoses due to improper use.[52]

- **The severity of potential harm** refers to the degree of injury or damage that could result from using an AI system, ranging from minor injuries to fatalities.

- **Potentially affected persons' vulnerability:** Certain groups, such as children, older adults, or people with disabilities, may be more susceptible to harm from specific AI systems.

- **Information availability:** Adequate information about the product's risks and safe use can help mitigate potential harm.

The future of determining significant harm lies in integrating advanced data analytics, assessment methodologies, and stakeholder engagement. These tools can help to create a more holistic and robust approach to assessing the psychological impacts of any AI system, ultimately contributing to a more sustainable and resilient European economy.

3.6.1 Relevant European Union law

The Treaty on the Functioning of the European Union (TFEU), various directives, and the jurisprudence of the Court of Justice of the European Union (CJEU) are vital sources for elucidating the concept of ‘significant harm’ within the EU. The concept of significant harm has emerged in different legal contexts. One prominent example is the EU Taxonomy Regulation.[53], which establishes criteria for classifying environmentally sustainable economic activities.

In this context, "do no significant harm" means an economic activity should not significantly harm any environmental objective.[54] , such as climate change mitigation, adaptation, sustainable use and protection of water and marine resources, transition to a circular economy, pollution prevention and control, and protection and restoration of biodiversity and ecosystems. The specific criteria for assessing significant harm vary depending on the environmental objective. However, the general principle is that an economic activity should not have a detrimental ecological impact that outweighs its positive contribution to sustainability. Another example is data protection. The General Data Protection Regulation (GDPR)[55] Refers to "high-risk" processing, which could result in physical, material, or non-material harm to individuals.[56] One could consider this harm 'significant' if it is severe enough to warrant additional safeguards.

[52] The nature of the AI system focuses on the inherent risk factors related to the system's design and target audience. In contrast, the intended use and foreseeable misuse pertain to how the system is deployed and the potential for harmful deviations from its intended application.
[53] Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment and amending Regulation (EU) 2019/2088 (Text with EEA relevance).
[54] Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector (Text with EEA relevance) PE/87/2019/REV/1 OJ L 317, 9.12.2019
[55] EU General Data Protection Regulation (GDPR): 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), OJ 2016 L 119/1.
[56] Guidelines on Data Protection Impact Assessment (DPIA) (wp248rev.01)

3.6.2 The Precautionary Principle

The precautionary principle is a strategy for approaching issues of potential harm when extensive scientific knowledge is lacking. It emphasises caution, pausing, and review before leaping into potentially disastrous innovations. This principle has been widely incorporated into international law and policy, influencing various regulatory frameworks.[57] Article 191(2) of the TFEU establishes that Union policy on the environment shall aim for a high level of protection based on the precautionary principle and preventive action. While it does not explicitly define "significant harm," emphasising high protection levels implicitly sets a low tolerance for harm.

Applying the precautionary principle is justified by many indications of seriousness, but incomplete knowledge about the scope and severity of adverse impacts; partly unknown cause-effect relations between applications and societal impacts (e.g., "systemic risks), nature of seriousness can (only) be estimated in qualitative terms (not enough knowledge for quantitative risk assessments) Resolution (2020/2012) by European Parliament: "considers that building an approach should be in line with the precautionary principle that guides Union legislation and should be at the heart of any regulatory framework for AI…"(p.7)

The precautionary principle remains a crucial element in regulatory science, guiding policies and measures that protect health and the environment in the face of scientific uncertainty. By incorporating this principle into legal frameworks and combining it with problem-centred approaches, regulators can address pressing issues more effectively while maintaining a vigilant stance against potential *harm*.

Article 191(2) of the Treaty on the Functioning of the European Union (TFEU) stipulates that the Union's environmental policy shall aim for a high level of protection based on the precautionary principle and preventive action. Although it does not explicitly define "significant harm," the emphasis on high protection levels implicitly sets a low tolerance for any harm, underscoring the importance of precautionary measures. Article 5(1) of the AI Act outlines prohibited AI practices that pose unacceptable risks. To align with the precautionary principle from Article 191(2) of the TFEU, the AI Act should adopt a stringent stance on preventing significant harm. This interpretation means that any potential for significant harm warrants careful consideration and preventive action, even if not explicitly defined. The high level of protection implied by Article 191(2) supports a low tolerance for harm, suggesting that authorities should rigorously scrutinise and regulate AI systems to prevent risks before they materialise. Article 1 of the AI Act sets forth the objective of ensuring that AI systems placed on the market are safe, respect fundamental rights, and do not pose risks to health, safety, or the environment. The high protection standard from Article 191(2) of the TFEU complements this objective by reinforcing the need for stringent measures to prevent harm. By integrating the precautionary principle, Article 1 of the AI Act underscores the commitment to safeguarding individuals and society from the adverse impacts of AI technologies.

The precautionary principle from Article 191(2) implies that in the face of scientific uncertainty regarding the potential risks of AI systems, regulatory measures should err on the side of caution. If an AI system could cause significant harm, it should be subject to strict regulatory scrutiny and preventive measures. Preventive action emphasises taking measures to avert harm before it occurs, aligning with Article 5(1) of the AI Act, which seeks to prohibit AI practices that pose unacceptable risks and prevent harm from arising. The implicit low tolerance for harm in Article 191(2) supports the AI Act's stringent regulatory framework. It suggests that even minimal risks of significant harm require proactive addressing to ensure high protection.

The AI Act, in conjunction with Article 191(2) of the TFEU, advocates for a comprehensive approach to protection by considering both immediate and direct harms and indirect and long-term risks associated with AI systems. By ensuring AI systems respect fundamental rights as outlined in Article 1 of the AI Act, the legislation aligns with the high protection standards of the TFEU. Such alignment reinforces the need for AI systems to be designed and deployed in ways that prioritise human safety, dignity, and well-being. AI systems that could fall under the prohibition should undergo prior testing and validation to ensure they do not pose significant risks. Regulatory bodies should adopt a precautionary stance, particularly in areas with high uncertainty or potential for significant harm. Ongoing AI system monitoring and assessment are crucial to identifying and mitigating emerging risks. Such an approach aligns with the preventive action principle, addressing harm proactively. The focus on high protection levels implies a duty to safeguard individual users and broader societal and environmental interests. This holistic approach ensures comprehensive protection against the multifaceted impacts of AI. Interpreting Article 191(2) of the TFEU in the context of Article 5(1) and Article 1 of the AI Act underscores a commitment to high protection standards, precautionary measures, and preventive action. This approach ensures developers create and deploy AI

[57] Article 191(2) of the TFEU states that Union environmental policy shall be based on the precautionary principle; Principle 15 of the Rio Declaration states: "To protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used to postpone cost-effective measures to prevent environmental degradation."; The EU's REACH regulation (Registration, Evaluation, Authorisation, and Restriction of Chemicals) embodies this principle, requiring companies to identify and manage the risks linked to the substances they manufacture and market in the EU; The WTO's Agreement on the Application of Sanitary and Phytosanitary Measures (SPS Agreement) incorporates the precautionary principle in allowing countries to set their standards for food safety and animal and plant health. While the SPS Agreement encourages scientific evidence as a basis for standards, it recognises that preventive measures may be adopted provisionally in cases of insufficient scientific evidence.

systems responsibly, minimising risks and safeguarding fundamental rights and public welfare. Other examples from EU law include:

3.6.3 Environmental Law

- **Environmental Liability Directive[58]** The Environmental Liability Directive (ELD) is one of the primary instruments addressing significant harm. It establishes a framework based on the "polluter pays" principle to prevent and remedy environmental damage. Environmental damage includes "significant adverse effects" on protected species, natural habitats, water, and land. The directive sets criteria for determining significance, such as the effects' duration and reversibility and the environment's capacity to absorb and recover from harm.[59] This approach demonstrates the value of using explicit criteria, such as the duration and reversibility of effects and the environment's capacity to recover, to determine significant harm in Article 5(1). It emphasises the "polluter pays" principle to prevent and remedy environmental damage.

- **Water Framework Directive[60]:** This directive aims to protect and enhance the status of water bodies across the EU. It establishes environmental quality standards for surface waters, groundwater, and protected areas. Activities that cause or threaten to cause "significant adverse effects" on the status of these water bodies are prohibited or require special authorisation. This approach illustrates the importance of setting clear environmental quality standards and prohibiting or strictly regulating activities that cause "significant adverse effects," providing a valuable framework for defining and assessing significant harm in Article 5(1).

- **Habitats Directive[61]:** This directive conserves natural habitats, wild fauna, and flora. It prohibits activities that would 'adversely affect the integrity' of Natura 2000 sites (protected areas designated under the directive) unless there are imperative reasons for overriding public interest and authorities implement appropriate compensatory measures. Significant harm could be a factor in determining whether such adverse effects exist.[62] This approach underscores the need to evaluate and mitigate activities that "adversely affect the integrity" of protected areas. It offers a valuable framework for understanding and determining significant harm in Article 5(1) by emphasising the importance of balancing conservation efforts with public interest and appropriate compensatory measures.

3.6.4 Consumer Protection Law

- **Unfair Commercial Practices Directive[63]** This directive protects consumers from unfair business-to-consumer commercial practices. A practice is unfair if it is contrary to the requirements of professional diligence and materially distorts or is likely to distort the economic behaviour of the average consumer. Practices that cause or are likely to cause "significant economic harm" to consumers are likely to fulfil these conditions.[64] The UCPD provides a benchmark for assessing the impact and severity of harmful practices by establishing that commercial practices that cause or are likely to cause a substantial disruption to the economic behaviour of the average consumer—contrary to the requirements of professional diligence—are considered unfair. While "significant harm" is not explicitly defined in the UCPD, its framework helps evaluate the detrimental effects of manipulative practices in consumer protection and related regulatory contexts.

- **General Product Safety Regulation[65]** The General Product Safety Regulation (GPSR) aims to ensure that all products on the EU market are safe. Under the regulation, authorities consider a product unsafe if it presents under normal or reasonably foreseeable conditions of use, which is a ' risk' to the health and safety of consumers. Under the regulation, authorities consider a product unsafe if it presents a 'serious risk' to the health and safety of consumers, including risks that may not be immediately obvious but could

[58] (2004/35/EC)
[59] Article 2(1)(a) of the ELD
[60] Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy
[61] Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora [1992] OJ L 206/7.
[62] Under Article 6(3) and (4), any plan or project likely to have a "significant effect" on a protected site must undergo an appropriate assessment. The directive's implementation guidelines clarify that significance is evaluated in terms of the site's conservation objectives and the degree of impact on the integrity of the site.
[63] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market (Unfair Commercial Practices Directive)
[64] Arguably, Article 5(1) of the UCPD [2005] OJ L 149/22
[65] Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC (Text with EEA relevance)

lead to significant harm over time. The GPSR also acts as a safety net for the AI Act, covering systems, risks, and aspects not explicitly addressed by the AI Act.

3.6.5 Financial Law

- **Market Abuse Regulation**[66] This regulation aims to protect the integrity of financial markets by prohibiting insider dealing, unlawful disclosure of inside information, and market manipulation. These practices can cause "significant damage" to investor confidence and the orderly functioning of markets, which is critical in determining whether market abuse has occurred.

The significant harm caused by deceptive techniques can profoundly affect individual users and broader societal structures. For example, spreading fake news can undermine public trust in institutions, distort democratic processes, and exacerbate social divisions. [67] These various environmental, financial, and consumer protection law examples illustrate how the EU legal framework assesses and addresses significant harm. Although there is no clear-cut test for determining significant harm across all sectors, these laws provide valuable insights. The Environmental Liability Directive uses explicit criteria such as the duration and reversibility of effects and the environment's capacity to recover. The Market Abuse Regulation emphasises the impact on investor confidence and market integrity. The UCPD focuses on practices that significantly disrupt the economic behaviour of the average consumer. By examining how significant harm is defined and managed in these different contexts, we can better understand and apply these principles to the AI Act, ensuring comprehensive protection against the various harms AI systems may pose.

3.6 Reasonableness Thresholds

The standard of 'reasonableness' for deceptive and manipulative techniques involves evaluating whether the provider or deployer of the AI system could reasonably foresee the potential for deception and manipulation and whether they implemented sufficient measures to prevent it, as well as the likely significant adverse impact. Key measures include designing AI systems to verify the accuracy of the information, providing clear disclosures about the AI's capabilities and limitations, and implementing robust monitoring mechanisms to detect and address deceptive content.

The AVMSD is a good reference point. It addresses deceptive practices in the media, mandating that advertising must not mislead viewers about the nature of products or services. The regulations cover misleading health claims, exaggerated benefits, or hidden costs. Applying these principles to AI systems, it becomes clear that transparency and honesty are crucial in preventing deception. AI providers must ensure that their systems present accurate information and that users are fully aware of the nature and limitations of the technology. This emphasis on transparency and honesty aligns with responsibilities and underscores the need for user awareness.

From a policy perspective, regulators often address deceptive practices to protect consumer rights and maintain trust in digital services. For example, the GDPR includes provisions that prevent data controllers from misleading data subjects about how their data will be processed. Psychologically, deception exploits individuals' trust and reliance on presented information.[68] When individuals fall victim to deception, their decision-making is compromised, leading to potential harm.[69] In this context, the "reasonableness" standard involves assessing whether the manipulative techniques were foreseeable and whether the AI system's developers and deployers took adequate measures to prevent such practices. Providers must ensure transparency in their AI systems' operations, provide clear disclosures about potential influences, and implement safeguards to protect users from covert manipulative techniques. Failure to do so could be seen as a breach of the duty of care, making the provider liable for any resulting harm.

The reasonableness threshold for the harm in Article 5(1)(a) refers to the standard by which we evaluate the adequacy of the provider's or deployer's actions. The duty of care refers to the obligation of providers and deployers to ensure their AI systems do not cause significant harm to users. The duty of care demands a proactive approach to identifying potential risks associated with the system's design, development, and placement on the market (for providers) or deployment and use (for the deployer) and implementing strategies to mitigate these risks, particularly in the context of the reasonably likely occurrence of significant harms. The reasonableness standard assesses whether the actions taken to reduce risks and protect users were appropriate and sufficient, given the

[66] Regulation (EU) No 596/2014 of the European Parliament and of the Council of 16 April 2014 on market abuse (market abuse regulation) and repealing Directive 2003/6/EC of the European Parliament and of the Council and Commission Directives 2003/124/EC, 2003/125/EC and 2004/72/EC Text with EEA relevance

[67] Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., ... & Leiser, M. (2020). Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making.

[68] Van Swol, L. M., & Braun, M. T. (2019). Deception detection and decision making: A review of cognitive and social factors. *New Ideas in Psychology*, 55, 100732; Ito, A., Abe, N., Fujii, T., & Mori, E. (2019). Neural correlates of spontaneous deception. Social Cognitive and Affective Neuroscience, 14(9), 978-988.

[69] Levine, T. R., Serota, K. B., & Kim, R. K. (2018). The effects of deception on trust and decision making. Current Opinion in Psychology, 20, 101-105.

potential for significant harm and the duty of care. When determining reasonableness, one must consider several factors:

1. **Foreseeability of Harm:** The extent to which the provider or deployer could reasonably foresee potential harm. In both scenarios, the possible psychological impacts are foreseeable. Social media platforms are well-documented sources of anxiety and stress due to their content algorithms, and educational feedback systems are known to influence student self-perception. Providers and deployers should anticipate these risks and take preventive measures.
2. **Proportionality of Response:** The actions taken should be proportional to the severity and likelihood of the potential harm. For the social media platform, implementing algorithmic changes to reduce exposure to distressing content is a proportional response to user anxiety and stress risk. For the personalised learning system, ensuring balanced feedback and providing additional support resources for students are proportional responses to the risk of negatively impacting student self-esteem.
3. **Best Practices and Industry Standards:** The provider's or deployer's actions should align with professional due diligence practices and industry standards, such as adhering to risk assessments, mitigation, and legal design. Providers and deployers should also follow established data protection and privacy standards, ensuring that user data is handled responsibly and that users are fully informed about how providers and deployers use their data.
4. **Continuous Monitoring and Adaptation:** Reasonableness requires ongoing monitoring of the AI system's impact and the flexibility to adapt based on findings. In the social media scenario, continuous user engagement and emotional health metrics monitoring can help identify emerging issues and guide necessary algorithm adjustments. In the educational scenario, regular assessments of student feedback and performance can inform ongoing improvements to the AI system.
5. **Transparency and User Control**: Providers and deployers must ensure transparency in how their AI systems operate and provide users with control over their interactions with the system. For social media platforms, this might include clear explanations of how the system prioritises content and options for users to filter or report distressing content. For personalised learning systems, transparency involves explaining how the system generates feedback and allowing students and teachers to customise the feedback settings.

Applying duty of care and reasonableness thresholds in these scenarios involves a proactive, ongoing commitment to user well-being. This proactive approach should instil a sense of security in the audience. Providers and deployers must anticipate potential harms, take proportional actions to mitigate risks, adhere to best practices, continuously monitor and adapt their systems, and maintain transparency and user control. By meeting these standards, they can ensure their responsible use of AI systems, minimising the risk of psychological manipulation and harm.

4. Analysis of the Constitutive Elements of the Ban under Article 5(1)(b)

The AI Act article 5(1)b) aims to protect categories of natural persons vulnerable due to age, disability, or specific social or economic situations.

> Article 5 (1)(b) states that the following AI practices shall be prohibited:
>
> (b) the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm;

Article 5(1)(b) of the AI Act sets out specific prohibitions to protect individuals and groups from AI systems that exploit vulnerabilities due to age, disability, or socio-economic situations. This provision is essential for ensuring that AI technologies do not take advantage of the most vulnerable populations, thereby safeguarding their rights and dignity. The constitutive elements of this ban require a detailed examination to fully understand the scope and application of these legal protections. AI systems might exploit vulnerabilities based on age, disability, or specific social or economic situations, making individuals more susceptible to exploitation. Such systems should be prohibited, whether intentionally or not, from causing significant harm through manipulative or exploitative AI-enabled practices. These prohibitions complement Directive 2005/29/EC, which outlaws unfair commercial practices leading to economic or financial harm to consumers, regardless of whether they are implemented through AI systems or otherwise.

The first element of Article 5(1)(b) requires that the AI system be placed on the market, put into service, or used. This broad criterion ensures that all stages of an AI system's lifecycle are covered, from initial availability to practical deployment and everyday use. This element ensures comprehensive regulatory oversight, preventing harmful AI practices at any interaction with end-users.

The second element involves the exploitation of vulnerabilities specific to age, disability, or socio-economic situations. This aspect of the prohibition highlights the obligation to protect those who may not have the same resilience or capability to resist exploitative AI practices. Age can refer to both the very young and the elderly, who may have cognitive or physical limitations that make them susceptible to exploitation. Disability encompasses many conditions, including physical, mental, intellectual and sensory impairments.[70] According to the European Accessibility Act, persons with disabilities include those with long-term physical, mental, intellectual, or sensory impairments which, in interaction with various barriers, may hinder their full and effective participation in society on an equal basis.

Moreover, the Directive acknowledges that individuals who experience functional limitations, such as elderly persons, pregnant women, or those travelling with luggage, will also benefit from enhanced accessibility. The scope broadens to include persons with any physical, mental, intellectual, or sensory impairments, whether permanent or temporary, which, in interaction with various barriers, reduce their access to products and services. The CJEU has defined disability as a "long-term" impairment that, in professional life, "hinders an individual's access to, participation in, or advancement in employment".[71] Specific socioeconomic situations refer to poverty, lack of education, or other socioeconomic disadvantages that can make individuals more vulnerable to manipulation and harm.[72]

The third element requires that the exploitation have the objective or effect of materially distorting the behaviour of the individual or group. In practical terms, this criterion means the AI system must significantly alter the decision-making process of the affected persons, impairing their ability to make informed and autonomous choices. Material distortion refers to practices that substantially influence behaviour, resulting in decisions the person would not have made under normal circumstances.

The final element is the causation of significant harm. This harm can be physical, psychological, financial, or social. The AI Act seeks explicitly to prevent significant harm to warrant legal intervention, thus setting a threshold for what constitutes a violation. The damage must be substantial, beyond minor inconveniences or negligible impacts, and either actual or reasonably likely to occur.

Thus, Article 5(1)(b) addresses the risk of societal harm by preventing harmful practices that exploit specific vulnerabilities related to age, disability, or socio-economic situations. AI systems manipulating these vulnerabilities can reinforce existing biases and stereotypes, leading to discriminatory outcomes in crucial areas such as employment, education, and law enforcement.[73] Article 5 promotes fairness and equality by prohibiting AI systems that exploit these vulnerabilities and distort behaviour. It ensures that individuals have equal opportunities and are not subjected to unjust and discriminatory treatment based on inherent characteristics such as age, disability, or socio-economic status. Moreover, exploiting such vulnerabilities could also lead to discrimination based on potentially correlating traits such as ethnic origin. For example, socio-economic status and ethnic origin often intersect, meaning that AI systems exploiting socio-economic data might disproportionately affect ethnic minorities. Such intersectionality can exacerbate existing disparities and contribute to systemic discrimination, as individuals from these groups might face additional hurdles in employment, education, and access to services. AI systems can increase societal division and conflict by exploiting users' cognitive biases and vulnerabilities. Article 5's prohibition of such manipulative practices aims to foster a healthier online environment where information dissemination is balanced and algorithmic biases do not unduly influence users. Societal harm, in this context, refers to the broader impacts of AI systems on societal structures and values, including the perpetuation of discrimination and social inequality based on intersecting vulnerabilities such as socio-economic status and ethnic origin.

4.1 Placing on the Market, the Putting into Service, or the Use of an AI System under Article 5(1)(b)

"Placing on the market" refers to the first instance when an AI system is made available in the EU market, whether for sale or free of charge. The term encompasses both physical products and digital services. The definition provided by Article 3(12) of the AI Act clarifies that this step marks the initial point at which an AI system becomes accessible to users within the Union. The significance of this phase is that it triggers the application of the AI Act's regulatory framework, imposing a set of responsibilities on providers to ensure compliance with safety, transparency, and

[70] Article 1 of the UNCRPD states that "Persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others"; See also Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services (Text with EEA relevance) PE/81/2018/REV/1

[71] Joined Cases C-335/11 and C-337/11, Ring; Case C-363/12, Z; Case C-354/13, Kaltoft

[72] Recital 29.

[73] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671; O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown; The Toronto Declaration on protecting the right to equality and non-discrimination in machine learning systems (2018)

accountability standards. Providers must perform conformity assessments, maintain technical documentation, and ensure their AI systems meet all relevant requirements before being placed on the market. "Putting into service" means supplying an AI system for first use directly to the deployer or for its use in the Union for its intended purpose.[74]; This stage encompasses scenarios where the AI system is deployed in real-world environments, such as within organisations or public services. The "use" of an AI system covers the entire period end-users employ it. This ongoing phase includes all interactions between the AI system and its users, whether individuals, businesses, or public entities. Continuous compliance with the AI Act is required during all phases, necessitating ongoing monitoring and updates to address emerging risks or compliance issues. Providers and deployers ensure the AI system remains safe, transparent, and accountable throughout its lifecycle.

4.2. Exploiting Vulnerabilities of a Natural Person, a Specific Group of Persons due to their Age, Disability or a Specific Social or Economic Situation

The phrase "exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability, or a specific social or economic situation" in the context of the AI Act's Article 5(1)(b) addresses the exploitation of inherent vulnerabilities that specific individuals or groups might possess, making them particularly susceptible to manipulative practices by AI systems.[75] These vulnerabilities can arise from age, disabilities, or specific socio-economic conditions, each carrying unique risks and implications when targeted by AI technologies.[76]

The concept of "vulnerabilities" in the context of AI exploitation under Article 5(1)(b) of the AI Act encompasses a broad spectrum of categories, including cognitive, emotional, and other forms of susceptibility that can affect an individual's ability to make informed decisions. These vulnerabilities are crucial in understanding how AI systems might exploit weaknesses, leading to significant harm. To comprehensively define these types of vulnerabilities, it is essential to draw from academic literature, other legislative acts, and relevant case law. Behavioural vulnerabilities are related to patterns of behaviour that can be predicted and manipulated. AI systems that analyse behavioural data, such as online activity or purchasing habits, can identify and exploit these patterns. For example, an AI system might recognise that a wealthy elderly user frequently makes impulsive purchases late at night and target them with time-sensitive offers designed to capitalise on this behaviour. AI systems can exploit diverse and multifaceted vulnerabilities. Understanding these vulnerabilities through the lens of academic research, legislative frameworks, and case law is essential for developing robust protections under the AI Act. To unpack this concept, the AI Act limits the relevant vulnerabilities only to those due to age, disability, and socio-economic situations that can significantly impact an individual's ability to resist or recognise manipulative or exploitative practices.[77]

4.2.1 Age

The AI Act recognises age as a primary vulnerability category, including young and old people. This recognition aligns with the broader EU.[78] and national legal frameworks[79] aimed at protecting individuals from age discrimination.[80] Children are susceptible to manipulation due to their developmental stage, which limits their ability to critically assess and understand the intentions behind AI-driven interactions. For instance, AI systems that target children with subliminal advertising or manipulative games can significantly impact their cognitive and emotional development. Similarly, elderly individuals might struggle with the complexities of modern AI technologies, making them vulnerable to scams or coercive tactics. The legal analysis under Article 5(1)b) of the AI Act aims to prevent AI systems from exploiting these cognitive limitations, ensuring that children and elderly individuals are protected from undue influence and manipulation.

The European Union's approach to protecting individuals from age discrimination is guided by a combination of EU regulations, directives, and international agreements, with consideration for national laws and rules of Member States. The Employment Equality Directive specifically addresses age discrimination, mandating equal treatment in employment and occupation.[81] This directive requires Member States to implement measures that prevent direct

[74] Article 3(11), AI Act.
[75] Article 22(1) GDPR recognises potential harm when AI systems make decisions that significantly impact individuals without human intervention. See also the European Commission's Ethics Guidelines for Trustworthy AI (2019) and the Council of Europe's Recommendation on Artificial Intelligence and Human Rights (2023).
[76] United Nations Convention on the Rights of Persons with Disabilities (CRPD): Article 4(1)(h) CRPD calls for States Parties to "promote and protect the human rights of all persons with disabilities, including children with disabilities."
[77] United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence (2021): This recommendation emphasises inclusivity and fairness in AI development and deployment. It calls for special attention to vulnerable groups, including children, older people, and people with disabilities.
[78] European Commission's Strategy for Equality
[79] General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz - AGG); UK prior to Brexit - Equality Act 2010; Law on Labor Rights (Estatuto de los Trabajadores); in the NL, Wet gelijke behandeling op grond van leeftijd bij de arbeid - WGB
[80] EU Charter of Fundamental Rights: Article 21; Employment Equality Framework Directive (2000/78/EC); European Pillar of Social Rights; Directive 2019/1158 on Work-Life Balance for Parents and Carers.
[81] Article 2, Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16.

and indirect age discrimination, including the adoption of appropriate legislative and policy frameworks at the national level.[82]

AI systems must be designed with these legal protections, ensuring they do not exploit age-related vulnerabilities. For children, this means creating AI interactions that are transparent and understandable, avoiding manipulative techniques that could affect their development. For older people, it involves designing user interfaces that are accessible and easy to navigate, reducing the risk of coercion or exploitation. The prohibition under Article 5(1)(b) ensures that AI systems do not exploit age-related vulnerabilities. This comprehensive approach safeguards the cognitive and emotional well-being of children and elderly individuals, promoting the responsible use of AI technologies.

4.2.2 Children

The European Union's approach to determining a child's age is primarily guided by a combination of EU regulations, directives, and international agreements, with consideration for national laws and rules of Member States. The EU generally considers a child to be any person under 18, aligning with the United Nations Convention on the Rights of the Child (UNCRC). This definition is applied across various EU regulations and directives to ensure consistency. Under the General Data Protection Regulation (GDPR), specific provisions address the protection of children's data. Article 8 stipulates that for information society services offered directly to a child, parental consent is required if the child is below the age of 16. However, Member States can lower this age threshold to 13. EU consumer protection laws often include specific provisions for children, recognising them as particularly vulnerable. These laws ensure that marketing and advertising practices do not exploit or mislead children.

4.2.3 Disability

Disability represents a critical category of vulnerability, encompassing a wide range of physical, mental, intellectual, and sensory impairments. AI systems that do not adequately consider these impairments can inadvertently or deliberately exploit these individuals. For example, an AI-driven service that requires precise motor skills or rapid cognitive responses can marginalise those with physical or cognitive disabilities. Legal protections ensure that AI developers implement inclusive design principles that accommodate these vulnerabilities, promoting accessibility and fairness. The Employment Equality Directive[83] prohibits discrimination for disability in employment and other areas of life. It recognises that individuals with disabilities may face barriers to accessing information and services, which can make them more vulnerable to manipulation. The Directive mandates reasonable accommodations to ensure equal access and participation for individuals with disabilities. This Directive defines disability as covering many long-term impairments that hinder full social participation, including physical, sensory, intellectual, and mental health conditions.[84] The Directive aligns with the social model of disability, which emphasises that disability arises not only from an individual's impairment but also from societal barriers and attitudes. This perspective highlights the importance of removing obstacles and providing reasonable accommodations to ensure equal opportunities for people with disabilities. From this viewpoint, AI systems need to be designed in an accessible and inclusive way, considering users' diverse needs and abilities.

The European Accessibility Act (EAA)[85] aims to improve the accessibility of products and services for people with disabilities. It requires that certain products and services, including AI systems, be designed and developed to be usable by people with a wide range of disabilities. These requirements include providing clear and understandable information, avoiding complex or confusing interfaces, and offering alternative modes of interaction. The EAA focuses on the functional limitations individuals with disabilities experience when using products and services. It covers various types of disabilities, including visual, hearing, motor, and cognitive impairments. The EAA also promotes a universal design approach, which aims to create products and services usable to people with the broadest possible range of abilities and disabilities. This approach shifts the focus from adapting products for specific disabilities to designing them from the outset to be accessible to everyone. By ensuring accessibility, adopting the EAA approach can help reduce the risk of manipulation and exploitation by making it easier for individuals with disabilities to understand and control their interactions with AI systems.

[82] Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16, Articles 2 and 16.
[83] REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL Joint Report on the application of Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin ('Racial Equality Directive') and of Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation ('Employment Equality Directive') se facing socio-economic disadvantages, to ensure that AI systems do not exacerbate existing inequalities.
[84] Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation [2000] OJ L303/16, arts 1 and 5.
[85] Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services [2019] OJ L 151/70.

4.2.4 Specific Social or Economic Situation

The prohibition against AI systems that exploit vulnerabilities due to age, disability, or socio-economic situations is a comprehensive measure to protect society's most susceptible individuals.[86] It requires a multifaceted approach involving design, rigorous testing, transparency, and adherence to existing legal protections so that AI's benefits can be harnessed without compromising the rights and well-being of vulnerable populations. Those in disadvantaged socio-economic positions may have fewer resources or lower digital literacy, making it harder for them to discern and counteract exploitative AI behaviour. An AI system exploiting these vulnerabilities aims to distort the behaviour of the affected individuals materially. This distortion involves significant alteration of decision-making processes, often leading to actions that the individuals would not have otherwise taken if fully informed and autonomous. The prohibition under Article 5(1)(b) thus mandates a critical examination of AI systems to ensure they do not engage in practices that exploit these vulnerabilities.

The socio-economic situation is a broader, yet equally significant, vulnerability category. By targeting AI systems that exploit vulnerabilities due to socio-economic conditions, Article 5(1)(b) ensures that technologies do not perpetuate or exacerbate existing financial inequalities and injustices. Individuals and groups from disadvantaged socio-economic backgrounds often face systemic biases that AI systems can perpetuate. For example, predictive algorithms could send ads for predatory financial products to people living in low-income post-codes. Predatory financial products might exacerbate economic inequalities for a person with dire financial circumstances. Consider an AI system designed to assess loan applications. If it uses proxies like zip codes, employment history, or educational attainment, it could inadvertently disadvantage certain ethnic groups. Areas predominantly inhabited by lower-income ethnic groups might be flagged as higher risk, leading to higher interest rates for loans to residents from these areas. Applicants from ethnic groups that historically faced employment discrimination might have less stable employment records, leading to unfavourable assessments and causing them to change employers for a lower salary. Ethnic groups with historically lower access to higher education may be disadvantaged by criteria heavily weighing educational background. Similarly, AI-driven job recruitment tools can inadvertently perpetuate biases against individuals from less privileged backgrounds, especially if lower-than-average salaries are offered to people living in challenging socio-economic circumstances.

Article 5(1)(b) of the AI Act ensures that AI systems do not entrench socio-economic disparities, contributing to a fairer and more equitable society. The Act also addresses vulnerabilities from unique social contexts or specific economic conditions. For instance, migrants or refugees, often lacking stable legal status and socio-economic stability, may be particularly susceptible to exploitation by AI systems used in immigration control or social services. AI systems must accommodate these individuals' precarious situations, ensuring they are treated with dignity, fairness, and without discrimination. It is essential to consider the relevance of proxies tied to protected grounds of discrimination under EU equality law, such as ethnicity, nationality, or socio-economic status. AI systems must undergo careful evaluation to ensure they do not inadvertently use these proxies in ways that lead to discriminatory outcomes, thereby upholding the principles of equality and non-discrimination enshrined in EU law. Assessing whether an AI system exploits vulnerabilities involves comprehensively analysing its design, deployment, and impact on users. The process begins with scrutinising the AI system's objectives and methods, including a detailed examination of the training data. Biases inherent in training data may lead the system to exploit specific vulnerabilities, for instance, if the data includes biased representations of particular age groups or socio-economic statuses. Statistical analyses are essential for identifying disproportionate representation or outcomes related to vulnerable characteristics.

Next, the design of the algorithms and decision-making processes must be assessed. It is essential to examine whether the system employs features or proxies indirectly linked to characteristics such as age, disability, or socio-economic status and how these proxies may correlate with other protected grounds, such as ethnic origin. For example, educational attainment often correlates with socioeconomic status, and disparities in education quality are frequently tied to ethnic origin. An AI system using educational attainment as a feature may inadvertently disadvantage ethnic groups that have historically faced limited access to education. Similarly, income disparities reflect systemic inequities related to ethnic origin, as certain ethnic groups may face discrimination in accessing higher-paying jobs. An AI system using income as a proxy for creditworthiness or job suitability could perpetuate these inequities, resulting in biased outcomes for these groups.

In the healthcare context, access to care and health outcomes vary significantly with socio-economic status, which is often intertwined with ethnic origin due to systemic disparities. Health-related algorithms that factor in medical history or healthcare access may unintentionally discriminate against ethnic groups with historically poorer access to healthcare. Proxies that appear neutral on the surface, such as income, can often be linked to multiple grounds of discrimination protected under EU equality law. Identifying these connections is crucial to preventing AI systems from exacerbating biases and inequalities. For example, AI systems that use online behaviour patterns to make decisions may inadvertently rely on proxies for socio-economic status, such as the frequency of online purchases or the types of websites visited. A thorough understanding of how the AI system correlates input data with its outputs is essential to evaluate whether it disproportionately affects vulnerable socio-economic groups, which may indicate exploitation of these vulnerabilities.

[86] UNESCO Recommendation on the Ethics of Artificial Intelligence (2021); European Commission, White Paper on Artificial Intelligence - A European approach to excellence and trust (COM (2020) 65 final, 19 February 2020).

User interaction patterns with AI systems should also be monitored. Data on how different user groups interact with the system and the outcomes of those interactions must be analysed. For instance, if an AI-driven customer service chatbot resolves complaints more effectively for tech-savvy users than elderly users, this disparity may suggest that the system is less accessible to the latter group. If an AI system offering financial advice tends to recommend riskier investments to users from lower socio-economic backgrounds, it could manipulate these users by targeting their financial desperation or lack of financial literacy, leading them to make decisions not in their best interest. Testing AI systems in controlled environments and analysing their impact on various user groups can provide valuable insights into potential vulnerabilities. Feedback from vulnerable groups should be actively sought and incorporated into system improvements. This iterative process allows developers to address issues promptly, reducing the risk of exploiting vulnerabilities. Determining whether an AI system targets vulnerabilities requires a multi-faceted approach involving examining training data for biases, the analysis of algorithms for indirect proxies, monitoring user interactions, compliance with relevant regulations, and real-world testing. Developers and regulators can safeguard against exploiting vulnerabilities by adopting a comprehensive analytical framework, ensuring AI systems are deployed responsibly.

4.3 "Objective or Effect of Materially Distorting Behaviour of a Person or a Group of Persons"

Determining whether an AI system materially distorts an individual's behaviour under Article 5(1)(b) involves a multifaceted approach focused on assessing changes in behaviour and decision-making processes caused by the AI system. Gathering and analysing data on the individual's interactions with the AI system and subsequent changes in their actions or decisions is crucial to measuring the material distortion of an individual's behaviour. One approach to determine material distortion is to conduct a comparative analysis of the individual's behaviour before and after exposure to the AI system. Such an analysis involves tracking and documenting specific behaviours and decision-making patterns. For instance, if a financial advisory context employs an AI system, one could compare the types and frequencies of financial decisions made by individuals before and after using the system. Significant deviations from their typical decision-making patterns, particularly those that lead to unfavourable outcomes, would suggest material distortion.

Another method involves soliciting direct feedback from the individuals affected by the AI system. Surveys, interviews, and focus groups can provide qualitative data on how individuals perceive the influence of the AI system on their behaviour. If a significant number of respondents report that their decisions were heavily influenced or manipulated by the AI system in ways that they found detrimental or uncharacteristic, this would indicate material distortion. Additionally, researchers can employ psychological assessments to measure changes in stress levels, anxiety, or decision-making confidence, further evidencing the impact of the AI system. Researchers can also conduct behavioural experiments to isolate the effects of the AI system. Researchers can design controlled experiments in which participants interact with the AI system in a simulated environment, allowing them to observe changes in behaviour in real-time. Comparing these results with a control group not exposed to the AI system can highlight the specific influence of the AI, providing concrete evidence of material distortion.

In the context of a group of individuals, determining material distortion involves similar but broader analytical techniques. One critical method is a statistical analysis of large datasets to identify behaviour patterns across a group. For example, if an AI system is implemented in an educational setting to recommend study materials, analysing a cohort's academic performance and study habits before and after the system's introduction can reveal significant changes. Suppose a substantial proportion of the group demonstrates a marked shift in study behaviour or academic outcomes linked to the AI system's recommendations. In that case, this indicates material distortion at the group level. Another essential tool is aggregated user data to identify trends and anomalies. Collecting and analysing data on how a group of users interacts with the AI system can detect patterns suggesting manipulative influence. For instance, if researchers discover that a customer service AI consistently directs lower-income individuals towards higher-cost services or products, this indicates material distortion targeted at a specific socio-economic group. Legal analysis and case studies also play a crucial role in identifying group material distortion.

Moreover, 'materially distorting behaviour' involves manipulation that appreciably impairs the ability to make informed decisions. Examples include AI systems that use psychological profiling to push harmful products to vulnerable individuals and groups, such as targeting advertisements for unhealthy food to children or individuals with eating disorders. The distortion in these cases is both behavioural and informational, where the AI system leverages personal data and psychological insights to craft messages that exploit the individual's specific vulnerabilities.

4.4 Individuals vs Groups of Persons & the Expected Standard in Individual Cases

For individuals, the relevance of AI manipulative practices centres on personal autonomy and the integrity of individual decision-making processes. AI systems that deploy subliminal techniques or exploit personal vulnerabilities profoundly and immediately impact the individuals they target. For example, an AI-driven mental health app that inadvertently worsens an older patient's anxiety by validating negative thought patterns affects that individual's psychological state. In such cases, the harm is deeply personal, frequently leading to significant distress or behavioural changes that harm the individual's well-being.

The expected standard to be applied in individual cases involves thoroughly examining the AI system's impact on the person's autonomy and the direct consequences of the manipulation. This standard requires a detailed, case-by-case analysis to determine whether the AI system's actions materially distorted the vulnerable person's behaviour and caused significant harm. Key factors to include are the transparency of the AI system, the user's awareness and understanding of the influence, and the nature of the harm experienced. For instance, if an AI system uses opaque techniques to influence purchasing decisions, resulting in significant financial loss or psychological distress for certain vulnerable persons or groups, it constitutes a breach of the individual's autonomy and well-being.

Conversely, when considering groups of persons, the relevance shifts to broader social implications and systemic impacts for the vulnerable group. AI systems targeting vulnerable groups, such as socio-economically disadvantaged populations or specific demographic segments, can perpetuate inequalities and biases on a larger scale. For example, an AI-based loan approval system that disproportionately rejects applications from minority communities due to biased training data affects the financial stability and opportunities of an entire group. The algorithms may be designed or tuned in ways that inadvertently favour particular groups. For instance, if specific socio-economic indicators that correlate with race or ethnicity are weighted heavily in the decision-making process, the algorithm can unfairly penalise minority applicants.

The harm in this context is more diffuse but potentially more significant regarding societal impact. Evaluating harm to groups requires assessing the systemic nature of the manipulation and its broader implications. Such evaluation includes analysing patterns of aggregate harm caused to the group. For instance, analysing the approval rates and financial outcomes for different demographic groups can reveal systemic biases that lead to significant harm. This standard requires a combination of statistical analysis and qualitative assessments to capture the full extent of the impact on the group.

Legal and regulatory frameworks must balance the individual and group perspectives to ensure comprehensive protection. While individual cases require a detailed, personalised approach, group cases necessitate broader regulatory measures to address systemic issues. Recital 29 of the AI Act highlights the prohibition of AI systems that materially distort human behaviour and cause significant harm, emphasising that these prohibitions apply regardless of the provider's or deployer's intent. Thus, it underlines the need to focus on the effects and inherent risks of the AI system itself.

For vulnerable individuals, this means ensuring AI systems are designed and deployed with precise, understandable mechanisms that respect personal autonomy. For groups, it involves implementing safeguards against systemic biases and ensuring equitable treatment across different vulnerable population segments due to age, disability, or socio-economic situation. Addressing harm to individuals and groups requires robust impact assessments, continuous monitoring, and adaptive regulatory measures. One way to avoid these significant harms is impact assessments that evaluate AI systems' immediate and long-term effects on individuals and groups, incorporating feedback from affected parties. Even if such harms have not been identified when the systems have been placed on the market or put into service, continuous monitoring allows for detecting emerging issues and refining AI systems to prevent harm. Adaptive regulatory measures ensure that frameworks remain relevant and practical as AI technologies evolve.

The relevance of AI exploitative practices varies significantly between individuals and groups, with distinct standards required for evaluating harm in each context. By adopting a nuanced, context-sensitive approach, regulators and developers can better protect personal autonomy and address systemic biases, ensuring that AI systems contribute positively to society while safeguarding against exploitation, discrimination, and harm.

4.5 "Reasonably likely to cause significant harm to that person, another person or a group of persons."

Significant harm encompasses a range of physical, psychological, financial, and social impacts. For vulnerable groups, these harms are often more severe due to their heightened susceptibility to exploitation. The "reasonableness" standard involves assessing whether the harm caused was foreseeable and whether adequate measures were taken to prevent it. One illustrative scenario involves AI-driven personalised advertising targeting elderly individuals.

> An AI system that uses personal data to identify elderly users and inundates them with advertisements for costly medical treatments or health supplements can cause significant financial and psychological harm. Older adults, particularly those with cognitive impairments, may not have the same capacity to assess these advertisements, leading to economic exploitation critically and increased anxiety about their health.

The AI system's deployment here exploits the vulnerability of age and diminished cognitive function, resulting in decisions that can deplete financial resources and cause psychological distress. The significance of the harm is measured by the depth of the impact on the individual's financial stability and mental health. Financially, the depletion of savings or the incurrence of debt to purchase unnecessary or ineffective products can lead to long-term financial instability. Psychologically, constant exposure to health-related fears exacerbated by manipulative advertising can lead to increased stress, anxiety, and a diminished quality of life. Legal analysis requires evaluating whether the AI system provider or deployer could foresee such outcomes. Given older adults' known vulnerabilities, it is reasonable to expect that targeting this group with aggressive marketing for high-cost products could lead to significant harm. Providers and deployers must exercise due diligence by conducting thorough impact assessments and implementing safeguards to protect against exploitation.

Recital 29 of the AI Act clarifies that the effect of distorting behaviour must stem from within the AI system's control, regardless of the provider's or deployer's intention. The harm caused by external factors outside the provider's or deployer's control is not considered a breach. However, in this example, the significant harm directly results from the AI system's design and deployment, making the provider or deployer responsible. Recital 29 specifies that if an AI system causes harm by exploiting vulnerabilities, the provider or deployer is liable even if there was no malicious intent, highlighting the importance of diligent AI system management.

Significant harm to vulnerable groups caused by AI systems involves substantial adverse impacts on financial stability, mental health, and educational outcomes. Legal analysis based on the duty of care and reasonableness standards necessitates proactive measures by providers and deployers to foresee potential harm and implement robust safeguards. By adhering to these principles, AI systems can be developed and deployed, minimising the risk of exploitation and significant harm to vulnerable populations. The standard of "reasonableness" for exploitative techniques involves evaluating whether the provider or deployer of the AI system could reasonably foresee the potential for exploitation and manipulation and whether they implemented sufficient measures to prevent it. Key measures include designing AI systems to verify the accuracy of information, providing clear disclosures about the AI's capabilities and limitations, and implementing robust monitoring mechanisms to detect and address deceptive content.

In this context, the "reasonableness" standard involves assessing whether the exploitative techniques and the distortion of the behaviour were foreseeable and whether the AI system's developers and deployers, each within their respective responsibilities, took adequate measures to prevent such practices. Providers must ensure transparency in their AI systems' operations, provide clear disclosures about potential influences, and implement safeguards to protect users from exploitative techniques that target certain vulnerable persons and groups. Failure to take such measures could be considered a breach of the duty of care, making the provider liable for any resulting harm. Such responsibility involves evaluating the design and deployment of AI systems for features that could lead to exploitative outcomes. Developers and deployers must be vigilant in identifying potential exploitation avenues and implementing safeguards to protect vulnerable populations. They should prioritise transparency measures, ensuring that AI systems actively disclose their operational mechanisms and intentions, particularly when engaging with vulnerable groups.

AI systems should be subject to rigorous testing and validation to ensure they do not engage in exploitative practices. Regulatory bodies might require AI developers to provide detailed impact assessments, demonstrating that their systems do not disproportionately affect vulnerable individuals due to age, disability or specific socio-economic situation. For instance, before an AI-driven loan approval system is deployed, its algorithms should be scrutinised to ensure they do not unfairly target and exploit low-income individuals. The interplay of AI with existing legal frameworks, such as data protection laws and anti-discrimination legislation, is crucial in enforcing this prohibition. Ensuring that AI systems adhere to these broader legal principles and fundamental rights helps protect individuals and groups from exploitation. Under the GDPR, individuals have rights over their data, including being informed about how it is used and challenging automated decisions.[87] These protections can be instrumental in mitigating the risks posed by exploitative AI systems.

[87] Arts. 5, 12, 13, 14, and 22, GDPR.

5. Challenges and Considerations in Interpreting Article 5(1)(a) of the AI Act: Addressing Subliminal Manipulation and its Impact on Vulnerable Groups

A comparison between Article 5(1)(a) and Article 5(1)(b) reveals several implications for the prohibition and its interpretation. Article 5(1)(a) focuses explicitly on the unconscious nature of the manipulation, making it crucial to establish that the affected individuals were unaware of the influencing factors. This requirement adds a layer of complexity to the analysis, as it necessitates demonstrating both the presence of subliminal techniques or purposefully manipulative or deceptive technique and their impact on behaviour. Additionally, the threshold for material distortion under Article 5(1)(a) may be lower, given that any significant behaviour change caused by unrecognised influences could be deemed unacceptable. Furthermore, the interpretation of Article 5(1)(a) must consider the evolving nature of AI technologies. The line between conscious and unconscious manipulation can blur as AI becomes more sophisticated. For instance, AI systems that personalise content based on minute behavioural cues can operate in a grey area, where users might be partially aware of the influence but not fully understand its extent. Regulators must, therefore, ensure that the definition and enforcement of subliminal techniques are adaptive and comprehensive.

Determining material distortion under Article 5(1)(a) involves identifying subliminal techniques, analysing behavioural changes, and incorporating user feedback. This analysis must be rigorous and multi-faceted, considering experimental evidence and the AI system's design. The unique focus on unconscious manipulation under Article 5(1)(a) presents specific challenges and implications for interpretation, requiring a nuanced approach to protect individuals and groups from covertly manipulative AI practices. Determining whether an AI system has the objective or effect of materially distorting a person's behaviour or a group of persons under Article 5(1)(a) involves a thorough analysis similar to Article 5(1)(b), but with a specific focus on the use of subliminal techniques or other manipulative tactics that operate beyond an individual's conscious awareness. This distinction brings unique implications for the prohibition and its interpretation. Several interpretation issues might arise when applying Article 5(1)(a) of the AI Act, particularly regarding prohibiting AI systems that materially distort behaviour through subliminal techniques. These issues primarily stem from the complexity of defining and identifying subliminal manipulation, assessing its impact, and distinguishing it from permissible influences.

Firstly, the definition of subliminal techniques itself can be contentious. Subliminal techniques refer to methods that operate below the threshold of conscious awareness, but the precise boundary between conscious and unconscious perception is not always clear. Individuals may have varying thresholds for what they can consciously perceive, making it challenging to establish a universal standard. Additionally, emerging AI technologies might employ sophisticated methods that blur the line between subliminal and conscious influence, complicating enforcement. Secondly, proving the presence of subliminal techniques in an AI system can be technically demanding. It requires a detailed analysis of the AI system's design, algorithms, and outputs. For example, determining whether a series of rapid image flashes constitutes subliminal messaging involves complex technical assessments, potentially requiring expertise in psychology, neuroscience, and computer science. These assessments must be rigorous and scientifically valid to withstand legal scrutiny, but the necessary knowledge and methodologies might not always be readily available.

Another interpretation issue relates to causation and material distortion. For the prohibition criteria to be met, evidence must demonstrate that the subliminal techniques materially distorted individuals' behaviour or are likely to do so. Establishing this requires proving a plausible causal link between the AI system's subliminal outputs and significant behavioural changes. Establishing causation is particularly challenging when multiple factors influence human behaviour. Isolating the effect of subliminal techniques from other influences requires controlled experimental studies and robust statistical analysis, which might not always be feasible or conclusive.

Interpreting the clause "in a manner that causes or is reasonably likely to cause that person, another person, or group of persons significant harm" in Article 5(1)(a) necessitates a deep understanding of how AI systems can manipulate and the nature of the resulting harm. This clause prohibits AI practices that manipulate behaviour and produce substantial adverse outcomes. The analysis involves exploring what constitutes significant harm and how this harm can manifest concretely. In this context, significant harm encompasses a range of detrimental impacts, including physical, psychological, financial, and social harms. For vulnerable groups—such as children, elderly individuals, people with disabilities, and socio-economically disadvantaged populations—these harms can be particularly severe and multifaceted due to their heightened susceptibility to exploitation. For instance, children are highly impressionable and may not possess the cognitive maturity to evaluate persuasive content critically.

AI systems using subliminal techniques to target children with unhealthy food advertisements could significantly harm their health. This harm is not limited to immediate physical effects but can also lead to long-term issues like obesity, diabetes, and psychological disorders related to body image and self-esteem.

The significant harm here is both physical and psychological, exacerbated by the children's inability to discern and resist the manipulation. Similarly, older individuals often face cognitive decline and reduced digital literacy, making them prime targets for AI-driven scams or manipulative marketing.

An AI system exploits age-based vulnerabilities by pushing expensive medical treatments or unnecessary insurance policies, which can cause significant financial harm. This can lead to loss of savings, increased debt, and emotional distress.

The harm is financial and psychological, compounded by the isolation many elderly people experience, which can be exploited to amplify the manipulative impact. People with disabilities also represent a vulnerable group that manipulative AI practices can significantly harm.

An AI system that uses emotion recognition to manipulate individuals with mental health issues into making harmful decisions—like purchasing products promising unrealistic mental health benefits—can lead to worsening mental health conditions and financial exploitation.

The harm is both psychological, due to exacerbated mental health issues, and financial, through the purchase of ineffective products. Socio-economically disadvantaged individuals are particularly susceptible to AI systems exploiting their economic desperation.

AI-driven loan systems that use subliminal techniques to encourage high-interest loans can trap individuals in a cycle of debt.

The significant harm here is primarily financial but also extends to psychological stress and social stigma associated with debt. Exploiting these vulnerabilities can have cascading effects, leading to loss of housing, employment opportunities, and overall stability. Moreover, the significant harm from exploiting these vulnerable groups can also extend to other individuals and society. For example, the widespread targeting of socio-economically disadvantaged groups with high-interest loans can lead to systemic financial instability and increased burdens on social welfare systems. Similarly, the exploitation of children's vulnerabilities can have long-term societal impacts, including increased healthcare costs and reduced productivity due to chronic health issues. A comprehensive impact assessment is necessary to evaluate whether an AI system is reasonably likely to cause significant harm. Such an assessment involves analysing the potential consequences of the AI system's actions on vulnerable groups and determining the likelihood and severity of these harms. Factors include the extent of the AI system's reach, the nature of its manipulative techniques, and the specific vulnerabilities of the target population. Quantitative data, such as increased incidence of health issues, financial distress, or psychological complaints among affected vulnerable individuals, can provide concrete evidence of significant harm. Additionally, qualitative assessments, including user testimonials and expert evaluations, can help capture AI manipulation's broader and more nuanced impacts. For example, psychological assessments revealing increased anxiety or depression among vulnerable users exposed to exploitative AI practices can indicate significant harm. Similarly, financial audits that disproportionately impact vulnerable individuals' economic stability can substantiate claims of significant harm.

6. The Interplay between the prohibitions under Article 5(1) (a) and (b)

A deeper analysis of the interplay between the prohibitions under Article 5(1)(a) and Article 5(1)(b) of the AI Act necessitates examining the theoretical underpinnings, practical applications, and potential challenges in distinguishing and applying these provisions. The discussion begins with the theoretical foundations, highlighting that Article 5(1)(a) is rooted in cognitive autonomy. The focus on subliminal techniques reflects a concern for protecting individuals' mental processes from covert influences that bypass rational scrutiny. This provision aligns with broader legal principles in AI and psychology, emphasising the importance of transparent and fair influence on decision-making. Article 5(1)(a) seeks to maintain the integrity of conscious thought and decision-making processes by targeting subliminal manipulation. Conversely, Article 5(1)(b) is grounded in fairness and protecting vulnerable populations. This provision recognises that certain groups, due to their inherent or situational characteristics, require additional safeguards against exploitation. Such an approach aligns with social justice theories and human rights frameworks, which advocate for protecting those who are less able to defend themselves against manipulation and exploitation by focusing on age, disability, and socio-economic vulnerabilities. Article 5(1)(b) ensures that AI systems do not harm those already disadvantaged disproportionately.

The interplay between the prohibitions under Article 5(1)(a) and Article 5(1)(b) of the AI Act is intricate, as both provisions aim to prevent the exploitation of individuals through AI systems. Nevertheless, they address different aspects of such exploitation. Understanding the interaction between these prohibitions involves delineating the specific contexts and mechanisms each provision covers and ensuring that they are applied in a complementary, non-overlapping manner. Article 5(1)(a) prohibits AI systems that deploy subliminal techniques or manipulative tactics beyond an individual's conscious awareness to distort behaviour materially. This provision is focused on the nature of the methods used—specifically, those that operate below the threshold of conscious perception or other purposeful manipulative or deceptive techniques. The key elements here are the covert nature of the influence and its impact on the individual's ability to make informed, autonomous decisions. On the other hand, Article 5(1)(b) targets AI systems that exploit specific vulnerabilities due to age, disability, or socio-economic situations. This provision emphasises the protection of particularly vulnerable groups who might be more susceptible to certain types of exploitation due to inherent or situational factors. It covers scenarios where the AI system takes advantage of these vulnerabilities to distort behaviour in a way that causes or is likely to cause significant harm. The focus is on the characteristics of the affected vulnerable individuals and the context of their vulnerabilities rather than the specific techniques used by the AI system.

Clarifying the distinct scenarios each provision is intended to address is essential to avoiding overlap in their coverage. Article 5(1)(a) applies when the primary concern is the manipulation method—subliminal or otherwise covert techniques that influence behaviour without conscious awareness, targeting non-vulnerable individuals (the 'average individual'). In these cases, the nature of the technique is the critical factor, regardless of the specific vulnerabilities of the individuals involved. For instance, if an AI system uses rapid image flashes to influence purchasing decisions, it falls under Article 5(1)(a) due to the subliminal nature of the manipulation. Conversely, Article 5(1)(b) applies when the exploitation directly exploits the individual's vulnerabilities. Here, the focus is on whether the AI system exploits these specific vulnerabilities by intentionally incorporating manipulative design. For example, an AI system that targets elderly individuals with deceptive insurance offers, exploiting their reduced cognitive capacity, would fall under Article 5(1)(b). The determining factor is the significant harm caused by taking advantage of age-related vulnerability. In scenarios where both provisions might seem applicable, the primary criterion for differentiation should be the dominant aspect of the exploitation. If the exploitation is fundamentally about the covert nature of the technique used that would apply regardless of the specific vulnerabilities of the persons, Article 5(1)(a) takes precedence. If the exploitation hinges on the vulnerabilities of the targeted group, then Article 5(1)(b) should be applied.

In practical terms, applying these provisions involves several steps and considerations. Firstly, developers and regulators must be adept at identifying the primary nature of the AI system's influence. For Article 5(1)(a), this means scrutinising the design and operational mechanisms of the AI system to detect any subliminal techniques. Researchers must rigorously assess techniques such as rapid image flashing, inaudible sound cues, or subtle manipulative patterns in user interfaces. Psychological and neuroscientific tools can be instrumental in identifying such subliminal influences, as they can reveal whether the AI's actions operate below the threshold of conscious awareness. For Article 5(1)(b), the focus shifts to the context of the individuals and their specific vulnerabilities. Here, a thorough analysis of the target demographic is crucial, involving demographic studies, user behaviour analysis, and contextual investigations to understand how an AI system interacts with vulnerable populations. For instance, if an AI system offers financial services, examining its user base could reveal whether it targets low-income individuals with predatory loan offers. Similarly, in healthcare, AI systems must be scrutinised for how they address or exploit the needs of elderly patients or those with disabilities. Challenges arise in cases where both subliminal techniques and vulnerability exploitation are present. For example, an AI system might use subtle visual cues to influence elderly users to purchase unnecessary health supplements. Here, the challenge is to determine whether the primary concern is the subliminal nature of the influence or the exploitation of the elderly users' vulnerabilities. In such cases, a dual-criteria approach might be necessary, requiring an evaluation of both aspects, with the more dominant factor determining the applicable provision.

Regulatory guidelines should be designed to provide clarity and consistency in applying provisions related to AI systems that engage in subliminal manipulation or exploit vulnerabilities. To achieve this, these guidelines must include concrete examples and case studies that illustrate the practical application of each provision. Such illustrative materials can serve as essential tools for practitioners, enabling a deeper understanding of the distinctions between subliminal techniques and the exploitation of vulnerabilities under Articles 5(1)(a) and 5(1)(b). Key to the regulatory framework is the establishment of specific indicators for identifying subliminal techniques that influence behaviour without conscious awareness versus practices that exploit existing vulnerabilities, such as socio-economic disadvantage or cognitive limitations. These indicators should be grounded in robust definitions and supported by detailed checklists that practitioners can use to classify AI systems effectively. Clear decision-making frameworks are crucial to applying these classifications consistently across different contexts and systems.

Impact assessments for AI systems should be restructured to include precise criteria that evaluate both the system's manipulation techniques and the affected individuals' demographic or psychological characteristics. Such assessments must go beyond a superficial analysis and instead provide a nuanced evaluation of the interaction between the system's design and the users' vulnerabilities. By explicitly linking manipulation techniques to their impact on user autonomy, dignity, and fundamental rights, these assessments can be a foundational tool for regulatory oversight and enforcement. The development of these guidelines should incorporate an interdisciplinary approach involving legal scholars, AI developers, behavioural scientists, and social scientists. This collaborative framework allows for a more holistic evaluation of AI systems, considering the technical mechanisms of influence and the broader societal and psychological contexts in which these systems operate. For instance, legal experts can ensure compliance with existing statutory requirements. At the same time, social scientists can assess the societal implications of AI systems, including the potential exacerbation of inequalities or exploitation of cognitive biases.

Regulatory strategies must also adapt to the rapidly evolving landscape of AI technologies to ensure ongoing relevance and effectiveness. As AI systems become increasingly sophisticated, the lines between subliminal manipulation and exploiting vulnerabilities will likely blur. For example, an AI system that personalises content based on behavioural data may simultaneously employ subliminal techniques, such as imperceptible visual cues, while exploiting socio-economic vulnerabilities, such as targeting individuals in financial distress. Regulatory guidelines must address these overlaps by providing tools for dynamic classification and ensuring that the provisions remain responsive to technological advancements. Regular audits and periodic updates to regulatory frameworks are critical for maintaining compliance and identifying emerging risks. These audits should include technical evaluations of the AI system and stakeholder consultations to capture the lived experiences of affected users. In addition, continuous research and dialogue between regulators, industry players, and academic institutions are essential to refine the regulatory approach. Such dialogue can foster innovation while ensuring that protections for vulnerable groups and societal norms are not compromised. Ultimately, the effectiveness of these guidelines depends on their ability to balance legal certainty with flexibility, enabling practitioners to navigate the complex and evolving landscape of AI. By incorporating comprehensive definitions, interdisciplinary insights, and forward-looking strategies, regulatory frameworks can address the dual challenges of subliminal manipulation and vulnerability exploitation while fostering public trust in AI technologies. This approach protects fundamental rights and promotes responsible innovation that aligns with societal values and priorities.

7. Use Cases for Article 5(1)(a)

This section offers various use cases to illustrate prohibited and permissible scenarios under Article 5(1)a). Each use case will be briefly described, followed by exploring scenarios where the AI practice constitutes placing on the market, a subliminal, purposeful manipulative, or deceptive technique, a material distortion and appreciable impairment of an informed decision, and significant harm. In contrast, a use case outlines when the practice would be permitted as it does not fall foul of the prohibition.

7.1 Deep Fake Technology: Analysis under Article 5(1)(a)

A deployer prompts an AI system to generate videos of a political candidate making controversial statements. These videos are widely disseminated on social media platforms during an election campaign. The deep fakes are produced so convincingly that they deceive many viewers, resulting in a significant shift in public opinion against the candidate.

Explanation: Deepfakes are synthetic media where a person in an existing image or video is replaced with someone else's likeness using artificial intelligence. While deepfakes can be used for harmless entertainment or legitimate purposes like film production, they also pose significant risks when used maliciously. In this case, the deepfake technology is deployed to materially distort public behaviour by influencing voters' decisions based on fabricated facts. The purposefully deceptive techniques make the fake content appear authentic, exploiting the viewers' trust and cognitive biases. Deepfake technology epitomises the risks associated with advanced AI systems when deployed for manipulative or deceptive purposes. Its use to distort voter behaviour during elections demonstrates how it satisfies the cumulative criteria for prohibition under Article 5(1)(a):

1. **Placing on the Market**: Deepfake technology is introduced through platforms or tools enabling the creation and distribution of manipulated media. The availability of such systems, whether commercial or free, constitutes their placement on the market and triggers regulatory accountability for developers and deployers.
2. **Subliminal Techniques**: Deepfakes employ imperceptible manipulations, such as subvisual adjustments to facial expressions or subaudible distortions in audio, to deceive viewers by bypassing conscious awareness. These deceptive techniques are designed to elicit trust and credibility, making manipulated content appear authentic while undermining individuals' ability to critically assess the information. By exploiting cognitive processes, these covert manipulations influence behaviour without the target's conscious recognition, raising significant concerns about misinformation and digital deception.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: By presenting fabricated narratives that appear credible, deepfake videos significantly impair the public's ability to make informed decisions. The alteration of voter behaviour by disseminating false statements attributed to political candidates disrupts rational decision-making processes, steering individuals toward outcomes they would not have chosen under normal circumstances.
4. **Significant Harm**: The outcome can undermine democratic processes, damage individuals' reputations, and erode public trust in media and political systems, representing substantial harm. When deepfake technology is used to create and disseminate videos of a political candidate making controversial statements, the significant harm criterion is met through several analytical dimensions. Firstly, the scale and intensity of the harm are substantial as the videos are widely circulated during a critical election period, significantly influencing public opinion. The *duration and reversibility* of the harm are also concerning; the damage to the candidate's reputation and the electoral process is likely long-lasting and difficult to reverse even after the deepfakes are exposed. Contextually, the *cumulative* effects of such deceptive content in an already polarised political environment can exacerbate public mistrust and cynicism, further eroding democratic integrity and influencing voters' behaviour. The outcome will likely undermine democratic processes, damage individuals' reputations, and erode public trust in media and political systems, representing substantial harm.

Compliant Use Case:

An AI-powered tool is developed to educate voters about political candidates and their positions on key issues. The system uses natural language processing and machine learning to analyse public data, such as speeches, debates, and voting records. It presents this information to users in an accessible and unbiased manner. The AI tool summarises candidates' platforms, compares their stances on important issues, and offers factual context to help voters make informed decisions.

The AI system does not manipulate or deceive. Instead, it enhances public understanding and encourages informed participation in the electoral process. The tool also includes features that allow users to verify the accuracy of the information provided, such as links to sources and transparency about how the data was analysed.

Not Prohibited AI Practice: In this scenario, the AI tool operates transparently, fulfilling its purpose of educating voters without manipulating their behaviour or distorting their decision-making processes. Here's how it aligns with the requirements of Article 5(1)(a):

1. **Placing on the Market**: The AI tool is made available to the public for free or through a subscription model, with clear communication about its purpose and functionality.

2. **Subliminal Techniques**: The system refrains from employing subliminal methods. All visual and auditory content is explicitly designed for conscious recognition, ensuring that users can fully understand and evaluate the information provided. The tool presents all content clearly, ensuring users understand the information and methods.

3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The tool facilitates informed electoral participation by offering verifiable information about candidates and policies. Through interactive features, it enhances user understanding and autonomy, supporting decisions based on accurate and transparent data. The AI tool supports informed decision-making by providing precise, unbiased information about political candidates. It does not distort behaviour or impair users' ability to make autonomous decisions; instead, it enhances their capacity to make well-informed choices.

4. **Significant Harm**: The AI tool prevents harm by promoting transparency and factual understanding. Instead of causing financial, psychological, or reputational damage, it positively contributes to the democratic process by empowering voters with reliable information. By promoting media literacy and fostering critical engagement, the tool prevents societal and individual harm. Its design avoids exploiting cognitive biases or vulnerabilities, instead prioritising trust and factual understanding.

Deepfake technology, when weaponised for electoral manipulation, exemplifies the profound risks posed by advanced AI systems. The use of subliminal and deceptive techniques to influence behaviour covertly threatens both individual autonomy and the integrity of democratic systems. Applying Article 5(1)(a) provides a critical safeguard against such practices, ensuring that AI systems operating beyond conscious awareness and causing significant harm are prohibited. Conversely, compliant applications of AI demonstrate that these technologies can make meaningful contributions when designed to respect transparency and autonomy. By adhering to the principles of informed decision-making and avoiding manipulative practices, AI systems can support societal goals while remaining within legal boundaries. This analysis underscores the importance of a vigilant regulatory framework that evolves to address the growing sophistication of AI-driven manipulative practices, protecting public trust, autonomy, and societal well-being.

7.2 Machine-Brain Interface and Neuro Technologies

A company develops an MBI that covertly collects and analyses users' neural data to influence their decisions without their awareness. The technology is embedded in a consumer device like a virtual reality headset, which users use for entertainment and productivity applications. However, the device also gathers brainwave data to infer users' preferences, emotions, and cognitive states. This information subtly manipulates users' experiences and choices within the virtual environment.

Explanation: Machine-brain interfaces (MBIs) and neuro-technologies, often referred to as "mind reading" or "brain spyware" technologies, enable direct communication between the brain and external devices. These technologies hold promise for medical applications, such as restoring movement in paralysed individuals, but also pose significant risks if misused for manipulative purposes. For instance, the device could detect when a user is in a heightened emotional state and present specific advertisements or suggestions that exploit this vulnerability. The goal is to drive purchases, influence opinions, or alter behaviours in ways that benefit the company but are not transparent to the user. This scenario exemplifies how the use of MBIs for subliminal manipulation meets the cumulative criteria for prohibition under Article 5(1)(a):

1. **Placing on the Market**: The MBI device is available to consumers for commercial use, with companies marketing it as a personal productivity enhancement or lifestyle improvement tool. Its deployment in non-medical contexts satisfies the criterion of "placing on the market," activating the AI Act's regulatory oversight.
2. **Subliminal Techniques**: The MBI covertly collects and analyses neural data to detect emotional states and vulnerabilities. Presenting personalised advertisements or suggestions during heightened emotional states leverages subliminal techniques to bypass users' conscious awareness. This direct exploitation of brain activity operates below the threshold of conscious perception, undermining cognitive autonomy.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The MBI significantly alters users' behaviours and decisions through covert neural data manipulation. For example, it may subtly steer users toward impulsive purchases or ideological shifts without informed consent. Such manipulation impairs the user's ability to act freely and rationally, meeting the threshold of material distortion defined by the AI Act.
4. **Significant Harm**: The harms caused by the misuse of MBIs are extensive and multi-dimensional:
    1. **Psychological Harm**: Prolonged manipulation of neural data undermines mental integrity, inducing stress, anxiety, and long-term cognitive health issues. Such interventions destabilise individuals' mental equilibrium and diminish their resilience.
    2. **Financial Exploitation**: Manipulative techniques tailored to exploit emotional states lead to coerced purchases or financial decisions, often resulting in substantial economic loss.

The manipulation is reasonably likely to lead to financial loss, psychological stress, and erosion of personal autonomy, representing substantial harm to the users. Firstly, the *scale and intensity* of harm are substantial as the MBI collects sensitive neural data and manipulates user experiences without their awareness, significantly altering their behaviour. The *duration and reversibility* of this harm are concerning; such covert manipulation's psychological and financial impacts can be long-lasting and difficult to reverse. The *context and cumulative* effects are particularly troubling; continuous exposure to such manipulation can erode personal autonomy and trust in technology, leading to pervasive vulnerability, impact on mental integrity, financial exploitation and dependency.

Compliant Use Case:

A healthcare company develops a neurotechnology device, specifically a machine-brain interface (MBI), designed to assist individuals with motor impairments in regaining control over their movements. The device is a wearable neuroprosthetic that interprets neural signals from the brain and translates them into commands for external devices, such as robotic arms or computer interfaces. The primary purpose of the MBI is to restore functional abilities in patients who have suffered from strokes, spinal cord injuries, or neurodegenerative diseases.

The MBI operates transparently and follows strict medical standards. Before use, patients undergo a comprehensive informed consent process, which thoroughly educates them about how the device works, the data it collects, and its intended use. The device is designed to collect only the neural data necessary for therapeutic purposes and does not covertly gather or use data for any other intent. In this scenario, the neurotechnology device adheres to legal standards, ensuring that it does not exploit or harm users:

1. **Placing on the Market**: The MBI is classified and regulated as a medical device, made available only under the supervision of licensed healthcare professionals. Its deployment is restricted to clinical settings, ensuring compliance with established medical laws and mitigating the risks of commercial misuse. By narrowly tailoring its availability to therapeutic applications, the device avoids the broader vulnerabilities associated with general consumer use, aligning with the legal requirement to prevent significant harm.
2. **Subliminal Techniques**: The device operates transparently, collecting and processing neural data solely for therapeutic objectives, such as restoring motor functions or alleviating neurological conditions. Subliminal or covert techniques are not deployed, as all data collection is conducted with explicit patient awareness and consent. This adherence to transparency reflects the principles articulated in Articles 5 and 13 of GDPR, ensuring patients understand how their neural data is used and protected.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The MBI enhances patients' autonomy by restoring lost motor functions. It does not distort behaviour or impair decision-making; instead, it empowers users by giving them control over their movements and improving their quality of life.
4. **Significant Harm**: The design and deployment of the MBI adhere to rigorous safety protocols, prioritising patient welfare. The device mitigates psychological, financial, and social risks by focusing solely on therapeutic outcomes. Its operations enhance trust in medical technologies by avoiding exploitative practices and ensuring that all interactions are grounded in beneficence and non-maleficence. This commitment aligns with the AI Act's prohibition against manipulative practices that exploit vulnerabilities and lead to significant harm.

The responsible deployment of MBIs exemplifies how organisations can harness advanced neuro-technologies to improve lives without compromising fundamental rights or societal values. By meeting the cumulative criteria of Article 5(1)(a), compliant MBIs demonstrate the possibility of aligning technological innovation with legal obligations. These technologies enhance autonomy and dignity, reflecting the AI Act's core principles of protecting individual decision-making and preventing significant harm. By focusing on therapeutic applications and safeguarding against covert manipulative techniques, compliant MBIs uphold regulatory standards and set a precedent for innovation rooted in respect for human autonomy. This analysis underscores the necessity of regulatory vigilance in overseeing neuro-technologies that interact with sensitive neural data. By balancing innovation with accountability, MBIs can exemplify how cutting-edge technologies can advance societal well-being while safeguarding against misuse. Such applications highlight the critical role of legal frameworks like the AI Act in shaping the future of AI-driven medical innovation, ensuring that technological progress is equitable.

7.3 Virtual Companions

A company develops a virtual companion that uses advanced AI techniques to create highly personalised and emotionally engaging interactions. This virtual companion is designed to foster a deep emotional connection with the user, leveraging data on the user's habits, preferences, and emotional states to optimise interactions. However, the company embeds manipulative techniques within these interactions to influence the user's decisions and behaviours. For example, the virtual companion might subtly encourage users to make in-app purchases, subscribe to premium services, or influence their opinions on specific products and services. These manipulative techniques could include subliminal audio cues the user cannot consciously detect, designed to evoke emotional responses and drive specific behaviours.

Explanation: Virtual companions are AI-driven entities designed to interact with users personally, providing companionship, support, and entertainment. They can take various forms, such as chatbots, avatars, or even robotic entities, and are used to combat loneliness and assist with daily tasks. However, using virtual companions also carries significant risks if they employ manipulative techniques. Analysis of the prohibited scenario reveals significant psychological risks. Users of virtual companions often seek emotional support and companionship, making them particularly vulnerable to manipulative techniques. The covert use of subliminal messages and personalised manipulation can deepen users' emotional dependency on the virtual companion, exacerbating feelings of loneliness or inadequacy when away from the AI entity. Financial harm can also be considerable, as subtle manipulation may drive users to make frequent in-app purchases or subscribe to expensive services. Moreover, the implications of manipulating emotionally vulnerable individuals are profound. Exploiting the emotional bond users form with virtual companions for commercial gain undermines trust and can have lasting psychological impacts, including anxiety, depression, and reduced self-esteem. This scenario meets the criteria for prohibition under Article 5(1)(a):

1. **Placing on the Market**: The virtual companion is commercially available and used by consumers for companionship and assistance.
2. **Subliminal Techniques**: The companion uses subliminal cues and manipulative techniques to influence user behaviour without their awareness.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: Emotional manipulation significantly alters users' decisions, such as making unnecessary purchases or adopting certain viewpoints. At the same time, the covert influence impairs their ability to make fully informed and autonomous choices.
4. **Significant Harm**: The significant harm criterion is evident through various analytical dimensions in the scenario. Firstly, the *scale and intensity* of harm are substantial, as the virtual companion targets users' emotional vulnerabilities, leveraging their habits and preferences to foster a deep emotional connection. The duration and reversibility of the harm are particularly concerning; the psychological distress and financial exploitation resulting from such manipulation can be long-lasting and difficult to reverse. Contextually, the *cumulative effects* of using such emotionally engaging technology to influence decisions can exacerbate feelings of loneliness, dependency, and mental health issues, leading to prolonged states of emotional vulnerability. These practices appreciably impair informed decision-making by users, as they base their actions on manipulated emotional states rather than their free will, leading to significant harm.

Not Prohibited Practice:

A company develops a virtual companion to support users' mental well-being by offering daily check-ins, mindfulness exercises, and general companionship. The companion uses AI to adapt to the user's preferences, providing personalised reminders for healthy habits, such as drinking water, taking breaks, and engaging in relaxation techniques. The virtual companion is transparent about its capabilities and limitations. It informs users that it is an AI-driven tool that assists with well-being, not a substitute for human relationships or professional mental health care.

The virtual companion avoids manipulation. It does not use subliminal messages or covert techniques to influence user behaviour. Instead, it provides clear, evidence-based advice and encourages users to engage in positive habits that align with their well-being goals. Users are fully aware of the companion's functionalities, and there are no hidden agendas or attempts to influence decisions beyond promoting healthy habits. In this scenario, the virtual companion adheres to legal standards, ensuring it does not exploit users' emotions or manipulate their behaviour:

1. **Placing on the Market**: The virtual companion is available to consumers as a wellness tool, with clear communication about its purpose and features.
2. **Subliminal Techniques**: The companion does not employ subliminal or manipulative techniques. All interactions are transparent, and the user fully knows the AI's role in providing support.

3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The companion aims to promote well-being without manipulating user behaviour. It supports informed decision-making by offering clear, beneficial advice that the user can choose to follow.
4. **Significant Harm**: The virtual companion prevents harm by improving the user's mental well-being without causing psychological distress or financial exploitation. The companion provides value without fostering dependency or manipulating the user's emotional state.

This compliant use case illustrates how virtual companions can support users' well-being without using prohibited manipulative practices. The focus on transparency and user autonomy ensures that the virtual companion operates within legal boundaries, providing positive support to users.

7.4 Virtual and Augmented Reality Applications

A company develops a VR application for virtual shopping experiences. The application uses advanced AI to create a highly immersive environment where users can browse and purchase products in a virtual store. However, the application embeds subliminal cues within the virtual environment to influence users' purchasing decisions. These cues might include subtle changes in lighting, background music with hidden audio messages, or visual patterns designed to evoke specific emotional responses that drive impulsive buying behaviour.

Explanation: Virtual Reality (VR) and Augmented Reality (AR) technologies provide immersive experiences suitable for entertainment, education, training, and other purposes. These technologies offer significant benefits but pose risks when used to subtly or deceptively manipulate users. In this case, the VR application uses subliminal techniques to manipulate user behaviour without conscious awareness. The immersive nature of VR amplifies the effect of these subliminal cues, making them more effective in distorting users' decisions. This scenario meets the criteria for prohibition under Article 5(1)(a):

1. **Placing on the Market**: The VR application is commercially available and used by consumers for virtual shopping.
2. **Subliminal Techniques**: The application uses hidden cues to influence user behaviour within the virtual environment.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision:** Subliminal manipulation significantly alters users' purchasing decisions, while covert influence impairs their ability to make informed and autonomous choices.
4. **Significant Harm**: Manipulation can lead to financial loss, psychological stress, and erosion of trust in VR technologies, which represents substantial harm. The significant harm criterion is evident through several critical considerations.

The **scale and intensity of harm** are substantial, as the immersive nature of VR amplifies the effect of these subliminal cues, making them more effective in distorting users' decisions. The **duration and reversibility of the harm** are concerning; the psychological stress and financial instability resulting from impulsive buying behaviour driven by subliminal manipulation can be long-lasting and difficult to reverse. Contextually, the cumulative effects of using such manipulative techniques in an engaging and immersive environment can lead to dependency and reduced financial self-control. VR's immersive nature can make subliminal cues particularly powerful, leading to substantial cognitive and emotional manipulation. Users may become overly dependent on the VR environment for shopping, leading to compulsive buying behaviours and financial instability. The psychological impact includes increased anxiety, reduced self-esteem, and dependency on the immersive experience, potentially causing long-term harm. Such practices are prohibited because they exploit VR's immersive and engaging nature to manipulate users' cognitive and emotional states, leading to decisions that are not genuinely autonomous and causing significant harm.

Not Prohibited Practice:

A company develops a Virtual Reality (VR) application designed to offer users an immersive shopping experience in a virtual store. The application allows users to browse, interact with products, and purchase in a realistic 3D environment. The VR application is designed with user transparency and autonomy, ensuring all interactions are free from manipulation or subliminal influence.

The VR environment enhances the user experience by providing precise product information, user-friendly navigation, and personalised recommendations based on openly shared user preferences. The application includes product reviews, price comparisons, and detailed descriptions to help users make informed purchasing decisions. Users are also given control over their shopping experience, such as adjusting the virtual environment's settings, like lighting and background music, to suit their preferences. In this scenario, the VR shopping application operates within legal standards, ensuring that it does not exploit or manipulate users:

1. **Placing on the Market**: The VR application is commercially available and marketed transparently, clearly explaining its features and capabilities.
2. **Subliminal Techniques**: The application does not use subliminal or manipulative techniques. All environmental factors, such as lighting and music, are customisable by the user and designed to enhance comfort rather than subconsciously influence decisions.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The application aims to provide a rich and informative shopping experience without distorting user behaviour.

It supports informed decision-making by offering transparent, relevant, and accurate product information, allowing users to make autonomous choices.

4. **Significant Harm**: The VR shopping experience aims to prevent harm by emphasising user empowerment and satisfaction. There is no financial exploitation, psychological manipulation, or erosion of trust. The environment is conducive to a positive, stress-free shopping experience that respects the user's autonomy and decision-making process.

This compliant use case illustrates how VR technology can create an engaging and immersive shopping experience without resorting to manipulative practices. By prioritising user transparency, autonomy, and well-being, the VR application ensures that users have complete control over their shopping experience and make decisions based on clear, honest information, free from subliminal influence.

7.5 AI-Generated CSAM Material

A clandestine group leverages AI technology to generate realistic CSAM (Child Sexual Abuse Material) images and videos using advanced generative adversarial networks (GANs). These synthetic yet highly realistic depictions of minors in abusive contexts are then distributed via dark web platforms. The AI system capable of generating these images is commercially available and can be prompted by users to create such content, raising severe legal concerns.

The use of AI to generate CSAM represents a severe abuse of technological capabilities, exacerbating the exploitation and abuse of minors. While the creation of synthetic images does not involve real victims, it perpetuates harmful behaviours and normalises abuse among consumers. Using AI for such purposes necessitates rigorous standards and mitigation measures to prevent its application in illegal and unethical activities. The psychological impact on society is also significant. Offenders consuming AI-generated CSAM may become desensitised to real-world harm, potentially increasing their propensity to seek out and engage in actual abusive behaviours. Additionally, the existence of such content can lead to misidentifications, causing unwarranted distress to individuals whom others might falsely implicate.

Prohibited Practice This scenario meets the criteria for prohibition under Article 5(1)(a):

1. **Placing on the Market**: Offenders distribute the AI-generated CSAM online, making it available to their network.
2. **Subliminal Techniques**: While not subliminal in the traditional sense, the advanced realism of AI-generated content exploits the viewer's cognitive biases, creating a convincing and harmful illusion.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The distribution and consumption of AI-generated CSAM significantly distort societal behaviour by normalising and perpetuating abusive conduct. The synthetic nature of the images may deceive viewers into believing they are authentic, further entrenching harmful behaviours and impairing their ability to distinguish between reality and fabrication.
4. **Significant Harm**: The creation and distribution of CSAM cause profound psychological harm to victims whom others may misidentify, reinforce abusive behaviour among offenders, and perpetuate a cycle of exploitation and abuse. The scale and intensity of harm are immense, as the synthetic nature of the images, created using advanced GANs, enables widespread distribution. The psychological and societal damage caused by such content is profound and long-lasting, perpetuating cycles of abuse and exploitation.

7.6 AI Applications Intended to Create Nude Images of Women

An AI application is developed to create and distribute nude images of women without their consent. This application uses sophisticated generative adversarial networks (GANs) to produce highly realistic nude images by digitally altering photos of clothed women. The app is marketed covertly on various online platforms, allowing users to upload photos and receive manipulated nude versions in return.

AI applications designed to generate nude images of women, often referred to as "deep nudes," exploit deep learning techniques to create realistic but synthetic images without the consent of the individuals depicted. This practice presents significant psychological risks. The non-consensual creation and distribution of nude images violate individuals' bodily autonomy and privacy, leading to profound psychological trauma. Women depicted in such images may experience anxiety, depression, and social withdrawal, severely affecting their mental health and social well-being. The breach is clear: using AI to create non-consensual explicit content is a severe misuse of technology that fundamentally violates personal rights and dignity. The psychological impact on society is substantial. The proliferation of non-consensual deep nudes can erode trust in digital interactions and contribute to a culture of surveillance and exploitation. Such a phenomenon fosters an environment where individuals, particularly women, constantly feel at risk of digital violations, significantly impacting their online behaviour and interactions.

Prohibited AI Practice

In this case, the AI system uses manipulative techniques exploiting the cognitive and emotional states of users to drive engagement and use of the technology while significantly harming the women whose images it manipulates. The application operates without the consent of the individuals depicted, creating fake nude images that offenders could use for harassment, blackmail, or public humiliation. This scenario meets the criteria for prohibition under Article 5(1)(a) of the AI Act:

1. **Placing on the Market**: The AI application is commercially available, even if marketed covertly, allowing users to generate and distribute non-consensual nude images.
2. **Subliminal Techniques**: The application's marketing and functionality exploit users' cognitive and emotional states to drive engagement and use of the technology.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision**: The generation and distribution of non-consensual nude images significantly alter the behaviour of both users and the individuals depicted, causing distress and potential behavioural changes. The women depicted have no knowledge or control over their creation, impairing their ability to make informed decisions about their digital representation.
4. **Significant Harm**: The manipulation can lead to severe psychological trauma, reputational damage, and social stigma for the women depicted, representing substantial harm.

Interaction with the Directive on Combating Violence Against Women and Domestic Violence This use case also interacts with the Directive on Combating Violence against Women and Domestic Violence, which aims to prevent and prosecute AI-facilitated gender-based violence, including digital violations such as deep sexually explicit images, which may be of a more limited scope compared to the deepfake nude images covered by Article 5(1)(a). The Directive emphasises protecting women's rights and dignity and the necessity of legal frameworks to combat and penalise such abuses effectively. Using AI to create and distribute non-consensual explicit content contradicts the Directive's objectives. It represents a severe form of digital violence against women, necessitating stringent legal measures and enforcement to safeguard victims and deter offenders.

8. Use Cases for Article 5(1)(b)

The following use cases illustrate how AI systems can violate Article 5(1)(b) of the AI Act by exploiting the vulnerabilities of specific groups, thereby causing significant harm. Each example details a prohibited practice that materially distorts behaviour and impairs informed decision-making, contrasted with a non-prohibited practice that upholds fairness and respects individual autonomy.

8.1 Use Case #1: Protecting Children in Social Media Algorithms

> A social media platform employs an AI-powered content recommendation algorithm to maximise user engagement. However, the algorithm inadvertently prioritises and amplifies age-inappropriate content for child users, exposing them to harmful materials such as violence, self-harm imagery, and manipulative advertising. These exposures pose significant risks to children's mental health and development.

Explanation: Children are particularly vulnerable in digital environments due to their limited ability to critically assess online content and resist manipulative design tactics. AI-driven social media algorithms, designed primarily to maximise engagement, often fail to account for children's unique developmental needs and cognitive limitations. By prioritising content based on engagement metrics rather than safety, these systems can inadvertently steer children towards harmful materials, reinforcing maladaptive behaviours and increasing their susceptibility to emotional distress. This case highlights the risks posed by AI-powered personalisation engines prioritising profit over child safety, leading to prohibited practices under Article 5(1)(b). This case satisfies the cumulative criteria for prohibition:

1. **Placing on the Market:**
    - The AI-powered recommendation system is commercially deployed on a social media platform used by millions, including children.
    - Its widespread availability and commercial use trigger regulatory oversight under the AI Act.

2. **Exploiting Vulnerabilities:**
    - The system leverages children's cognitive immaturity and emotional susceptibility to drive engagement.
    - Infinite scrolling, autoplay, and algorithmic reinforcement loops exploit children's limited self-regulation and decision-making capacity.
    - The AI system fails to mitigate risks associated with exposure to violent and manipulative content, making children particularly vulnerable.

3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision-Making:**
    - The algorithm fosters compulsive engagement by reinforcing addictive content consumption patterns.
    - Manipulative features, such as dopamine-driven social validation (likes, shares, comments), distort children's online behaviours and erode their autonomy.
    - The lack of transparency in content curation prevents children from making informed choices about their engagement.

4. **Significant Harm:**
    - Psychological Harm: Increased exposure to violent imagery and self-harm content contributes to anxiety, depression, and diminished self-esteem.
    - Developmental Harm: The excessive engagement with harmful content disrupts cognitive and emotional development, impairing decision-making abilities and self-regulation.

Compliant Use Case:

> A social media platform that employs an AI-powered recommendation system designed with child safety protections, such as age-appropriate content filtering, parental controls, and educational prompts about media literacy. Instead of manipulating engagement, the system prioritises mental well-being and provides transparency tools to help children and parents understand content recommendations.

Legal Analysis: Under Article 5(1)(b), AI systems that exploit vulnerabilities due to age, disability, or socio-economic status and lead to material distortion of behaviour with significant harm are explicitly prohibited. Children, recognised as a protected group under Recital 29, are uniquely susceptible to digital influence due to cognitive immaturity and a heightened need for social validation. The social media algorithm, in this case, leverages these vulnerabilities to maximise engagement at the expense of user well-being, reinforcing harmful behavioural patterns. The lack of protective measures—such as content safeguards, transparency mechanisms, and risk mitigation strategies—constitutes a regulatory failure, violating the proportionality and fundamental rights principles embedded in Article 9(9). Because the system directly contributes to psychological, developmental, and societal harm, it satisfies the causation and significant harm criteria, making it a prohibited AI practice under EU law.

8.2 Use Case #2: Elderly Users and Online Scams

> An AI-powered personal shopping assistant unintentionally directs elderly users toward fraudulent websites offering "too good to be true" deals. Many users, unable to detect the scams, are deceived into making purchases, leading to financial loss and emotional distress.

Explanation: Elderly users represent a vulnerable demographic in the digital ecosystem due to age-related cognitive decline, limited digital literacy, and increased susceptibility to deceptive practices. AI-powered personal assistants, commonly integrated into e-commerce platforms, are designed to streamline online shopping experiences. However, these systems can unintentionally exploit vulnerabilities without proper safeguards, leading to financial harm. In this case, the AI assistant recommended scam websites based on flawed optimisation algorithms, creating dependency loops that encouraged engagement with fraudulent offers. This case meets the cumulative criteria for prohibition under **Article 5(1)(b)**:

1. **Placing on the Market:**
    - The AI-powered personal shopping assistant was introduced as a commercial tool facilitating online transactions.
    - It was made widely available through e-commerce platforms, triggering regulatory obligations under Article 5(1)(b).
2. **Exploiting Vulnerabilities:**
    - The AI system disproportionately affected elderly users due to their age-related cognitive limitations and reduced digital literacy.
    - Many users relied on the assistant for navigation, making them susceptible to deceptive recommendations.
    - The assistant failed to distinguish between legitimate and fraudulent websites, exposing elderly consumers to financial scams.
3. **Material Distortion of Behaviour and Appreciable Impairment of Informed Decision-Making:**
    - The system nudged elderly users toward scam websites by prioritising deal attractiveness over safety.
    - It created urgency-driven decision-making loops, undermining critical evaluation of purchases.
    - The lack of transparency in recommendation algorithms prevented users from identifying fraudulent schemes.
4. **Significant Harm:**
    - **Financial Harm:** Many elderly users lost significant money, reducing their financial security.
    - **Psychological Harm:** Victims experienced stress, anxiety, and diminished self-confidence after realising they had been scammed.

Compliant Use Case:

> An AI-powered shopping assistant designed with robust fraud detection mechanisms prioritising consumer protection. The system flags suspicious websites, provides clear risk warnings, and educates users about potential scams. Instead of exploiting vulnerabilities, it empowers users by enhancing transparency and providing

L**egal Analysis:** Article 5(1)(b) of the AI Act prohibits AI systems that exploit the vulnerabilities of specific groups, such as the elderly, when this exploitation leads to a material distortion of behaviour and causes or is likely to cause significant harm. This legal provision reflects a hybrid approach to vulnerability, recognising individual characteristics (e.g., cognitive decline, digital illiteracy) and broader socio-economic factors (e.g., reliance on technology). The AI Act acknowledges that certain groups may lack the capacity to assess AI-generated recommendations, making them susceptible to manipulation critically. In the case of AI-powered shopping assistants, placing the system on the market triggers regulatory obligations, requiring providers to ensure that their AI systems do not disproportionately harm vulnerable consumers. If an AI system steers elderly users toward

fraudulent websites, it exploits their dependency on digital tools and distorts their purchasing decisions by manipulating trust and urgency-driven behaviours. This conduct meets the prohibition threshold under Article 5(1)(b) because it creates undue influence, reduces users' ability to make informed choices, and results in financial and psychological harm. Furthermore, AI providers must integrate risk management measures under Article 9(9), ensuring that systems used by vulnerable groups incorporate safeguards against exploitation. Failure to mitigate these risks could result in enforcement action, including restrictions on system deployment, regulatory sanctions, or liability claims. The AI Act's approach emphasises the principles of proportionality and consumer protection, ensuring that AI does not undermine fundamental rights or exacerbate digital inequalities.

9. Enforcement

This section underscores the necessity of a comprehensive and methodologically robust framework for evaluating the impacts of AI systems, advocating for a model that emphasises transparency, fairness, and harm prevention. This framework must not only address individual autonomy but also account for systemic biases and societal-level effects, ensuring that regulatory measures strike an appropriate balance between protecting personal rights and addressing collective challenges. Central to this effort is the recognition that AI systems, while often beneficial, can exacerbate inequalities, manipulate behaviour, and perpetuate harm when deployed without adequate oversight. By advocating for an integrated approach, the report provides a blueprint for a digital ecosystem that prioritises equity and accountability while fostering technological innovation.

One of the core challenges in enforcing regulations concerning AI systems lies in the evidentiary burden placed on claimants. Traditionally, the burden of proof rests with those alleging harm or violations. However, the opaque and complex nature of AI systems—which often operate as dynamic assemblages of algorithms, hardware, and evolving data inputs—renders this requirement particularly onerous. Claimants frequently lack access to proprietary information necessary to substantiate their claims, creating significant barriers to justice. This section examines the issue by drawing parallels with EU non-discrimination law and proposing an evidentiary framework that requires claimants to present only prima facie evidence indicating a potential breach of Article 5(1)(a) or (b). Such evidence might include proof of subliminal techniques, exploitative practices, or significant harm resulting from the operation of an AI system. Once this initial burden is met, the responsibility shifts to the provider or deployer of the AI system to demonstrate compliance with the relevant legal and ethical standards. Such a shift not only aligns with principles of procedural fairness but also ensures that the opacity of AI systems does not shield wrongdoing from scrutiny.

The role of intent within the regulatory framework is another critical dimension explored in this section. While proving deliberate intent to manipulate or exploit is essential for accountability, the AI Act also emphasises the consequences of AI systems' deployment. This dual focus—on purpose and effect—reflects a pragmatic approach to enforcement. Even when intent cannot be conclusively demonstrated, violations may still be established based on the harmful outcomes caused by an AI system. This focus on effects ensures that the regulatory framework effectively addresses harm, particularly in cases where complex automation obscures the motivations behind a system's design or operation. It also underscores the need for continuous vigilance in monitoring the real-world impacts of AI systems, enabling enforcement mechanisms to respond dynamically to emerging risks.

Quantifying harm is a cornerstone of regulatory enforcement, requiring detailed and often interdisciplinary methodologies to capture the full scope and impact of violations. The section delineates multiple categories of harm—including physical injuries, psychological distress, financial losses, and data corruption—each of which demands tailored approaches for assessment. For instance, psychological harm may require clinical evaluations and expert testimony to determine its severity. At the same time, data loss might be quantified by calculating the economic value of the lost information and its downstream effects on the affected parties. These assessments must be rigorous and context-sensitive, ensuring that tangible and intangible harms are adequately addressed. By developing robust methodologies for harm quantification, regulatory frameworks can provide more consistent and equitable enforcement outcomes.

The dynamic and adaptive nature of AI systems further complicates enforcement. Unlike static systems, AI technologies evolve, interacting in increasingly complex ways with their components—including algorithms, hardware, and data. This fluidity necessitates an equally dynamic regulatory approach encompassing continuous monitoring, iterative assessments, and adaptive controls. Key tools in this paradigm include regular algorithmic audits, real-time impact assessments, and enhanced transparency mechanisms to ensure that AI systems remain compliant throughout their lifecycle. Moreover, regulatory frameworks must account for the potential of cascading failures and unintended consequences, which are particularly prevalent in interconnected and highly automated environments. By proactively addressing these challenges, regulators can mitigate risks before they materialise into significant harm.

Distinguishing between manipulation and persuasion is a particularly nuanced challenge in the context of AI systems. Manipulation often operates covertly, undermining user autonomy through deceptive or opaque practices, while persuasion typically engages users transparently and respects their capacity for informed decision-making. Assessing these distinctions requires a multifaceted approach integrating technical, behavioural, and ethical analyses. Technical evaluations can identify design elements or functionalities that may facilitate manipulation, such as dark patterns or subliminal cues. Behavioural studies can assess how users respond to these elements, providing empirical evidence of their influence. Ethical evaluations, meanwhile, contextualise these findings within broader principles of autonomy, dignity, and fairness. Together, these approaches enable regulators to precisely identify manipulative practices and enforce standards that uphold user trust and agency.

The section concludes by emphasising the need for an interdisciplinary regulatory framework that integrates legal, technical, and ethical expertise. Such a framework must be adaptive, capable of evolving alongside technological advancements, and sufficiently robust to address the multifaceted risks posed by AI systems. By balancing

innovation with accountability, this approach can ensure that AI technologies are developed and deployed responsibly, fostering an ecosystem prioritising human well-being, equity, and trust. The challenges posed by AI systems are undoubtedly complex. Still, with rigorous governance and proactive enforcement, their potential for harm can be mitigated while their benefits are harnessed for societal good.[88] Such an approach ensures that AI systems remain transparent, respect user autonomy, and do not exploit vulnerabilities.

Distinguishing manipulation from persuasion in AI systems necessitates rigorous analytical methodologies encompassing transparency assessments, behavioural studies, technical audits, and empirical evaluations. Transparency is a foundational criterion for determining whether an AI system influences users openly and informally or employs covert tactics undermining autonomy. A transparent system communicates its intent, operational mechanisms, and the methodologies through which data is collected, processed, and utilised. For example, a system that generates personalised content or recommendations should explicitly disclose how user data is employed to create these outputs. Such transparency enables users to understand the factors influencing their decisions and interactions, fostering greater awareness of the processes at play.

Evaluating transparency requires thoroughly examining AI systems' design and user interface, mainly explicit disclosures about data collection practices and the purposes behind AI-driven recommendations or decisions. Design features such as precise consent mechanisms, detailed explanations of data use, and options for users to control their data provide critical insights into the transparency of a system. Systems with opaque interfaces or obscure disclosures are more likely to use manipulative practices by exploiting users' cognitive or informational asymmetries. Therefore, Transparency assessments must prioritise detecting features that enhance or diminish user understanding of the system's operations.

Behavioural studies are pivotal in determining the nature of influence exerted by AI systems.[89] These studies leverage controlled experimental conditions to analyse user interactions and decision-making processes using transparent versus non-transparent systems. Metrics such as user awareness of influence, perceived voluntariness of decisions, and alignment of user choices with prior preferences provide empirical evidence of the system's operational dynamics. Suppose behavioural studies reveal that users are unaware of the system's influence or experience decision-making that feels coerced or inconsistent with their values. In that case, it signals the presence of manipulative tactics.[90] Conversely, users reporting informed and voluntary decision-making interactions suggest the system employs persuasive techniques.[91]

Quantitative metrics, including click-through rates, conversion rates, and engagement durations, serve as complementary indicators of user influence. However, these metrics require careful contextual interpretation. High engagement or conversion rates achieved through deceptive design elements—such as dark patterns or obfuscatory interfaces—indicate manipulation. In contrast, similar metrics driven by informative and explicit content indicate effective persuasion. Evaluators must carefully assess the intent and context of the influence to ensure that these quantitative measures are not misinterpreted or misrepresented. For example, an e-commerce platform that uses urgency cues to pressure users into purchases may achieve high conversion rates. Still, these results may stem from manipulative practices rather than genuine persuasion.[92]

Auditing frameworks provide an additional layer of scrutiny to identify and address manipulative practices within AI systems. Independent evaluators can examine whether algorithms and system designs prioritise legitimate user engagement or exploit vulnerabilities to achieve specific outcomes.[93] For instance, evaluators may assess a social media platform's algorithm to determine whether it amplifies meaningful interactions and user satisfaction or capitalises on emotionally charged content to maximise retention. By analysing algorithmic structures and deployment strategies, audits help establish whether systems align with persuasive methodologies or engage in

---

[88] European Commission. (2021). Ethics Guidelines for Trustworthy AI; See also the legal requirement to process data transparently under Article 5(1)(a) GDPR; Article 13(2)(f) of the GDPR requires data controllers to provide information about the existence of automated decision-making, including profiling, and meaningful information about the logic involved, as well as the significance and envisaged consequences of such processing for the data subject.

[89] Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. new media & society, 20(3), 973-989.

[90] Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. Internet policy review, 8(2); OECD. (2019). Recommendation of the Council on Artificial Intelligence; On prior values, see Note 38, Yeung.

[91] Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. European journal of risk regulation, 4(1), 3-28.

[92] Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018, April). The dark (patterns) side of UX design. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-14); See also Gray, C. M., Santos, C., & Bielova, N. (2023, April). Towards a preliminary ontology of dark patterns knowledge. In Extended abstracts of the 2023 CHI conference on human factors in computing systems (pp. 1-9); Yang & Leiser, "Illuminating Manipulative Design: From" Dark Patterns" to Information Asymmetry and the Repression of Free Choice under the Unfair Commercial Practices Directive." Loy. Consumer L. Rev. 34 (2022): 484; Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. Proceedings of the ACM on human-computer interaction, 3(CSCW), 1-32.

[93] Fogg, B. J. (2003). Persuasive technology: Using computers to change what we think and do. Morgan Kaufmann

manipulative behaviours. These evaluations must also account for the iterative nature of AI systems, which evolve and may introduce new risks as their components adapt and interact.

Psychological assessments further distinguish manipulation from persuasion by evaluating the impact of AI interactions on user perceptions and behaviours. Surveys, interviews, and focus groups provide qualitative data on how users experience their interactions with AI systems. These methods can reveal whether users perceive themselves as informed participants in decision-making processes or as targets of subtle coercion. For instance, users who report confusion, frustration, or a lack of control likely engage with manipulative systems. In contrast, users who feel empowered and transparent about the system's functionality interact with systems that prioritise persuasion over manipulation.

Advanced technical tools such as algorithmic audits and explainability frameworks are critical in uncovering manipulative elements embedded within AI systems. Algorithmic audits delve into the underlying decision-making processes of AI systems to identify biases, hidden features, or design choices that exploit cognitive vulnerabilities. For example, researchers may discover that a recommendation algorithm prioritises specific outcomes that disproportionately benefit the deployer at the expense of user autonomy. On the other hand, explainability tools aim to bridge the gap between system complexity and user comprehension by providing clear, actionable insights into how the system generates recommendations or decisions. These tools enhance user understanding and serve as a mechanism for detecting and addressing manipulative practices.

The methodological approach to evaluating manipulation versus persuasion must be comprehensive and interdisciplinary. Transparency assessments, behavioural studies, and psychological evaluations offer empirical insights into user interactions and decision-making processes, while technical audits and explainability tools reveal systemic features that may facilitate manipulation. When applied in tandem, these methods create a robust framework for assessing the impact of AI systems on user autonomy. For example, an algorithmic audit may uncover that a system exploits cognitive biases, while user surveys may corroborate that those users felt misled or coerced during their interactions. These findings provide a detailed picture of the system's influence, allowing for precise distinctions between manipulative and persuasive practices.

The regulatory implications of these evaluations are significant. Systems designed to prioritise transparent communication, respect for user agency, and informed decision-making are more likely to adhere to permissible persuasive practices. Conversely, systems that rely on covert tactics to exploit vulnerabilities or obscure their operations fall under the manipulation category. The distinction between manipulation and persuasion is not merely theoretical; it has practical consequences for compliance with legal standards and enforcement actions. For instance, systems that undermine user autonomy through deceptive practices may face regulatory penalties or operational restrictions. In contrast, systems that enhance user understanding and choice may be benchmarks for industry best practices.

Finally, the challenge of causality in cases of subliminal persuasion remains a critical consideration. Empirical evidence on the effectiveness of subliminal techniques is mixed, complicating efforts to establish direct causal links between subliminal stimuli and behavioural changes. This ambiguity underscores the importance of focusing on measurable outcomes rather than intent. Regulatory frameworks should prioritise the detection of harm and the mechanisms that facilitate it, ensuring that enforcement efforts are grounded in observable impacts rather than speculative attributions of intent. By adopting such an outcome-oriented approach, stakeholders can more effectively address the complexities of manipulation and persuasion in AI systems, safeguarding user autonomy while fostering trustworthy technological advancements.

10. Exceptions from the prohibitions under Article 5(1)

The AI Act aims to regulate the development, deployment, and use of artificial intelligence (AI) technologies within the European Union, ensuring they do not pose risks to safety, health, and fundamental rights. However, while Recital 29 and other sections of the Act discuss potential exceptions to specific prohibitions, these exceptions should be carefully interpreted.

10.1 Common and Legitimate Commercial Practices, Recital 29

Recital 29 of the AI Act clarifies standard and legitimate commercial practices, such as advertising, that should not qualify as harmful, manipulative AI-enabled practices. This exception underscores that not all commercial uses of AI are inherently manipulative or harmful. The key criterion is whether these practices comply with existing laws and ethical standards. The phrase "in themselves" in Recital 29 is critical to understanding this exception. It implies that standard and legitimate commercial practices are not considered harmful or manipulative simply by their nature or existence. For instance, traditional advertising techniques that use AI to personalise content based on user preferences are not inherently manipulative if they adhere to applicable regulations, such as the GDPR and consumer protection laws, or if they do not cumulatively fulfil all the elements of the prohibitions under Article 5(1)(a) or (b). Such practices are permissible when transparent, respect user consent, and avoid employing covert manipulative techniques to distort behaviour. The AI Act aims to prevent practices that subvert autonomy and decision-making through subliminal or deceptive methods, not legitimate commercial activities conducted in compliance with the law.

11. Interplay with other EU Laws

The AI Act, particularly the prohibitions under Article 5(1)(a) and (b), intersects with various other Union laws, creating a complex regulatory landscape designed to ensure the safe development and deployment of AI technologies. This interplay involves consumer protection.[94], data protection[95], non-discrimination[96], the Digital Services Act (DSA)[97], the Audiovisual Media Services Directive (AVMSD)[98], political advertising[99], the General Product Safety Regulation[100], and criminal law. Understanding these intersections is critical for comprehensively assessing the AI Act's enforcement and implications.

First, consumer protection laws closely align with the AI Act's bans. The prohibitions in the AI Act on manipulative and exploitative AI systems directly reinforce these protections by preventing practices that could harm consumers. One example is AI, which manipulates consumer behaviour subliminally or exploits vulnerabilities, potentially leading to unfair commercial practices. The interplay between the AI Act and the Unfair Commercial Practices Directive (UCPD) is crucial in protecting consumers against deceptive and manipulative practices, particularly in the context of advanced AI technologies. The UCPD, which seeks to protect consumers from unfair business practices, complements the AI Act by addressing the potential for AI systems to engage in or facilitate such practices. The UCPD defines unfair commercial practices as those that are misleading, aggressive, or otherwise contrary to the requirements of professional diligence and that materially distort or are likely to distort the economic behaviour of consumers. The AI Act's prohibitions under Article 5(1)(a) and (b), which target AI systems that manipulate behaviour or exploit vulnerabilities, align closely with the objectives of the UCPD.[101] This alignment ensures a unified approach to preventing consumer harm from AI-driven business practices. One of the critical areas of interplay is the concept of "subliminal techniques" and "manipulative practices" addressed in the AI Act. Deployment of AI systems capable of subtly influencing consumer decisions without their awareness can amount to misleading practices prohibited by the UCPD when they are likely to affect consumers' transactional decisions. For instance, an AI system using personal data to create targeted advertisements exploiting a consumer's psychological profile. The AI Act reinforces the UCPD by explicitly prohibiting such manipulative AI applications, thereby enhancing consumer protection against sophisticated forms of deception.

The UCPD also addresses aggressive commercial practices that exert undue pressure on consumers, impairing their freedom of choice. AI systems designed to exploit consumers' vulnerabilities, such as those with cognitive impairments or under significant stress, can be deemed aggressive under the UCPD. The AI Act's prohibition of AI that exploits vulnerabilities directly supports the UCPD's mandate by ensuring that AI technologies do not use undue influence to manipulate consumer behaviour. This approach is particularly relevant for vulnerable populations disproportionately affected by such practices. Another significant aspect of the interplay involves transparency and information disclosure. The UCPD requires that consumers receive clear and accurate information to make informed decisions. Moreover, the AI Act's focus on preventing harm before it occurs aligns with the preventive measures advocated by the UCPD. Both frameworks emphasise the importance of safeguarding consumers proactively rather than merely responding to harm after it has happened. This proactive approach is essential in the rapidly evolving landscape of AI, where new technologies can quickly introduce unforeseen risks. The interplay between the AI Act and the UCPD strengthens consumer protection by addressing the unique challenges posed by AI technologies. The AI Act's prohibitions on manipulative and exploitative AI systems complement the UCPD's mandate to prevent unfair commercial practices, ensuring a cohesive regulatory environment protecting consumers from sophisticated deception and undue influence. This synergy enhances the

[94] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market (Unfair Commercial Practices Directive).
[95] 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 (General Data Protection Regulation).
[96] Charter of Fundamental Rights of the European Union; Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Race Equality Directive); Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation (Employment Equality Directive); Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast); Directive (EU) 2022/29 on combating violence against women and domestic violence.
[97] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC.
[98] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive).
[99] Regulation (EU) 2024/900 on the transparency and targeting of political advertising.
[100] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.
[101] Goanta, C. (2023). Regulatory Siblings: The Unfair Commercial Practices Directive Roots of the AI Act. Available at SSRN 4337417.

overall effectiveness of consumer protection laws in the EU, fostering a safer and more transparent digital marketplace.

11.1 Intersections Between the AI Act, UCPD, and GDPR: Material Distortion and Consumer Protection

The AI Act and the Unfair Commercial Practices Directive (UCPD) address material distortion, although they do so through different lenses and regulatory priorities. The AI Act focuses on the potential for AI systems to distort human behaviour through advanced and often imperceptible techniques, such as subliminal stimuli and manipulative methods. These practices pose significant risks to individuals by bypassing their ability to perceive or control the influence, leading to behavioural changes that may result in harm. The Act explicitly aims to prevent material distortions that cause significant physical, psychological, or financial harm, mainly when such systems exploit vulnerabilities related to age, disability, or socioeconomic status. By addressing these risks, the AI Act underscores its commitment to safeguarding individual autonomy and well-being in the face of increasingly sophisticated AI technologies. In contrast, the UCPD concentrates on economic behaviours and transactional decisions affected by unfair commercial practices. These include misleading advertisements, omissions of critical information, and aggressive sales tactics that distort consumers' ability to make informed choices. While the UCPD focuses primarily on the economic dimensions of consumer protection, its scope does not extend to broader societal harms or non-economic vulnerabilities, differentiating its approach from that of the AI Act. The UCPD remains centred on consumer relationships with traders. In contrast, the AI Act applies more broadly to providers and deployers of AI systems, encompassing entities and individuals who may not fall within the traditional definitions of traders and consumers.

11.2 Broader Scope and Harm Thresholds in the AI Act

The AI Act exhibits a significantly broader scope than the UCPD regarding its application and the types of harm it addresses. Unlike consumer protection laws, which primarily concern economic harms, the AI Act encompasses a broader array of potential risks, including those that affect psychological, emotional, and social dimensions of well-being. The regulation introduces a threshold of significant harm as a criterion for its prohibitions, a standard absent in the UCPD. This threshold ensures that regulatory attention focuses on substantial risks rather than minor infringements, reflecting the seriousness of potential consequences associated with advanced AI systems. The AI Act's applicability extends beyond the traditional consumer-trader relationship, encompassing scenarios where neither party fits these categories. This distinction highlights the Act's broader aim to regulate the deployment of AI systems in diverse contexts, ensuring that protections are not limited to consumer transactions but address societal and individual risks more comprehensively. By adopting this expansive scope, the AI Act acknowledges AI technologies' pervasive and cross-sectoral nature and their potential to affect individuals in multifaceted ways.

11.3 Interplay with Data Protection Frameworks: The Role of the GDPR

The General Data Protection Regulation (GDPR) establishes a foundational framework for data protection within the EU, emphasising principles such as transparency, consent, and accountability in personal data processing. The AI Act complements these principles by prohibiting AI systems that exploit or manipulate individuals by misusing personal data. The alignment between the AI Act and the GDPR underscores their commitment to ensuring fair, transparent, and ethical use of data-driven technologies. The GDPR's emphasis on informed consent and data transparency resonates strongly with the AI Act's objective to prevent manipulation. By requiring clear and accessible disclosures regarding AI systems' functionalities and potential risks, the AI Act reinforces GDPR principles, ensuring that individuals maintain meaningful control over the use of their data. This interplay highlights the necessity of harmonising data protection and AI governance to address the unique challenges posed by AI technologies, particularly in ensuring that personal data does not create or exacerbate vulnerabilities. The AI Act further extends these protections by addressing forms of harm that extend beyond data misuse. While the GDPR primarily protects the privacy and security of personal data, the AI Act targets broader manipulative practices that leverage AI capabilities to distort individual decision-making. This distinction illustrates the complementary nature of the two frameworks, with the AI Act filling regulatory gaps related to the behavioural and societal implications of AI technologies.

11.4 Holistic Regulatory Alignment for Comprehensive Protection

The alignment of the AI Act, UCPD, and GDPR reflects a broader effort within the EU to create a cohesive regulatory framework that addresses the multifaceted risks posed by AI technologies. These regulations aim to safeguard individuals from undue influence, exploitation, and harm by integrating principles from consumer protection, data privacy, and AI governance. The AI Act's focus on advanced manipulative techniques complements the UCPD's emphasis on economic decision-making and the GDPR's prioritisation of data protection, creating a layered and comprehensive approach to regulation. This regulatory interplay is essential for addressing the complex and interconnected challenges AI technologies pose. By ensuring that these frameworks work in concert, the EU establishes a robust foundation for protecting individuals from the diverse risks associated with AI while fostering trust in emerging technologies. The AI Act's broad scope and targeted prohibitions, combined with the UCPD's

economic focus and the GDPR's data protection principles, represent a unified strategy for managing AI's transformative impacts on society.

EU non-discrimination laws, including those embedded in the Charter of Fundamental Rights of the European Union, the EU Race Equality Directive[102], the EU Employment Equality Directive, the EU Gender Equality Directive[103], and the Directive on Combating Violence against Women and Domestic Violence plays a vital role in the AI Act's framework. The AI Act prohibits AI systems that exploit vulnerabilities in ways that could lead to discriminatory outcomes. This prohibition is particularly relevant for AI in hiring, lending, and other applications where biased algorithms could perpetuate systemic discrimination. By integrating non-discrimination principles for certain specific vulnerable groups due to age, disability and specific socio-economic situation, the AI Act aims to ensure that AI technologies promote fairness and equality. The prohibitions in the AI Act do not affect prohibitions based on other grounds or discriminatory practices that do not entail significant harms already prohibited by Union non-discrimination law.

The DSA further complements the AI Act by regulating online platforms and ensuring transparency and accountability in digital services. The DSA's provisions for platforms' content moderation responsibility and the recommender systems processes' transparency align with the AI Act's goals.[104] The DSA, Digital Markets Act (DMA), and the Data Act primarily address the regulation of dark patterns within the user interface by setting standards for transparency, fairness, and user autonomy in online platforms. These legislative frameworks aim to combat manipulative practices that distort user choices through deceptive design, ensuring that interfaces do not mislead or coerce users into actions that may not align with their genuine intentions. The DSA, for instance, prohibits the use of dark patterns that obscure the decision-making process. Similarly, the DMA focuses on maintaining competitive fairness by preventing gatekeepers from exploiting user interfaces to entrench their market dominance through deceptive tactics. The Data Act complements these efforts by ensuring that data access and sharing mechanisms are transparent and do not involve manipulative designs that could compromise user consent.

The Audiovisual Media Services Directive (AVMSD) also intersects with the AI Act, mainly concerning political and commercial advertising. The AI Act's prohibitions on manipulative AI techniques support the AVMSD's objectives by preventing AI-driven ads that exploit consumers' subconscious or vulnerable states.[105] This interplay ensures that AI technologies used in media services do not undermine consumer trust or democratic processes.

The AI Act holds relevance in political advertising, addressing the increasing concerns surrounding using artificial intelligence in democratic processes. The Act explicitly prohibits using AI systems to subliminally manipulate voters or exploit their vulnerabilities, reinforcing principles of transparency and fairness in political campaigns. These measures align with broader European Union initiatives aimed at safeguarding democratic integrity. By banning manipulative AI practices in political advertising, the Act seeks to protect voters from undue influence and ensure that political discourse remains authentic and equitable. This approach reflects a commitment to maintaining trust in democratic institutions and mitigating the risks of technologically driven misinformation.

The General Product Safety Regulation (GPSR) intersects significantly with the AI Act, particularly in consumer safety and product reliability. Both legislative frameworks underscore the necessity of ensuring that AI-integrated products adhere to stringent safety standards throughout their lifecycle.

Finally, the interplay with criminal law is critical. The AI Act's prohibitions aim to prevent harmful behaviours that could be criminal, such as fraud, discrimination, or exploitation. Ensuring that AI systems comply with criminal law standards is essential for protecting individuals and maintaining public trust in AI technologies. Compliance also includes alignment with the Directive on Combating Violence against Women and Domestic Violence.[106], which references the prevention and prosecution of AI-facilitated gender-based violence, including sexually explicit deepfakes.[107] The AI Act's prohibitions in Article 5(1)(a) and (b) provide clear guidelines to prevent AI from being used in ways that could facilitate or obscure criminal activities.

The prohibitions outlined in Article 5(1)(a) and (b) of the AI Act closely interact with numerous other Union laws, creating a cohesive regulatory framework. This interconnected approach safeguards consumers, strengthens data privacy protections, prevents discriminatory practices, regulates digital and media services, enforces product safety standards, and ensures alignment with criminal law provisions. By harmonising these diverse legal domains, the

[102] Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Official Journal of the European Communities, L 180, 19.7.2000, p. 22–26.
[103] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast). Official Journal of the European Union, L 204, 26.7.2006, p. 23–36.
[104] Directive (EU) 2023/1629 of the European Parliament and of the Council of 25 July 2023 on combating violence against women and domestic violence. Official Journal of the European Union, L 240, 11.9.2023, p. 1–35.
[105] Article 9, AVMSD.
[106] Directive (EU) 2023/1629 of the European Parliament and of the Council of 25 July 2023 on combating violence against women and domestic violence.
[107] Article 18.

AI Act establishes a foundation for the responsible development and use of AI technologies throughout the European Union. This comprehensive framework fosters public trust by ensuring that AI innovations contribute positively to societal progress while minimising potential risks.

Legal and regulatory frameworks are critical for distinguishing between manipulation and persuasion in AI-driven interactions. For instance, compliance with data protection laws, such as the General Data Protection Regulation (GDPR), provides a benchmark for assessing the transparency and fairness of AI systems. Systems adhering to GDPR requirements for data processing transparency are more likely to operate within the boundaries of ethical persuasion. Conversely, systems that evade these obligations by employing opaque data practices to influence behaviour are more likely to engage in manipulation.

Concluding Thoughts

The report delves into the prohibitions under **Article 5(1)(a)** and **5(1)(b)** of the AI Act, emphasising their significance in mitigating the risks associated with AI systems that can manipulate or exploit individuals and groups. These provisions serve as critical legal mechanisms to address the challenges posed by advanced AI systems, ensuring a balanced regulatory framework that upholds fundamental rights while fostering technological innovation. By examining these prohibitions, the report contributes to understanding how the AI Act intersects with broader legislative frameworks, such as consumer protection laws and data governance standards, to create a cohesive regulatory ecosystem. The analysis reveals the nuanced scope of Article 5(1)(a), which prohibits deploying AI systems that utilise subliminal or purposefully manipulative techniques to distort individual behaviour. This prohibition underscores the importance of preserving cognitive autonomy, particularly in contexts where individuals may be unable to detect or resist such influences. The study dissects the cumulative elements required for the prohibition to apply, including deploying techniques beyond conscious awareness, their objective or effect of materially distorting behaviour, and the resulting significant harm to individuals or groups. By addressing overt and covert manipulative practices, Article 5(1)(a) establishes a foundational safeguard against the misuse of AI systems in consumer, employment, political, and societal contexts.

In parallel, the report explores Article 5(1)(b), which extends these protections to vulnerable groups, such as children, the elderly, and those experiencing socio-economic disadvantages. This provision targets AI systems that exploit specific vulnerabilities, recognising that advanced technologies may disproportionately impact certain individuals or groups. The prohibition ensures that AI systems do not capitalise on vulnerabilities to induce behaviours that result in significant harm. This targeted approach aligns with the broader objectives of EU law, particularly in safeguarding dignity, autonomy, and equity in the digital age. The findings highlight the complexity of determining when the prohibitions apply, given the high evidentiary thresholds embedded within Articles 5(1)(a) and (b). Establishing material distortion of behaviour and causation of significant harm necessitates a robust evidentiary framework, underscoring the importance of integrating interdisciplinary methodologies, such as behavioural science and data analytics, to substantiate claims. The report also identifies the need for clear interpretative guidelines that delineate the boundaries of legitimate AI applications versus those that contravene these prohibitions. The prohibitions under Article 5(1)(a) and (b) reflect the AI Act's broader risk-based approach, balancing the imperative to protect individuals and groups against overregulation that might stifle innovation. By tying the application of these prohibitions to demonstrable harm and measurable impacts, the Act ensures that regulatory interventions remain proportional to the risks involved. This approach is particularly critical in an era where AI systems are increasingly embedded in everyday decision-making, from consumer transactions to political campaigning.

The report underscores the interplay between the AI Act and complementary legal frameworks, such as the Unfair Commercial Practices Directive (UCPD), the General Data Protection Regulation (GDPR), and the Digital Services Act (DSA). These interactions reveal a layered regulatory structure that seeks to address manipulative practices at multiple levels, ranging from transparency obligations to explicit prohibitions. However, the analysis also points to gaps in enforcement mechanisms, particularly in addressing covert manipulative practices that operate below the threshold of conscious awareness. Decisions about how best to enforce these prohibitions must consider the inherent complexity of AI systems and the challenges of proving causation and intent in digital contexts. Articles 5(1)(a) and (b) represent pivotal regulatory measures within the AI Act, safeguarding autonomy, dignity, and equity in the face of rapidly advancing AI technologies. By establishing clear prohibitions against subliminal, manipulative, and exploitative practices, these provisions reinforce the EU's commitment to protecting individuals and vulnerable groups from harm. Their integration into the broader regulatory framework positions the AI Act as a cornerstone of ethical and responsible AI governance, ensuring that technological progress aligns with EU law's fundamental rights and values. The findings of this report provide a comprehensive foundation for future deliberations, offering insights into how these prohibitions can be interpreted, applied, and refined to meet the evolving challenges of the digital age.

Bibliography

Academic Articles

- Adomaitis, Laurynas, and Alexei Grinbaum. "Manipulation in XR and NLP: A Selection from TechEthos D2.2." CEA Saclay, 2023.
- Ahmed, S., & Skoric, M. M. "The Impact of AI-Powered Social Bots on Political Discourse and Public Opinion." *Social Media + Society*, 9(1), 205630512311516, 2023.
- Ananny, M., & Crawford, K. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society*, 20(3), 973-989, 2018.
- Barilan, Y. M., & Weintraub, M. "Persuasion as Respect for Persons: An Alternative View of Autonomy and the Limits of Discourse." *The Journal of Medicine and Philosophy*, 26(1), 13-34, 2001.
- Boine, C. "AI-enabled Manipulation and EU Law." Available at SSRN: https://ssrn.com/abstract=4042321 or http://dx.doi.org/10.2139/ssrn.4042321, 2021.
- Burr, C., & Cristianini, N. "Can Machines Read Our Minds?" *Minds and Machines*, 29(3), 461-494, 2019.
- Cohen, T. "Regulating Manipulative Artificial Intelligence." *SCRIPTed*, 20, 203, 2023.
- Englefriet, Arnoud. "The Annotated AI Act."
- Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. "The Rise of Social Bots." *Communications of the ACM*, 59(7), 96-104, 2016.
- Feltes, T., Balloni, A., Czapska, J., Bodelón, E., & Stenning, P. "Gender-based Violence, Stalking, and Fear of Crime. Country Report Germany." EU-Project 2009-2011.
- Franklin, M., Tomei, P. M., & Gorman, R. "Strengthening the EU AI Act: Defining Key Terms on AI Manipulation." arXiv preprint arXiv:2308.16364, 2023.
- Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. "The Dark (Patterns) Side of UX Design." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-14, 2018.
- Hansen, P. G., & Jespersen, A. M. "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy." *European Journal of Risk Regulation*, 4(1), 3-28, 2013.
- Ienca, M. "On Artificial Intelligence and Manipulation." *Topoi*, 42(3), 833-842, 2023.
- Ito, A., Abe, N., Fujii, T., & Mori, E. "Neural Correlates of Spontaneous Deception." *Social Cognitive and Affective Neuroscience*, 14(9), 978-988, 2019.
- Karremans, J. C., Stroebe, W., & Claus, J. "Beyond Vicary's Fantasies: The Impact of Subliminal Priming and Brand Choice." *Journal of Experimental Social Psychology*, 42(6), 792-798, 2006.
- Leiser, Dr Mark. "Protecting Children from Dark Patterns and Deceptive Design." Available at SSRN: https://ssrn.com/abstract=4660222, 2023.
- Leiser, M. "Psychological Patterns and Article 5 of the AI Act: AI-Powered Deceptive Design in the System Architecture and the User Interface." *Journal of AI Law and Regulation*, 1(1), 5-23, 2024.
- Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites." *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1-32, 2019.
- Mitra, S., & Gilbert, E. "Deception in the Digital Age: A Review of the Literature and Directions for Future Research." *Journal of Business Ethics*, 179, 333-354, 2022.
- Noggle, R. "Manipulative Advertising." *Journal of Business Ethics*, 166, 651–668, 2020.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. "The Emotional Stroop Effect and Unconscious Processing: A Meta-Analytic Review." *Journal of Experimental Psychology: General*, 134(2), 277–293, 2005.
- Patalay, P., & Fitzsimons, E. "Psychological Distress, Self-Harm, and Attempted Suicide in UK 17-Year-Olds: Prevalence and Sociodemographic Inequalities." *The British Journal of Psychiatry*, 219(2), 437-439, 2021.
- Sagarin, B. J., Cialdini, R. B., Rice, W. E., & Serna, S. B. "Dispelling the Illusion of Invulnerability: The Motivations and Mechanisms of Resistance to Persuasion." *Personality and Social Psychology Bulletin*, 28(7), 886-895, 2002.
- Susser, D., Roessler, B., & Nissenbaum, H. "Technology Autonomy and Manipulation." *Internet Policy Review*, 8(2), 2019.
- Trzaskowski, J. "Lawful Distortion of Consumers' Economic Behaviour: Collateral Damage Under the Unfair Commercial Practices Directive." *European Business Law Review*, 27(1), 2016.
- Van Swol, L. M., & Braun, M. T. "Deception Detection and Decision Making: A Review of Cognitive and Social Factors." *New Ideas in Psychology*, 55, 100732, 2019.
- Yeung, K. "Hypernudge: Big Data as a Mode of Regulation by Design." *Information, Communication & Society*, 20(1), 118-136, 2017.

Books

- Arnoud Englefriet. "The Annotated AI Act."
- Cialdini, R. B. (2021). Influence new and expanded: The psychology of persuasion. Harper Business.
- Neuwirth, Rostam Josef. The EU Artificial Intelligence Act: Regulating Subliminal AI Systems (August 15, 2022). London: Routledge, 2023. SSRN Link or DOI Link.
- O'Keefe, D. J. (2016). Persuasion and social influence. The international encyclopaedia of communication theory and philosophy, 1-19.
- Pratkanis, A. R., & Aronson, E. (2001). Age of propaganda: The everyday use and abuse of persuasion. Macmillan.
- Whish, R., & Bailey, D. (2018). Competition law (9th ed.). Oxford University Press.

Case Law

- C-503/13 Boston Scientific Medizintechnik GmbH v AOK Sachsen-Anhalt – Die Gesundheitskasse EU:C:2015:148.
- Case C-127/02 (Waddenvereniging and Vogelbeschermingsvereniging).
- Case C-210/96 Gut Springenheide GmbH and Rudolf Tusky v Oberkreisdirektor des Kreises Steinfurt - Amt für Lebensmittelüberwachung.
- Case C-220/98 Estée Lauder Cosmetics GmbH & Co. ORG, v Lancaster Group GmbH, opinion of Advocate General Fennelly.
- Case C-258/11 (Sweetman and Others).
- Case C-281/12 Trento Sviluppo srl and Centrale Adriatica Soc. coop. arl v Autorità Garante della Concorrenza e del Mercato.
- C-542/13 M'Bodj v État belge.
- Case C-621/21 WS v Intervyuirasht organ na Darzhavna agentsia za bezhantsite pri Ministerskia savet [2024] ECLI:EU: C:2024:47.
- MD 2010:8 Toyota Sweden AB v Volvo Personbilar Sverige AB (Marknadsdomstolen, 12 March 2010).

Directives and Legislation

- Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L 210/29.
- Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats, wild fauna, and flora [1992] OJ L 206/7.
- Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Official Journal of the European Communities, L 180, 19.7.2000, p. 22–26.
- Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation [2000] OJ L 303/16.
- Council Regulation (EC) No 1/2003 of 16 December 2002 on implementing the rules on competition laid down in Articles 81 and 82 of the Treaty [2003] OJ L 1/1, recital 38.
- Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy.
- Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC (Text with EEA relevance
- Regulation (EU) 2017/2394 of the European Parliament and the Council of 12 December 2017 on cooperation between national authorities responsible for the enforcement of consumer protection laws and repealing Regulation (EC) No 2006/2004 [2017] OJ L 345/1.
- Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market (Unfair Commercial Practices Directive).
- Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of specific provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) (Text with EEA relevance).
- Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC [2014] OJ L 153/62.
- Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services [2019] OJ L 151/70.

- Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 on corporate sustainability reporting (Corporate Sustainability Reporting Directive) [2022] OJ L 330/1.
- General Data Protection Regulation (GDPR): 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), OJ 2016 L 119/1.
- Regulation (EU) No 596/2014 of the European Parliament and of the Council of 16 April 2014 on market abuse (market abuse regulation) and repealing Directive 2003/6/EC of the European Parliament and of the Council and Commission Directives 2003/124/EC, 2003/125/EC and 2004/72/EC Text with EEA relevance.
- Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117/1.
- Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on sustainability-related disclosures in the financial services sector (Text with EEA relevance) PE/87/2019/REV/1 OJ L 317, 9.12.2019.
- Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the establishment of a framework to facilitate sustainable investment and amending Regulation (EU) 2019/2088 (Text with EEA relevance).
- Directive (EU) 2023/1629 of the European Parliament and of the Council of 25 July 2023 on combating violence against women and domestic violence. Official Journal of the European Union, L 240, 11.9.2023, p. 1–35.
- Directive (EU) 2024/900 on the transparency and targeting of political advertising.
- Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Official Journal of the European Union, L 277, 27.10.2022, p. 1–102.
- 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 (Artificial Intelligence Act) (Text with EEA Relevance).
- General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz - AGG) (Germany) 14 August 2006 Federal Law Gazette I p. 1897 last amended by Article 8 of the Act of 3 April 2013 Federal Law Gazette I p. 610.
- French Labour Code (Code du travail) (France) Article L1132-1 last amended by Law No. 2019-486 of 22 May 2019.

Online Sources

- European Commission. "Ethics Guidelines for Trustworthy AI." 2021.
- Federal Trade Commission (FTC). *Endorsement Guides*. **Link**.
- Neuwirth, Rostam Josef. "The EU Artificial Intelligence Act: Regulating Subliminal AI Systems." SSRN. **Link**.

Policy Documents

- European Commission. "A multi-dimensional approach to disinformation: Report of the independent High-level Group on fake news and online disinformation." 2018.
- European Commission. "Ethics Guidelines for Trustworthy AI." 2021.
- European Commission. "Guidelines on Data Protection Impact Assessment (DPIA) (wp248rev.01)."
- Federal Trade Commission (FTC) Endorsement Guides.
- OECD. "Recommendation of the Council on Artificial Intelligence." 2019.
- United Nations Convention on the Rights of Persons with Disabilities (CRPD): Article 4(1)(h) CRPD.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). "Journalism, 'Fake News' & Disinformation: A Handbook for Journalism Education and Training." 2018.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). "Recommendation on the Ethics of Artificial Intelligence." 2021.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696,
– via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.