The Many Meanings of Vulnerability in the AI Act and the One Missing

Federico Galli, Claudio Novelli*

ABSTRACT: This paper reviews the different meanings of vulnerability in the AI Act (AIA). We show that the AIA follows a rather established tradition of looking at vulnerability as a trait or a state of certain individuals and groups. It also includes a promising account of vulnerability as a relation but does not clarify if and how AI changes this relation. We spot the missing piece of the AIA: the lack of recognition that vulnerability is an inherent feature of all human-AI interactions, varying in degree based on design choices and modes of interaction. Finally, we show how such a meaning of vulnerability may be incorporated into the AIA by interpreting the concept of "specific social situation" in Article 5 (b).

KEYWORDS: Vulnerability; AI Act; AI; Human-Computer Interaction; Specific Social Situation.

1. Introduction: Vulnerability and the AI Act

The uptake of AI and digital technologies, coupled with the increased awareness of their risks to human beings, has revamped the interest in the concept of vulnerability within the legal field. The discussion has taken place both at empirical and regulatory levels.

[*] *Federico Galli: Research assistant, University of Bologna. Mail: federico.galli7@unibo.it; Claudio Novelli: Claudio Novelli: postdoctoral researcher, University of Bologna. Mail: claudio.novelli@unibo.it. While the research results are based on a combined effort, Sections 2, 4 and 5 should be attributed to Claudio Novelli, while Sections 3, 6-8 should be attributed to Federico Galli. Introduction and conclusion are shared. Federico Galli was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (GA. 833647) and by the PRIN 2022 PNRR Project DAFNE (P2022R7RS9) under the MUR National Recovery and Resilience Plan funded by the European Union – Next Generation EU. The article was subject to a double-blind peer review process.*

At the empirical level, research has focused on the impact of the deployment of AI in specific areas, like finance[1], social networks[2], and dispute resolution[3]. Moreover, research has shown that AI systems can exacerbate existing inequalities and disproportionately affect already disadvantaged groups in society, such as gender minorities, low-income individuals, and those with limited competence with technology[4].

At the regulatory level, the discussion is pivoting around a key question: to what extent the vulnerability concept can represent a normative benchmark for different AI-powered contexts and practices, thereby requiring enhanced protection[5]. In other words, this would mean establishing new legal standards and safeguards designed to protect individuals and groups susceptible to harm due to AI technologies.

Some recent EU regulatory initiatives establishing legal frameworks for AI development and use have increasingly referred to the concept of vulnerability[6]. Among these, the recently adopted EU Artificial Intelligence Act (henceforth, AIA)[7] seems to be the one taking the idea of vulnerability most seriously.

While the research results are based on a combined effort, Sections 2, 4 and 5 should be attributed to Claudio Novelli, while Sections 3, 6-8 should be attributed to Federico Galli. Introduction and conclusion are shared. Federico Galli was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (GA. 833647) and by the PRIN 2022 PNRR Project DAFNE (P2022R7RS9) under the MUR National Recovery and Resilience Plan funded by the European Union – Next Generation EU.

[1] E. Mogaji, T.O. Soetan, T.A, Kieu, *The implications of artificial intelligence on the digital marketing of financial services to vulnerable customers*, in *Australasian Marketing Journal*, 29, 3, 2021, 235.

[2] See, among many studies, N. Bol, J. Strycharz, N. Helberger, B. van de Velde, C. H de Vreese, *Vulnerability in a tracked society: Combining tracking and survey data to understand who gets targeted with what content*, in *New Media & Society*, 22, 11, 2020, 1996.

[3] F. Casarosa, *Access to (Digital) Justice: Is There a Place for Vulnerable People in Online Dispute Resolution Mechanisms?*, in *Journal of European Consumer and Market Law*, 13, 3, 2024, 126.

[4] P. Kerrigan, M. Barry, *Automating vulnerability: Algorithms, artificial intelligence and machine learning for gender and sexual minorities*, in P. Aggleton, R. Cover, C.H. Logie, C.E. Newman, R. Parker (eds.), *Routledge Handbook of Sexuality, Gender, Health and Rights,* London, 2023, 164; M. Gilman, *POVERTY LAWGORITHMS A Poverty Lawyer's Guide to Fighting Automated Decision-Making Harms on Low-Income Communities*, Data & Society Research Institute, 2020, in https://datasociety.net/wp-content/uploads/2020/09/Poverty-Lawgorithms-20200915.pdf (last accessed: 29/11/2024); C. Wang, S.C. Boerman, A.C. Kroon, J. Möller, C. de Vreese, *The artificial intelligence divide: Who is the most vulnerable?*, in *New Media & Society*, 24 February 2024, 1-23.

[5] See the discussion around digital vulnerability in private and consumer law, e.g., N. Helberger, M. Sax, J. Strycharz, H.W. Micklitz, *Choice architectures in the digital economy: Towards a new understanding of digital vulnerability*, in *Journal of Consumer Policy*, 44, 4, 2021, 175; M. Grochowski, *Does European contract law need a new concept of vulnerability?* in *Journal of European Consumer and Market Law*, 10, 44, 2021, 133; F. Galli, *Algorithmic Marketing and EU Law on Unfair Commercial Practices*, Berlin/Heidelberg, 2022, 181-207. An equivalent debate has taken shape in the constitutional/administrative law sphere: S. Ranchordas, *Empathy in the Digital Administrative State,* in *Duke Law Journal*, 71, 6, 2021, 1341; S. Ranchordas, *The Invisible Citizen in the Digital State: Administrative Law Meets Digital Constitutionalism*, in C. van Oirsouw, J. de Poorter, I. Leijten, G. van der Schyff, M. Stremler, M. de Visser (eds.), *European Yearbook of Constitutional Law* (forthcoming*,* 2024).

[6] More or less extensive references to vulnerability are contained in the Digital Services Act, the Digital Markets Act, the Data Act, the Regulation on Political Advertising, and the Cyber Resilience Act. For a comparative review, see M. Sax, N. Helberger, *Digital Vulnerability and Manipulation in the Emerging Digital Framework*, in *Digital Fairness for Consumers*, A report commissioned by BEUC, The European Consumer Organisation, 2024, 11.

[7] 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

Downloaded from www.biodiritto.org.
ISSN 2284-4503

"Vulnerability" is mentioned 19 times, 12 of which are in the Recitals and 7 in the binding part of the text, in several thematic areas. For example, AI systems exploiting some individual's vulnerabilities are classified as a prohibited practice (Article 5(1)(b)). Vulnerability is also a parameter for the European Commission to update the list of high-risk AI systems (Article 7(h)). The extent to which the high-risk AI system impacts minors and other vulnerable groups is one of the steps under the risk-management system (Article 9(9)). Within the context of regulatory sandboxes in the AIA, individuals in a condition of vulnerability due to their age or disability must be appropriately protected (Article 60(4)(g)). When dealing with AI systems presenting risk, market surveillance authorities must pay particular attention to the risks that AI systems present to vulnerable groups (Article 79(2)). The AI Office and Member States should facilitate the drawing up of codes of conduct, inter alia, on assessing and preventing the negative impact of AI systems on vulnerable persons or groups (Article 95). However, the practical implementation of these provisions remains unclear, particularly regarding how the AI Office will fulfil this role and coordinate with other bodies established under the AIA[8].

Despite these many occurrences, the AIA does not provide a unified definition of "vulnerability," thus leaving the term open to interpretation in each instance it is adopted. One may even doubt whether all occurrences of the term refer to the same concept.

In this paper, we review the different meanings of vulnerability contained in the AIA. We show that the AIA follows a rather established tradition of looking at vulnerability as a trait or a state of certain individuals and groups. It also includes a promising notion of vulnerability as a relation, but it does not clarify if and how AI changes this relation. Then, we spot the missing piece of the AIA, namely an idea of vulnerability as a characteristic of all AI-human relations, which manifests depending on different design features and interaction modes. To address this gap, we argue how such a view of vulnerability may be incorporated into the current text of the AIA by interpreting the concept of "specific social situation" contained in Article 5 (b).

2. The Underlying Meaning: Vulnerability as a Key Factor in the AIA's Objectives and Risk Analysis

The absence of an explicit and unified definition of vulnerability in the AIA does not preclude inferring it from the text, where the term is repeatedly used with different referents.

The AIA offers a nuanced account of human vulnerability in interactions with AI systems, as highlighted by the combined normative meaning of Article 5 and various Recitals, notably 5 and 48. These sections emphasise the power, knowledge, and agency imbalances between individuals and AI technology providers. Consequently, the AIA aims to protect individuals who depend on AI systems to fulfil a purpose or exercise a right, acknowledging their potential vulnerability. The AIA's normative references to

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), OJ 12.7.2024.

[8] C. Novelli, P. Hacker, J. Morley, J. Trondal, L. Floridi, A *Robust Governance for the AIA: AI Office, AI Board, Scientific Panel, and National Authorities*, in *European Journal of Risk Regulation,* 2024, 1-25.

vulnerability collectively present a multifaceted view that includes factors such as age, health status, financial situation, and level of social inclusion[9].

From this perspective, vulnerability has a dual dimension: it is a *general* condition, where merely possessing fundamental rights increases the risk of negative impacts from AI systems, and a *specific* condition, which depends on the right-holder individual situation (e.g., age, health, education).

As a general condition, vulnerability plays a specific role in achieving the AIA's objectives. As stated in its first recitals, the AIA aims to safeguard fundamental rights such as human dignity, democracy, equality, and the rule of law, which form the bedrock of the EU's approach to AI governance. When AI systems engage with fundamental rights, such as those in employment or law enforcement, the severity of adverse consequences may increase. These consequences can vary significantly based on the specific conditions or traits of the affected individuals or groups. Thus, any interference with fundamental rights resulting from AI deployment must be justifiable.

European legal culture and its case law are heavily influenced by the belief that resolving conflicts between fundamental rights and competing interests (or among rights) – such as those arising from the deployment of AI systems – is inherently complex and requires a balanced approach. This is because they are typically contained in legal principles, which are designed to be open-ended, explicitly value-driven, defeasible optimisation directives that can be realised in various ways and to varying extents (unlike legal rules).[10] They must coexist as far as possible. Thus, conflicts involving these principles are addressed through a proportionality procedure. This procedure facilitates the balancing and trade-offs of these rights in specific situations.

In the AIA, this trade-off procedure takes the form of a risk-based regulation[11]. AI systems are classified according to their varying risk levels. So, for instance, systems that pose unacceptable risks are prohibited because their (prospected) benefits do not outweigh the (potential) harm they may cause to fundamental rights. High-risk systems require more stringent legal safeguards before being brought to market[12].

In this risk-based regulatory architecture, vulnerability constitutes a key component of AI risk. This interpretation aligns with established risk science methodologies and prominent policy reports, such as those by the Intergovernmental Panel on Climate Change (IPCC). In these contexts, vulnerability is a critical factor in evaluating risk magnitude, as it impacts both the likelihood and severity of

[9] Among the most prominent and influential proponents of such a "universal approach to vulnerability" is Martha Fineman. According to Fineman, «human vulnerability arises in the first place from our embodiment, which carries with it the imminent or ever-present possibility of harm, injury, and misfortune». It follows that if vulnerability is embodied, «we can attempt to lessen risk or act to mitigate possible manifestations of our vulnerability» but «the possibility of harm cannot be eliminated». According to this understanding of vulnerability, vulnerable subjects are not the exception; they are the rule. See, M. FINEMAN, *The Vulnerable Subject: Anchoring Equality in the Human Condition*, in *Yale Journal of Law and Feminism*, 20, 1, 2008, 9.

[10] R. ALEXY, *On the Structure of Legal Principles*, in *Ratio Juris*, 13, 2000, 294 ss.; R. ALEXY, *Constitutional Rights, Balancing, and Rationality,* in *Ratio Juris*, 16, 2003, 131-140.

[11] In essence, this is a cost-benefit analysis inspired by the precautionary principle. Given the nature of the regulation, this proportionality procedure is merely outlined in the AIA itself, with the majority of the assessment and balancing to be carried out during the implementation and enforcement phases, primarily by the courts.

[12] Many of these high-risk systems are enumerated in Annex III of the Regulation, reflecting AI applications that align with core European values.

consequences of a risk event[13]. By factoring the susceptibility of individuals, communities, or regions to adverse effects from hazard sources, alongside other risk components like exposure and response mechanisms, policymakers can develop a more accurate understanding of specific risk scenarios and tailor regulations accordingly. Essentially, the vulnerability in the AIA normative philosophy and architecture is an AI system's risk amplifier.

To illustrate this briefly, consider the AIA's attention to physical and mental disabilities. So, for instance, AI systems used in healthcare may not be designed to accommodate individuals with disabilities, limiting their access and potentially perpetuating bias against those with pre-existing conditions. This is even clearer in cases of malicious intent, such as AI systems designed to exploit emotional triggers and manipulate users into sharing personal information. Individual vulnerability in these cases – in its dual dimension – contributes to signalling the risk level of an AI system and triggers higher standards and increased responsibility.

3. Vulnerability as a Trait or a Situation of Persons/Groups that Can be Exploited

One explicit reference to vulnerability is contained in Article 5, which prohibits certain AI practices. Among the latter, Article 5, lit. b, prohibits «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».

The prohibition refers to "any kind of vulnerability", but in reality, it scopes out only some sources[14] of vulnerability: 1) age, 2) disability, and 3) a specific social or economic situation. The list of sources seems to be exhaustive. Recital 29 only clarifies that "disability" must be interpreted in line with the notion of "people with disability" contained in Directive 2019/882[15].

This meaning of vulnerability shares many similarities, both in the conceptual framework and in the literal wording[16], with the Directive 2005/29/CE on unfair commercial practices[17]. The Directive

[13] N.P. Simpson, K.J. Mach, A. Constable, J. Hess, R. Hogarth, M. Howden, J. Lawrence, R.J. Lempert, V. Muccione, B. Mackey, M.G. New, *A framework for complex climate change risk assessment*, in *One Earth*, 4, 4, 2021, 489; C. Novelli, F. Casolari., A. Rotolo, M. Taddeo, L. Floridi, *AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AIA*, in *Digital Society*, 3, 13, 2024, 1-29; A.W. Coburn, R.J.S. Spence, A. Pomonis, *Vulnerability and Risk Assessment, Disaster management training programme*, Cambridge, 1994.

[14] We shall also refer to them as "vulnerability drivers".

[15] Recital 29 clarifies that the concept of «disability» must be interpreted in line with the notion of «people with disability» contained in Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services, that is, «people 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».

[16] See the in-depth analysis by C. Goanta, *Regulatory Siblings: The Unfair Commercial Practices Directive Roots of the AI Act*, in I. Graef, B. van der Sloot (eds.), *The Legal Consistency of Technology Regulation in Europe*, London, 2024, 71.

[17] 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 and amending Council Directive 84/450/EEC,

recognises that factors like age, mental capacity, and credulity can make certain consumers more susceptible to unfair commercial practices. Specifically, it prohibits practices that exploit these vulnerabilities in a way that traders should reasonably anticipate[18].

It seems, however, that the AIA made some important steps forward, which were probably informed by the contemporary debate on vulnerability.

First, the understanding of vulnerability has transformed. It is no longer seen solely as an inherent trait of specific individuals or groups. Instead, it is now recognised as a situational and context-dependent condition that can potentially affect all human beings. This aligns with vulnerability theory, as articulated by scholars like Florencia Luna[19], which emphasises that inherent human vulnerability, stemming from our physical and social nature, is amplified by situational and structural factors. According to Luna, multiple and different layers of vulnerability may overlap. Some of them may be related to problems of knowledge, others to possible violations of human rights, to temporary situations that individuals find themselves in, or to the characteristics of the person involved.

Secondly, among the contextual drivers, the AIA considers both cognitive impairment due to external pressure[20] and socio-economic factors. This move reflects an upgrade in the awareness that vulnerability can arise from broader social and economic contexts, not merely from individual characteristics. Scholars such as Jonathan Herring argue that socio-economic conditions significantly impact an individual's susceptibility to harm, advocating for broader protections[21]. the AIA acknowledges that vulnerability often stems from systemic inequalities and external pressures beyond individual characteristics.

However, Article 5 remains quite generic on the concrete states of vulnerability, i.e., what specific traits or situations categorise individuals and groups as vulnerable in relation to each source. Regarding "specific social or economic situation", Recital 29 only provides two examples, namely "persons living in extreme poverty" and "ethnic or religious minorities". It remains unclear what other types (if any) of a "specific social situation" may result in a vulnerability state, especially whether they include not only enduring situations but also transient states (e.g., temporary unemployment, recent migration, or short-term financial crises). Moreover, no consideration is given on how the vulnerability traits and situations potentially amplify or conversely alleviate in combination with each other. Certain individuals may possess a combination of personal, social and economic vulnerabilities that makes them more susceptible to exploitation (e.g., children living in poverty), while others with similar conditions may

Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council.

[18] A similar reference is contained in Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council.

[19] F. LUNA, *Elucidating the concept of vulnerability: layers not labels*, in *Int J Fem Approaches Bioethics*, 2, 1, 2009, 121, where is argued that the concept of vulnerability should be understood as a complex and multi-layered phenomenon rather than a simplistic label, advocating for a nuanced approach that takes into account the varying degrees and contexts of vulnerability in bioethical discussions.

[20] Article 5, lit. a) AIA.

[21] J. HERRING, *Vulnerability Adults and the Law*, Oxford, 2016.

have support systems that mitigate these risks (e.g., well-educated children from high-income families).

While the AIA highlights the importance of protecting vulnerable groups, it lacks clear guidelines for defining and identifying such groups. In this context, a group can be broadly defined as a collection of individuals who share certain characteristics or experiences that render them susceptible to exploitation or harm. Relevant characteristics may include demographic factors such as age, disability, and economic status, as well as social conditions like ethnicity, religion, or even transient circumstances such as recent migration or temporary financial crises. Identifying vulnerable groups requires a subtle understanding of how different vulnerabilities intersect and amplify each other. For instance, children living in poverty or elderly individuals with cognitive decline may represent groups with compounded vulnerabilities. However, the AIA falls short in providing specific mechanisms or criteria for recognising such groups or assessing the varying degrees of vulnerability within and across these groups.

Finally, Article 5 does not explain the exact role of AI in exploiting vulnerability. In particular, it needs to be clarified whether exploitation should be understood as an information-based process or whether it suffices for exploitation to manifest that harm to vulnerable individuals and groups occurs. In other words, does Article 5 require that the AI system possesses – either because it is provided with such knowledge or because it was learned by interacting with individuals or groups – information about a vulnerable state and uses it to make a recommendation, decision, etc.?[22] Or, is it enough that the exploitation occurs as a result of the AI system's actions, even if the system does not recognise or process the vulnerability "intentionally"?

For example, consider an AI-driven advertising platform that targets ads for payday loans to users based on their online behaviour and financial data. If the system identifies a user struggling financially and then bombards them with high-interest loan ads, this constitutes information-based exploitation. The AI system is leveraging the user's financial vulnerability to the advantage of the loan company, which profits from the user's desperation and lack of alternatives.

On the other hand, imagine an AI system designed to recommend healthcare services. This system might inadvertently harm financially vulnerable users by recommending expensive treatments without considering their economic constraints. This could lead these individuals to incur debt or forgo necessary care due to cost. Here, the AI system did not specifically leverage the information on economic vulnerability, but the harm still manifests due to the system's actions.

4. Vulnerability as a Feature of (High-Risk) AI Systems

AI systems, like humans, have vulnerabilities that can be exploited. This often-overlooked aspect of AI vulnerability is crucial because these weaknesses can interact with and worsen existing human vulnerabilities.

[22] We are not in any way referring to "mental processes" that imply an intentional state taking place in an AI system.

There are emerging parallels between human and AI vulnerability[23], and this can also be seen in the AIA. As anticipated, Article 5 portrays human vulnerability as involving susceptibility to harm due to endogenous or exogenous factors, with exploitation involving using these weaknesses to the detriment of the vulnerable party. Similarly, AI systems can be exploited to produce harmful outcomes, such as adversarial attacks (external) or the exploitation of biases (internal). Thus, it is not unlikely that the two concepts may influence each other in the implementation of the AIA. The concept of AI/ICT vulnerability is well-established in cybersecurity[24], where systems are continually assessed for weaknesses that could be exploited by malicious actors. We foresee the possibility that computer scientists look at more human-related accounts of vulnerability in the same vein as technical vulnerability[25].
Moreover, exploiting AI systems can directly impact human well-being, creating a cascade effect. For instance, manipulating an AI used in healthcare can lead to misdiagnoses and harm patients. This is why one of the essential requirements is a vulnerability assessment and mitigation of the systems both for high-risk systems (Article 15(5) and Recital 76) and systemic-risk General-Purpose AI Models (Article 55 and Recital 110). On the other hand, human vulnerabilities can amplify AI vulnerability when vulnerable individuals unknowingly provide data, which the AI system then learns from and perpetuates, leading to even more pronounced biases and errors. This interconnectedness may create feedback loops where human vulnerabilities influence AI outcomes, and flawed AI systems exacerbate human vulnerabilities.
However, it is not clear why AI vulnerability issues and related cybersecurity countermeasures should concern only systems/models classified as having high-risk/systemic risk. All AI systems, regardless of their risk classification, have the potential to harbour vulnerabilities that can be exploited, leading to significant consequences. For example, even low-risk applications (e.g., deep fakes or emotion categorisation systems) can become entry points for broader cyberattacks or can perpetuate subtle biases that have far-reaching implications. Focusing solely on high-risk AI systems may neglect the wider landscape of AI vulnerabilities even in benign or less risky applications.

5. Vulnerability as a Trait (and a Situation?) of Affected Persons That Can Be Impacted

A third meaning of vulnerability can be located in the provider's risk-management system obligation in Article 9, applicable to high-risk systems. Accordingly, the provider of a high-risk AI system must identify, assess and mitigate risks to health, safety and fundamental rights to individuals, including giving due consideration «to whether in view of its intended purpose, the high-risk AI system is likely to have an adverse impact on persons under the age of 18 and, as appropriate, other vulnerable groups»[26].

[23] See, e.g., the reasoning line in the blog post by Chief Research Officer at Women in AI NPO, M. Tschopp, *Vulnerability of humans and machines – A paradigm shift*, June 2022, available https://www.scip.ch/en/?labs.20220602 (last visited 27/07/2024).
[24] H. Yupeng, K. Wenxin, Q. Zheng, L. Kenli, Z. Jiliang, G. Yansong, L. Wenjia, L. Keqin, *Artificial Intelligence Security: Threats and Countermeasures*, in *ACM Computing Surveys*, 55, 1, 2021, 1-36.
[25] This may lead to an information-based interpretation of Article 5, lit. b) AIA.
[26] Article 9(9), AIA.

A similar account of vulnerability is also contained in other risk-oriented provisions of the AIA: Article 60(4) about the conditions for testing high-risk AI systems in the real world outside regulatory sandboxes («…the subjects of the testing in real world conditions who are persons belonging to vulnerable groups…»); Article 79 on investigation activities by market surveillance authorities on systems presenting particular risks («...Particular attention shall be given to AI systems presenting a risk to vulnerable groups»); Article 95 on codes of conduct, which applies to systems other than high-risk AI system (…«assessing and preventing the negative impact of AI systems on vulnerable persons or groups of vulnerable persons…»).

An assessment of the likely impact on vulnerable groups was also part of the deployer's obligation of a fundamental right impact assessment (FRIA), as proposed by the European Parliament in the former Article 29. Interestingly, this version also included "marginalised groups"[27]. The reference to vulnerable and marginalised groups was discarded in the final version of the current Article 27, though it is not implausible that measuring the impact on disadvantaged groups will be in the end part of the content of the FRIA[28].

Regardless of the actors to which these different provisions refer (i.e. providers, market surveillance authorities, Member States and providers' associations, deployers), the vulnerability concept, in this third account, provides the benchmark for ex-ante assessing and mitigating the risk of high-risk AI systems.

This notion builds the underlying meaning of vulnerability presented in Section 2, i.e., vulnerability as a component of AI risk, but in addition, looks at vulnerability as a trait of certain groups, similar to Article 5. Compared to the average affected person, vulnerable groups are indeed expected to suffer greater harm in the event of damage caused by an AI system. Therefore, keeping the probability constant in the overall risk assessment, systems that can harm vulnerable groups are inherently riskier because the severity of the expected harm is greater. It is explained why providers and authorities are encouraged to pay more attention to those risks in the mitigation phase or in enforcement actions.

Unlike Article 5(b), however, vulnerability here is not exploited by the AI provider, but it is "impacted". This means that the harm occurs as a potential side-effect of the AI system's operation rather than as a direct exploitation. For example, an AI system designed for automated hiring processes may inadvertently disadvantage individuals with disabilities by not properly accounting for gaps in employment history related to medical treatment. Although the AI provider does not "exploit" the vulnerabilities of disabled applicants, the system's design and deployment may still result in adverse impacts on disabled groups. The focus of the provider, therefore, should be on preventively recognising and mitigating these unintended effects to ensure that AI systems do not result in disproportionate harm, even in the absence of intentional exploitation.

The idea of vulnerability as something whose impact can be predicted is supported by a risk-based theory of vulnerability, which emphasises the importance of understanding and addressing the specific

[27] Article 29a of the Amendment of the European Parliament to the AIA.

[28] This can happen, for example, if the AI Office guidelines to FRIA (Article 27(5)) will accommodate a broad interpretation of «categories of natural persons and groups likely to be affected by its use in the specific context» (Article 27(1), lit. c).

risks that different groups face[29]. In technology regulation, this approach means identifying how socio-technological systems might inadvertently harm vulnerable populations and implementing measures to mitigate these risks.

This view of vulnerability is also increasingly influencing legal scholars. For example, Gianclaudio Malgieri explored how GDPR provisions and data protection impact assessment can integrate vulnerability as a critical factor in assessing risks and designing protections[30]. Malgieri argues that recognising vulnerability as a condition that can be impacted by data processing activities allows for a more nuanced and effective regulatory response that goes beyond merely preventing exploitation. A similar discussion is now taking place with the AIA's FRIA[31].

A separate question pertains to whether this different meaning of vulnerability refers to the same vulnerable entities as Article 5. Article 9 refers only to minors and "as appropriate" to other "vulnerability groups"; thus, it does neither include "individuals" nor explicitly refer to "social and economic situations". We can assume that the interpretation of "as appropriate" follows the purpose of the high-risk AI system. For instance, in biometric systems, vulnerable groups may be ethnic minorities, who may be disproportionately misidentified due to biases in the training data; in educational systems, minors and disabled; in employment and worker management systems, women; in justice administration, ethnic groups or already convicted persons. Instead, Article 60 refers only to age and disability, thus reflecting Article 5 only regarding personal traits and not situations. Article 79 generally refers to "vulnerable groups" and not "persons". Article 95 refers to "vulnerable persons and groups, including people with disability".

6. Vulnerability as a Power Relation

Finally, a fourth meaning of vulnerability is contained in Article 7(2) of the AIA. Here, vulnerability features are one of the criteria (lit. h) that the European Commission can consider when amending Annex III on high-risk AI system applications. In particular, the Commission can consider «the extent to which there is an imbalance of power, or the persons who are potentially harmed or suffer an adverse impact are in a vulnerable position in relation to the deployer of an AI system, in particular due to status, authority, knowledge, economic or social circumstances, or age».

This fourth account views vulnerability as a power imbalance where the less powerful entity is more susceptible to harm. Martha Fineman's "universal vulnerability approach" offers valuable insight,

[29] P. Blaikie, T. Cannon, I. Davis, B. Wisner, *At risk: natural hazards, people's vulnerability and disasters*, London, 2004.

[30] G. Malgieri, *Vulnerability and Data Protection Law*, Oxford, 2023, where the Author explores the intersection of vulnerability and data protection, arguing for the incorporation of vulnerability as a key consideration in data protection frameworks, and proposing legal mechanisms to better protect vulnerable individuals in the digital age.

[31] G. Malgieri, C. Santos, *Assessing the (Severity of) Impacts on Fundamental Rights*, 25 June 2024, Available at SSRN, https://ssrn.com/abstract=4875937 (last visited 27/07/2024). See, also, A. Mantelero, *The Fundamental Rights Impact Assessment (FRIA) in the AIA: roots, legal obligations and key elements for a model template*, in *Computer Law & Security Review*, 54, 2024, 1-18.

Downloaded from www.biodiritto.org.
ISSN 2284-4503

emphasizing that while vulnerability is a universal human experience, its extent varies and is shaped by social, political, and relational factors[32].

The idea of vulnerability as connected to power is also explored in socio-political literature. For example, political philosophers like Estelle Ferrarese extensively explored the dynamics of power and its relation to vulnerability. In his work, Ferrarese defines vulnerability as "an exposure to another's power to act"[33] and emphasises how power relations are embedded in social structures and institutions, affecting individuals' ability to protect themselves from harm. On similar lines, Judith Butler's concept of "precariousness" also aligns with this view, highlighting how social structures create conditions of vulnerability for certain groups while privileging others[34].

All this perspective underscores the importance of considering if and how AI systems can perpetuate or exacerbate these power imbalances.

The AIA clarifies that relations of vulnerability may derive from positional differences in terms of status, authority, knowledge, economic and social circumstance, or age. Vulnerability by virtue of age appears as in Article 5, although here, what seems to count is the asymmetry of experience rather than the intrinsic cognitive limitations of minors (and adults?). Socio-economic elements bear relevance too, but reference is made to "circumstances" and not to "specific situations", thus suggesting that transient aspects can also matter.

Finally, reference is made to the concepts of "status", "authority" and "knowledge". The three concepts are not defined. Yet, while "status" typically refers to an individual's social or professional position within a hierarchy or society and "authority" relates to the power or right to give orders and enforce obedience[35], "knowledge" pertains to the information, understanding, and skills that different individuals and organisations possess.

Arguably, examples of vulnerable relations depending on "status" and "authority" can be found in Recitals 58 and 60, which motivate the inclusion of AI systems used in essential services and benefits and in migration and border control management in the high-risk class. Here, different categories of people (namely, citizens and migrants) are deemed vulnerable to public entities (namely, public administration for social security and public authorities for border controls), which means that they can suffer negative consequences depending on the outcome of their decision.

As in the case of Article 5, the role AI plays in the vulnerability relation is not clear. Indeed, the way Article 7 is framed seems to look at the vulnerability condition as a characteristic of the relationship between the deployer and the affected person, *regardless* of the use of AI systems. The examples contained in Recitals 58 and 60 again may provide some clarification. Recital 58 highlights that AI

[32] M. FINEMAN, *The Vulnerable Subject*, *op. cit*.; see also R. GOODIN, *Protecting the Vulnerable: A Re-analysis of our Social Responsibilities*, Chicago, 1985

[33] E. FERRARESE, *Vulnerability and Critical Theory*, Leiden, 2018, 12, where the Author argues that vulnerability, as susceptibility to a harmful event, is, above all, a breach of normative expectations. She demonstrates that these expectations are not mental phenomena but are situated between subjects and must even be conceived as institutions.

[34] J. BURLET, *Precarious Life: The Powers of Mourning and Violence*, London, 2004

[35] See, e.g., H.L.A. HART, *The Concept of Law*, Oxford, 1961, 20, where, based on John Austin, Hart ties the concepts of authority and command: «To command is characteristically to exercise authority over men, not power to inflict harm, and though it may be combined with threats of harm a command is primarily an appeal not to fear but to respect for authority».

systems used by public administrations for social security services and benefits create a vulnerable relationship where citizens, due to their dependency on these services, are particularly susceptible to the decisions made by these systems. But do citizens depend on social security services only when AI is involved? Recital 60 addresses the use of AI in migration and border control management: are migrants and asylum seekers relying on public authorities to determine their right to enter or remain in a country only when AI is deployed? In our view, these cases clearly suggest that a relational vulnerability does not originate in the use of AI but in specific power relations – which is, in the end, what the same AIA concludes[36].

From a legal point of view, this notion of vulnerability is as innovative in its conceptual underpinning as it is limited in its application. As said above, relational vulnerability can (not shall) only guide the European Commission to review the list of high-risk AI systems in Annex III limited to areas already present. This means that the relational account of vulnerability does not provide any directly actionable protection to people in a vulnerable relationship.

Article 7(2), however, can play an additional role in shaping the implementation of the AIA: it may serve as a hermeneutic key for national courts and other enforcement authorities to interpret the notion of "high-risk" and the respective use cases, possibly using analogy[37]. This means that high-risk systems included in Annex III are there, also because a vulnerable relation is at play between the deployer and the potentially affected person.

7. The Missing Piece: Vulnerability Stemming from Human-Computer Interaction

Overall, among the many meanings of vulnerability, two stand out[38]: vulnerability as a feature of individuals and groups and vulnerability as a relation between organisations and persons. While the first aligns with traditional legal perspectives, the second offers a more nuanced view by considering power dynamics. In both cases, however, the AIA fails to clarify what exact contribution AI does bring into the picture.

We argue that a diverse, albeit essential, account of vulnerability is missing in the regulatory picture of the AIA: vulnerability as an inherent relation between AI systems and humans. This account shifts the focus from identifying and mitigating individual or situational vulnerabilities to evaluating how AI design and interaction paradigms impact human rights and other fundamental values. Interactional versions of vulnerability are currently discussed in the Human-Computer Interaction (HCI)

[36] Cf. also with the various references in the AIA to «power imbalance», such as in Recital 44 regarding the prohibited use of AI systems to infer emotions in the workplace and in Recital 59 on the use of AI by law enforcement authorities.

[37] This interpretation is supported by the fact that the criteria included in Article 7(2) are similar to the risk criteria used by the European Commission in the Impact Assessment accompanying the AIA to provide evidence for the list of high-risk AI systems included in Annex III. See Commission Staff Working Document Impact assessment – Annexes accompanying the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, SWD/2021/84 final, 40.

[38] We do not consider here AI vulnerability analysed in Paragraph 3.

community[39], but they have been largely neglected in mainstream legal-philosophical analysis, which arguably provided the intellectual milieu for the AIA.

The HCI theory suggests that the interactive design features of an AI system are relevant in establishing vulnerability relations. In HCI, "design" is a multifaceted concept encompassing the aesthetic and functional aspects of AI systems and the cognitive, emotional, and social dimensions of user interaction[40]. It involves creating interfaces and interactions that are intuitive, accessible, and responsive to user needs while considering the broader context in which these systems operate.

To understand better this account of vulnerability, we unpack its main design components: (a) the purpose of the system, (b) context of use, (c) autonomy level, (d) interaction modes, and (e) physical appearance.

First, according to the HCI literature, the purpose for which an AI device is designed and its role is crucial in understanding its impact on user vulnerability. The "purpose" has a social meaning: it abstracts away from specific, fixed and predictable uses (as used for Annex III of the AIA) and includes different "types of social interaction" an AI system is expected to engage with[41].

For example, an AI system may have therapeutic or care purposes, such as supporting mental health, emotional well-being, or physical rehabilitation. It may engage in advisory interactions, providing recommendations and guidance in various information-related contexts. AI systems may also be designed for behavioural change, aiming to influence user habits and decisions, or – the distinction may sometimes be subtle – for interactive and engaging purposes, like entertainment and immersive experiences. Other types include assistive interactions that enhance human capabilities, collaborative interactions that facilitate teamwork and productivity, and monitoring and surveillance interactions that ensure security and health monitoring.

In any case, each of these social purposes may entail its consequences in terms of vulnerable relations. For instance, AI systems delivering therapeutic and care purposes, such as in mental health support, present specific vulnerabilities related to emotional manipulation and dependency[42]. Individuals may develop a deep emotional attachment to AI systems, potentially leading to an over-reliance on these devices for emotional support, which can result in neglecting human relationships and becoming more isolated. Additionally, the sensitive personal data shared during therapeutic sessions may be

[39] See, for instance, the 4TU Virtual Symposium on Vulnerability and Human-Computer Interaction, organised by the University of Twente on 2 December 2021. The only exception in the legal community is provided by the 2nd Conference of the Italian PRIN Project DIVE (Digital Vulnerability in European Private Law) dedicated to Human Vulnerability in Interaction with AI.

[40] See, among others, the J. Preece, H. Sharp, Y. Rogers, *Interaction Design: Beyond Human-Computer Interaction*, Hoboken, New Jersey, 2015. The HCI's view on design has its roots in the socio-materiality of technology and ecological psychology of James J. Gibson, Donald A. Norman, and Jeff Raskin, who emphasised the importance of affordances and user-centred design principles in understanding and improving human interactions with technology. In this regard, we refer to the foundational reading by D. Norman, *The Design of Everyday Things*, New York, 1988.

[41] We follow the approach in C. Burr, N. Cristianini, J. Ladyman, *An Analysis of the Interaction Between Intelligent Software Agents and Human Users*, in *Minds and Machines*, 28, 2018, 735, distinguishing types of interaction with artificial agents based on different types of goals, such as coercion, persuasion, nudging, trading.

[42] A. Fiske, P. Henningsen, A. Buyx, *Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy*, in *Journal of Medical Internet Research*, 21, 5, 2019, e13216.

vulnerable to misuse or breaches, raising significant privacy concerns. A virtual therapist providing cognitive behavioural therapy must be designed to safeguard user data and ensure the integrity of the therapeutic process to prevent harm.

Behavioural change AI systems, which aim to influence user habits and decisions, also introduce particular vulnerabilities. These systems often employ persuasive techniques[43] to motivate users towards specific behaviours, such as adopting healthier lifestyles or making environmentally friendly choices. While beneficial, there is a risk that users may feel manipulated or coerced into behaviours they are not fully comfortable with, potentially undermining their autonomy and consent. Furthermore, the continuous monitoring required for these systems to provide feedback and guidance can raise issues of data sensitivity and privacy. For example, a fitness app that tracks physical activity and provides personalised workout plans must handle user data with utmost care to prevent unauthorised access and ensure user trust[44].

Behavioural change seems to be the only "social function" addressed in the AIA. Article 5(a) deals with the manipulative potential of many AI applications and outlaws using subliminal techniques beyond a person's consciousness. The meaning of "subliminal techniques" is unclear, as it is the meaning of "awareness" concerning practices that operate beyond it. One might wonder if, considering this vague terminology, techniques such as the use of digital architectures to induce certain harmful behaviours in users (dark patterns)[45] or personalised and adaptive recommendations that lead individuals to irrational and impulsive choices (so-called hyper nudging) could be included[46]. The risk is that, although commendable in its objective, the ban on manipulation remains a statement of intent.

Secondly, AI-human vulnerable relations may depend on the context of use. AI integrated into private spaces, like homes, can create intimate relationships with users, leading to high levels of dependency and inner bonding[47]. Research shows how smart home devices that control lighting, heating, and security create a seamless and convenient living environment but also pose risks to privacy and data security.

In contrast, AI systems in public or semi-public spaces, such as schools, workplaces, or hospitals, interact with a broader user base and must confront varying levels of trust and dependency, which are

[43] B.J. Fogg, *Persuasive technology: using computers to change what we think and do*, Ubiquity, 5, 2002, 89 ss.. More recently, the edited book by P. de Vries, H. Oinas-Kukkonen, L. Siemons, N.B.D Jong, L. van Gemert-Pijnen (eds.), *Persuasive technology: Development and implementation of personalized technologies to change attitudes and behaviors,* Berlin/Heidelberg, 2017.

[44] See, for instance, E. A. Edwards, J. Lumsden, C. Rivas, L. Steed, L. A. Edwards, A. Thiyagarajan, R. Sohanpal, H. Caton, C. J. Griffiths, M. R. Munafò, S. Taylor, *Gamification for health promotion: systematic review of behaviour change techniques in smartphone apps*, in *BMJ Open*, 6, 10, 2016, e012447.

[45] For an interpretation of Article 5 AIA in light of dark patterns-types of influence, see the recent piece by M. Leiser, *Psychological Patterns and Article 5 of the AIA: AI-Powered Deceptive Design in the System Architecture and the User Interface*, in *Journal of AI Law and Regulation*, 1, 1, 2024, 5.

[46] S. Faraoni, *Persuasive Technology and computational manipulation: hypernudging out of mental self-determination*, in *Frontiers in Artificial Intelligence*, 6, 2023, 1216340.

[47] H.R. Pelikan, M. Broth, *Why that now? How humans adapt to a conventional humanoid robot in taking turns-at-talk*, in *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, May 7-12, 2016), 2016, 4921, in which the Author examines the interaction dynamics and the adjustments humans make in response to the robot's timing and conversational cues, thereby providing insights into the challenges and nuances of human-robot communication in social contexts.

influenced by the institutional context. For instance, educational AI tools that assist in personalised learning can significantly impact student performance and engagement but also raise concerns about conforming practices to algorithm-driven learning paths and instil an increased feeling of loneliness and a lessened sense of belonging to a learning group[48].

Following an HCI account of vulnerability, a third determinant of vulnerable relations is the degree of autonomy granted to AI systems versus the level of human supervision.

Highly autonomous AI, such as self-driving cars, require users to place significant trust in the system's decision-making capabilities, which can be both empowering and anxiety-inducing[49]. For example, research shows that while autonomous systems can increase efficiency and convenience, they also raise concerns about accountability and the potential for errors[50]. Conversely, AI systems with substantial human oversight, like decision-support tools in medical settings, ensure a higher degree of control and reliability but may also suffer from reduced efficiency and increased cognitive load on human operators, which might also be described in terms of vulnerability[51].

The extent to which such nuances will be considered in the implementation of the AIA is not clear. As known, the definition of "artificial intelligence" in the Regulation contemplates machine-based systems with various levels of autonomy. However, the autonomy level does not seem to be directly correlated with the stringency of the regulatory measures imposed[52]. This raises questions about whether the specific challenges and vulnerabilities associated with highly autonomous systems are being adequately addressed in the regulatory framework.

Fourthly, the modes of interaction between AI systems and users—verbal, visual, physical, or a combination thereof—play a pivotal role in shaping the user experience and associated vulnerabilities. Verbal interactions, facilitated by voice assistants, can create a sense of conversational ease and familiarity but also introduce risks related to misinterpretation and the nuances of human language[53]. Visual interaction modes, such as those used in augmented reality and virtual reality, offer immersive

[48] P. Prinsloo, M. Khalil, S. Slade, *Vulnerable student digital well-being in AI-powered educational decision support systems (AI-EDSS) in higher education*, in *British Journal of Educational Technology*, 5, 2024, 2075 ss.

[49] P.A. Hancock, *Avoiding adverse autonomous agent actions*, in *Human–Computer Interaction*, 37, 3, 2022, 218. The Author offers the metaphor of "isles of autonomy", illustrating how autonomous systems may initially be supported by human operators, but over time, they are expected to become increasingly independent and integrated, reducing the need for human intervention.

[50] R.G. Dutta, X. Guo, Y. Jin, *Quantifying trust in autonomous system under uncertainties*, in *29th IEEE International System-on-Chip Conference (SOCC)*, 2016, 362.

[51] S. Daronnat, L. Azzopardi, M. Halvey, M. Dubiel, *Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration*, in *Frontiers in Robotics and AI*, 8, 2021, 642201.

[52] The only two exemptions are provided by the possibility of the provider to self-exempt from high-risk category pursuant the presumption stated in Article 6(3), for example, when the «AI system is intended to perform a narrow procedural task» (lit. b) and by the possibility granted by Article 7(2) to the Commission to amend Annex III also considering «the extent to which the AI system acts autonomously and the possibility for a human to override a decision or recommendations that may lead to potential harm» (lit. d).

[53] H.A. Voorveld, T. Araujo, *How social cues in virtual assistants influence concerns and persuasion: the role of voice and a human name*, in *Cyberpsychology, Behavior and Social Networking*, 23, 10, 2020, 689.

experiences that can enhance learning and entertainment but may also lead to over-reliance on virtual environments and potential disconnection from reality[54].

Finally, HCI literature has long stressed that the physical appearance of computational systems bears relevance in determining when a vulnerable human-AI relation exists[55]. The physical appearance of AI systems, whether embodied or disembodied, significantly influences the "social role", reliability, and the bond that individuals form with these systems. Embodied AI, such as humanoid robots, can evoke strong emotional responses and social bonding due to their human-like features[56]. Conversely, disembodied AI, like virtual assistants (e.g., Siri or Alexa)[57] and chatbots (e.g., ChatGPT, Gemini, or Claude), may foster a different type of interaction that relies on the perceived intelligence, responsiveness, and personalisation of the system rather than its physical presence. These systems often employ sophisticated conversational interfaces that create an illusion of understanding and empathy, leading users to engage with them as if they were interacting with a knowledgeable and reliable companion. This form of impersonation, where the AI mimics human-like conversational skills, can generate a sense of trust and emotional connection despite the absence of a physical body[58].

This idea of vulnerability as deception is limitedly expressed in the AIA. Only Article 50, containing transparency obligations for some AI systems considered as "low-risk", accepts that vulnerability may originate from the deceiving effect of human-like, anthropomorphised interactions. The provision requires that users be informed when they interact with an AI system rather than a human being to prevent deception and undue emotional attachment.

At the same time, Article 50 takes an optimistic stance on transparency, which contrasts with insights from HCI literature. This suggests that automating human likeness poses ethical and social questions

[54] D. van Heugten-van der Kloet, J. Cosgrave, J. van Rheede, S. Hicks, *Out-of-body experience in virtual reality induces acute dissociation*, in *Psychology of Consciousness: Theory, Research, and Practice*, 5, 4, 2018, 346.

[55] L.R. Caporael, *Anthropomorphism and Mechanomorphism: Two Faces of the Human Machine*, in *Computers in Human Behavior*, 2, 3, 1986, 215.

[56] For instance, in marketing studies, anthropomorphism is typically leveraged to entice an empathetic stance over clients and a heightened predisposition to buy. See, P. Aggarwal, A.L. McGill, *Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products*, in *Journal of Consumer Research*, 34, 4, 2007, 468.

[57] We recall, for example, the first announcement by Amazon in September 2019 on the new improvements to Alexa's voice, including the new celebrity-guest-voice skill featuring Samuel L. Jackson's voice. C. Gartenberg, *All the new features are coming to Alexa, including a new voice, frustration mode, and Samuel L. Jackson*, in *The Verge*, January 2019, https://www.theverge.com/2019/9/25/20883751/amazonalexa-voice-languages-natural-bi-lingual-frustration-support-new-features (last accessed 28/07/2024).

[58] See, among others, E. Go, S.S. Sundar, *Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions*, in *Computers in Human Behavior*, 97, 2019, 304. The study highlights several interesting findings, including the "compensation effect", where high anthropomorphic visual cues can make up for low message interactivity and vice versa. It also identifies an "expectancy violation" effect when identity cues are paired with interactive messaging, suggesting that revealing the chatbot's non-human identity can either meet or disrupt user expectations, depending on how it is communicated.

that go well beyond merely informing users[59]. Transparency could ultimately be counterproductive for AI deployers, as it risks endangering user engagement and trust[60].

8. Room for Manoeuvre: Looking at HCI as a "specific social situation"

The absence of an HCI perspective on vulnerability in the AIA does not preclude the possibility of interpreting and enforcing its provisions through such a lens. In fact, an HCI outlook on vulnerability is compatible with the AIA's fundamental view of AI as a product. Integrating HCI perspectives can enrich the understanding and regulation of AI systems, ensuring that they are designed and deployed in ways that prioritise user well-being and safety. The AIA predominantly treats AI as a product, focusing on the technical specifications, risk management, and compliance measures required to ensure its safe use. In this context, the HCI perspective brings to the forefront the interactions between humans and AI systems, emphasising the importance of design features and user experiences in shaping vulnerability. Consequently, viewing AI through the lens of HCI enriches the product-oriented approach by also considering AI as a service.

In the previous paragraph, we pointed out some provisions of the Regulation that may accommodate an HRI view of vulnerability. We argue now that an opening point in the AIA that allows the re-incorporation of a more structured view of AI-human interaction vulnerability is the concept of "specific social situation" contained in Article 5.

In sociology, a "social situation" is variously referred to as the condition in which individuals interact and form relationships[61]. For example, referring to the social situation of people with a mental health condition in the 60s American society, the famous sociologist Erving Goffman described "social situations" as structured interactions where individuals manage their self-presentation and deal with the expectations of near others[62]. This is part of what Goffman famously coined as the "interaction order", an order which includes the norms that dictate how people present themselves and respond to others in various contexts.

[59] J. PORRA, M. LACITY, M.S. PARKS, *Can Computer Based Human-Likeness Endanger Humanness? – A Philosophical and Ethical Perspective on Digital Assistants Expressing Feelings They Can't Have*, in *Information Systems Frontiers*, 22, 2020, 533.

[60] For example, this may happen in business-consumer relations, where research suggests that undisclosed chatbots are as effective as proficient workers and four times more effective than inexperienced workers in engendering customer purchases and that a disclosure of chatbot identity before the machine–customer conversation reduces purchase rates by more than 79.7%. See, X. LUO, S. TONG, Z. FANG, *Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases*, in *Marketing Science*, 38, 6, 2019, 937.

[61] It is beyond the scope of this discussion to reference the extensive literature on social situations, which encompasses seminal works such as Harold Garfinkel's ethnomethodological studies, Georg Simmel's analysis of social forms, and significant contributions from scholars including Herbert Blumer, Alfred Schutz, Pierre Bourdieu, and Norbert Elias, among others.

[62] E. GOFFMAN, *Asylums. Essays on the social situation of mental patients and other inmates*, New York, 1961, 144: «By the term social situation I shall refer to the full spatial environment anywhere within which an entering person becomes a member of the gathering that is (or does then become) present. Situations begin when mutual monitoring occurs and lapse when the next to last person has left».

Perhaps one of the most analytical account of social situations was given by social psychologists Michael Argyle, Adrian Furnham, and Jean Ann Graham[63], who describe social situations as comprising several key features: 1) goals of persons, 2) rules, that is, the beliefs that regulate peoples' behaviours within the situation; 3) roles defining the expected behaviours and responsibilities; 4) repertoire of elements relevant to the goals; 5) sequence of behaviour that need to be completed in a particular order; 6) shared concepts necessary for managing tasks and achieving goals; 7) environmental setting; i.e. physical environment, including boundaries, props, and modifiers, which influences behaviour and interaction in a situation; 8) language and speech, with specific vocabulary and speech patterns that may need to be adapted based on the context; 9) difficulties and skills, i.e. some situations require specific social, perceptual, or linguistic skills, and the challenges faced in these contexts can offer insights into social interaction processes.

Following this framework, an AI-human interaction can be effectively analysed as a social situation.

The AI-human interaction sets roles between the AI system and the human user. For example, an AI might assume the role of an advisor, assistant, or companion, while a human may take on the role of a decision-maker, dependent user, or learner. These roles come with specific expectations and responsibilities, much like roles in traditional social situations, influencing how the interaction unfolds and how the user perceives the AI system's capabilities and trustworthiness.

Humans engage with AI systems to achieve specific objectives like obtaining information or completing tasks through specific interaction sequences. As seen in the previous paragraph, these goals or modes of interaction, as well as the predictability or variability of these users' behavioural sequences, can introduce potential vulnerabilities, especially in those relations where humans become overly reliant on the AI system.

Environmental settings and language and speech are also relevant in AI-human interactions. The virtual environment or interface in which the interaction occurs can affect the user experience, just as the physical setting influences traditional social situations. Language use, including vocabulary and tone, is tailored to the interaction context, whether formal, casual, or technical, and can vary widely depending on the user's expectations and the AI's design.

The reference to a "specific social situation" in Article 5 of the AIA is sufficiently flexible to accommodate a tailored assessment of AI-human interactions. When providers and deployers must comply with the prohibition, namely assess when the AI systems exploit vulnerabilities due to a specific social situation, they may collaboratively consider aspects of vulnerability and assess the level of vulnerability in AI-human relations. This extensive interpretation of Article 5, lit. b) allows the reincorporation of an HCI view of vulnerability into the AIA.

Following such an approach, providers and deployers could be required to adopt measures to mitigate vulnerability by focusing on design features and interaction paradigms. While we cannot elaborate here on the exact nature and details of such measures, they may involve the continuous monitoring and evaluation of AI systems to understand their interactions with persons and the social situations they create. Contextual analysis of the specific environments in which AI systems are deployed and user behavioural patterns should be deemed essential to appropriately tailor design and regulatory responses.

[63] M. Argyle, A. Furnham, J.A. Graham, *Social Situations*, London, 1981.

9. Conclusion

In this paper, we reviewed the different meanings of vulnerability contained in the AIA. Our analysis reveals that the Act predominantly aligns with risk science literature, where vulnerability is seen as a factor influencing the overall magnitude of risk associated with an event. Additionally, the Act reflects an established tradition of viewing vulnerability as a trait or state of certain individuals and groups. This traditional perspective considers vulnerability as an inherent characteristic of specific demographics, such as the elderly, children, or economically disadvantaged groups, who are more susceptible to harm due to their particular conditions. The AIA also incorporates a promising notion of vulnerability as a relational concept, recognising that vulnerability can arise from the power dynamics between organisations and individuals, such as the dependency of citizens on public administrations for social security services or the precarious position of migrants in relation to border control authorities. However, the AIA falls short of clarifying the specific role AI plays in these interactions and how it may alter the dynamics of vulnerability.

We identified a critical missing meaning in the AIA: vulnerability as an intrinsic feature of all AI-human relations, which manifests depending on different design features and interaction modes. This perspective extends beyond the traditional meaning of vulnerability as merely an inherent trait or a relational dynamic and considers how the design and deployment of AI systems themselves can create or exacerbate vulnerability. Factors such as the purpose of the interaction, the context of use, the mode of interaction, the autonomy of the AI system, and the physical appearance of systems may contribute to determining the extent to which users may become vulnerable when engaging with these systems. Finally, we proposed that this different meaning of vulnerability can be integrated into the current text of the AIA by interpreting the construct of "specific social situation" in Article 5, lit. b) more broadly. By expanding this interpretation to cover the specific contexts and interaction paradigms facilitated by AI systems, the AIA can more effectively address the nuances of vulnerability in AI-human interaction. This holistic approach would not only protect traditionally vulnerable groups but also recognise and mitigate the new forms of vulnerability emerging from constituting relations with advanced AI technologies. In the future, this integration may prove essential for creating norms that ensure equitable deployment of AI systems and pay respect to the inherently weaker human conditions, especially vis-à-vis certain AI advanced technologies.