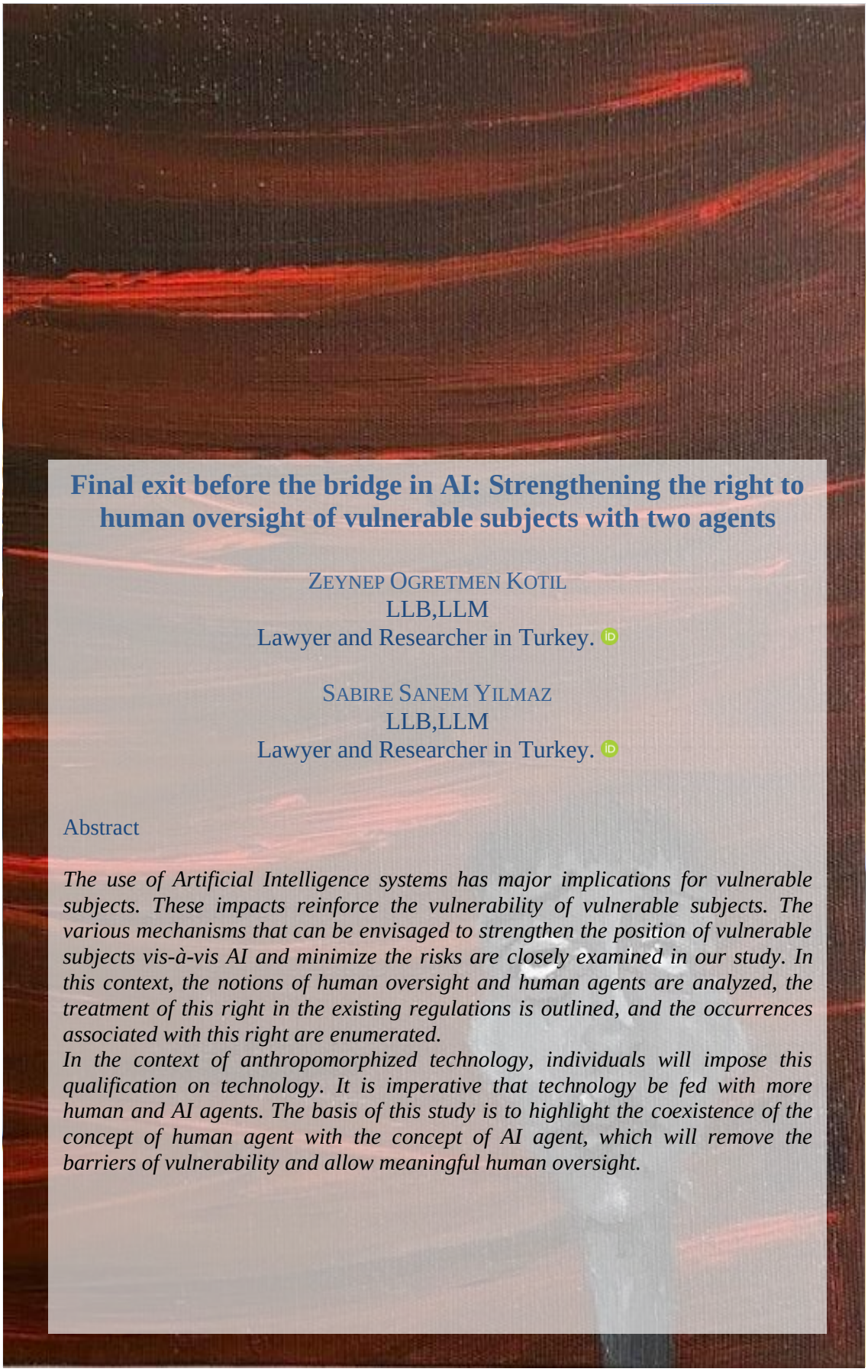
S Sanem Yılmaz, Zeynep Ö Kotil, 'Final exit before the bridge in AI: Strengthening the right to human oversight of vulnerable subjects with two agents' (February 2024) EJPLT Observatory. Available at: www.ejplt.tatodpr.eu

EJPLT
ISSN 2704-8012



Final exit before the bridge in AI: Strengthening the right to human oversight of vulnerable subjects with two agents

ZEYNEP OGRETMEN KOTIL
LLB,LLM
Lawyer and Researcher in Turkey.

SABIRE SANEM YILMAZ
LLB,LLM
Lawyer and Researcher in Turkey.

Abstract

The use of Artificial Intelligence systems has major implications for vulnerable subjects. These impacts reinforce the vulnerability of vulnerable subjects. The various mechanisms that can be envisaged to strengthen the position of vulnerable subjects vis-à-vis AI and minimize the risks are closely examined in our study. In this context, the notions of human oversight and human agents are analyzed, the treatment of this right in the existing regulations is outlined, and the occurrences associated with this right are enumerated.
In the context of anthropomorphized technology, individuals will impose this qualification on technology. It is imperative that technology be fed with more human and AI agents. The basis of this study is to highlight the coexistence of the concept of human agent with the concept of AI agent, which will remove the barriers of vulnerability and allow meaningful human oversight.

© The author(s) 2023, published by Suor Orsola Benincasa Università Editrice.
This contribution is licensed under a Creative Common Attribution 4.0 International Licence CC-BY-NC-ND, all the details on the license are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Keywords: Vulnerability; Human Oversight; Human Agent; Sense of Agent; Anthropomorhized Technology.

Summary: Introduction: Vulnerable Subjects, Vulnerability, AI - Defining the Vulnerability from Zero to +1. – 1. The Interactions of Vulnerable Subjects with AI – 2. The Key Element: Human Oversight – 2.1. What is "Human Oversight"? – 2.2 Reflections of Human Oversight in Legal Regulations – 2.2.1. Reflections in the GDPR – 2.2.2. Reflections in the AI Act – 3. The Phenomena That Best Define the Right: Transparency and Accountability – 4. Would it be possible to regulate the right of vulnerable subjects to request human oversight? – 4.1. Analysis of the Current Situation – 4.2. A Fundamental Rights Advocacy on Vulnerability: From Human Agent to Human Oversight – Conclusions.

Introduction: Vulnerable Subjects, Vulnerability, AI - Defining the Vulnerability from Zero to +1

The Oxford English Dictionary defines "vulnerability" as the quality or state of being exposed to the possibility of being attacked or harmed, either physically or emotionally. [1] We would like to assert that we are discussing a significant and comprehensively defined vulnerability. We encounter vulnerable people at every stage of our lives, but some are particularly vulnerable. Furthermore, we know that children, women who are exposed to violence, people who are constantly discriminated against, and people with disabilities are among the most vulnerable groups.[2] On the contrary, technology is progressively enhancing the variety and extent of vulnerable groups on a daily basis, and the escalating economic and social disparities that are causing this increase are a result of the digitization of the world.[3] It is challenging to establish the boundaries of vulnerability in the digital world, and we possess the expertise to mitigate it more effectively than the decisions made by the authorities. To be a human being is to be a particular kind of animal whose existence and flourishing depend on social relations with others, including relations of care. When one disregards the notion of vulnerability in society and fails to incorporate it into their political and legal regulations, they are also imposing a significant burden on society. This responsibility drives society to escape from vulnerability or being human.[4]

While the state remains the primary mechanism through which humans are most vulnerable, the situation has shifted towards Artificial Intelligence (AI). The advancement of AI technology renders humans more vulnerable than the state. Based on the findings of the Belmont Report and the Helsinki Declaration, it is evident that the utilization of vulnerable groups as subjects, particularly in clinical research, has been a significant challenge to the notion of exploiting their weaknesses.[5] It is noteworthy that

[1]Oxford Reference, <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803120303277> accessed 8 December 2023.
[2] MA Fineman, 'Vulnerability and Inevitable Inequality' (2017) 4(3) Scandinavian University Press 133-149.
[3] P Rebughini, 'A Vulnerable Generation? Youth Agency Facing Work Precariousness' (2019) 1 Papeles del CEIC 1-17.
[4] ibid 1-17.
[5] G Malgieri, J Niklas, 'Vulnerable Data Subjects' (2020) Special Issue on Data Protection and Research <https://ssrn.com/abstract=3569808> accessed 10 December 2023.

in a world where it holds little significance to remind individuals who have been denied their opportunities, it is a dream that they strive to realize in the age of technology. The more serious problem is that vulnerable subjects are open to more manipulation while making the most of it. Individuals tend to avoid following paths they are unfamiliar with; however, an algorithm can effectively direct individuals towards a distinct route while presenting the impression that they are following their familiar route. Many people who follow this path are unaware that personal data, such as track, walking mode, and heart rate, are recorded in this way.

For instance, we inquire about the rationale behind the vulnerability experienced by individuals with mental disabilities. It can be asserted that a person with a mental disability lacks three essential abilities: intuitive analysis, critical thinking, and creative problem-solving. If we were to take these abilities from a person who does not possess mental disabilities, they would eventually lose their ability to make decisions.[6]

According to the 2001 report, Ethical and Policy Issues in Research Involving Human Participants[7], people do not have the will to make independent decisions in scientific research and are even more vulnerable. In addition to the traditional definition of vulnerability, we will also deliberate on the significance of human oversight and agency in assessing the increasing vulnerability of individuals and their interdependence, as well as the evolution of vulnerability in the past and present.

AI systems are designed to mimic the human intelligence and make decisions[8]. Despite the long-held belief that AI systems would eventually replace humans in decision-making, contemporary research has revealed that human cognition is reliable, versatile, and efficient despite its shortcomings[9]. The era of AI will necessitate the enhancement of human agency and the right to human oversight, which necessitates individuals' vulnerability as a prerequisite for their right to self-determination in the neoliberal environment. The vulnerability of an individual who lacks any form of intervention or control over their age will emerge as a vulnerability. Hence, it is imperative to note that within the regulations of this domain, individuals are susceptible to vulnerability. In this context, it will be imperative to enhance the capacities of subjects who are subject to gender conflicts, specifically women and individuals who have been exploited, and to mitigate the agencies' vulnerability.[10]

In the "Human, Vulnerability and Personal Integrity" report of the UNESCO International Bioethics Committee:[11]

---

[6] B Berberian, M Pagliari, V Chambon, 'What is new with Artificial Intelligence? Human–agent interactions through the lens of social agency' (2022) Front Psychol. 13 <https://doi.org/10.3389%2Ffpsyg.2022.954444> accessed 10 December 2023.

[7] Report and Recommendations of the National Bioethics Advisory Commission, 'Ethical and Policy Issues in Research Involving Human Participants' (2001) <https://bioethicsarchive.georgetown.edu/nbac/human/overvol1.pdf> accessed 10 December 2023, 1-228.

[8] B Leong, 'The Threat of AI: Fact or Fiction', Human Decisions Thoughts on AI (Unesco Publishing, 2018) 60; High-Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines' (2019) <https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines> accessed 8 December 2023, 3-4.

[9] CH Besseyre Des Horts, 'Will robots soon replace humans?', Human Decisions Thoughts on AI (Unesco Publishing, 2018) 64-65; R Srinivasan, 'The Ethical Dilemmas of Artificial Intelligence', Human Decisions Thoughts on AI (Unesco Publishing, 2018) 106-107.

[10] A Ariza-Montes, H Han, M Arraño-Muñoz, M Irshad, MK Rehmat, MM Alam, SF Ahmad, 'Impact of artificial intelligence on human loss in decision making, laziness and safety in education' (2023) 10(1) Humanities and Social Sciences Communications <doi: 10.1057/s41599-023-01787-8> accessed 10 December 2023.

[11] International Bioethics Committee, 'The Principle of Respect for Human Vulnerability and Personal Integrity: report of the International Bioethics Committee of UNESCO (IBC)' (2013) <https://unesdoc.unesco.org/ark:/48223/pf0000219494> accessed 10 December 2023.

"Elderly people may be more vulnerable because of their reduced physical and sometimes mental capacities. Persons with disabilities need help to access and sustain the exercise of their self-determination. Persons with mental disorders may not be able to defend themselves or claim their rights. These can all be considered as "natural" determinants of special individual vulnerabilities."

In UNESCO's Universal Declaration of Bioethics and Human Rights, Article 5 of the declaration on the origin of vulnerable groups is referred to under the heading of autonomy and individual responsibility. Accordingly, "*The autonomy of persons to make decisions, while taking responsibility for those decisions and respecting the autonomy of others, is to be respected. For persons who are not capable of exercising autonomy, special measures are to be taken to protect their rights and interests."* And crucial Article 8 says: *"In applying and advancing scientific knowledge, medical practice and associated technologies, human vulnerability should be taken into account. Individuals and groups of special vulnerability should be protected, and the personal integrity of such individuals respected."*[12]

The McKinsey Global Institute Report discusses the fact that 30% of jobs in America have already been automated and that 12 million people will be unemployed by 2030. It was said that the most damaged were blacks, Spanish workers, non-graduates, and women.[13]

How does Habermas' concept of a "reasonable person" relate to the concept of a "vulnerable subject" in AI society? It is now quite easy to manipulate even a person with full willpower with subliminal techniques, but it has also been the subject of the AI Act, which states that manipulating a vulnerable subject with the missing side increases the risk.[14] For this reason, the criteria determined by Habermas for a reasonable person are defined as people or groups at risk in the AI Act. One of the discussions that were held during the draft stage of the AI Act centered on the issue of children's opposition to systems that disrupt the awareness of individuals with disabilities. It has been stated that these systems are completely prohibited. In this context, the notions of disruption and deformity of consciousness are enumerated. It is customary for vulnerable subjects who are in their unconsciousness to exhibit greater deformities. The circumstance to be considered here is that of practices that deform the consciousness of people we do not call vulnerable, albeit temporarily. The reason we draw attention to this issue is that it is imperative to revisit the notion of vulnerability for systems that are beyond the scope of previous knowledge. Individuals who are exposed to systems that infiltrate consciousness ought to be deemed vulnerable as well.

Our greatest experience in the recent past was closing the COVID-19 process and the vulnerabilities created by this process. We have witnessed a period in which the vulnerabilities of countries with low numbers of patients with a social understanding of the state and welfare level have increased further. It is necessary to read this period well in terms of the fundamental rights and freedoms of vulnerable subjects. When the number of vulnerable subjects increased even more during this period, health sector workers who faced difficulties in material supply gained first-order vulnerable subject status. When the vulnerabilities of high-risk systems reach a level that will harm them,

---

[12] ibid

[13] McKinsey Global Institute, 'Generative AI and Future of Work in America' (2023) <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america> accessed 10 December 2023.

[14] JJ Chriss, 'Review essay of Jurgen Habermas's Between Facts and Norms' [1998] Sociology &Criminology Faculty Publications 100.

or when they are forced to use the system, users may lose their decision-making mechanisms. Let us consider the order of an important drug in the health sector, and the procurement process of this drug is managed by an algorithm. In the scenario where this algorithm makes a biased decision and does not transport the drug to a black person, it is not difficult to predict how serious the situation will be. [15]

When we consider the concept of vulnerability, we should focus on a wider range, unlike legal regulations. Current uncertainty could distract us from the wider assessment of vulnerability, focusing on the rights of the real vulnerability, but also clarifying the definition of vulnerability in AI. When we consider the concept of vulnerability, we should focus on a wider range, unlike legal regulations. Current uncertainty could distract us from the wider assessment of vulnerability, focusing on the rights of the real vulnerability, but also clarifying the definition of vulnerability in AI. [16] A comprehensive approach requires that we recognize the ways in which power and privilege are conferred through the operation of societal institutions, relationships, and the creation of social identities, sometimes in an unjust manner. Since the law should recognize, respond to, and perhaps redirect unjustified inequality, the critical issue must be whether the balance of power struck by the law was warranted. In the era of AI, vulnerability will also entail more marginalization, social injustice, inequality, discrimination, and unemployment.

Certain individuals are prone to vulnerability due to their affiliation with a particular group. This is effortless for us to guess. For instance, children fall under this category, and to mitigate their vulnerability, the General Data Protection Regulation (GDPR) mandates the use of language that is child-friendly. We believe that vulnerability is layered in these texts, where parents are empowered with the rights of their children. For instance, it is evident that the child is innately vulnerable; however, what if the parent is also vulnerable at this juncture? The task of closing the vulnerability is not always an easy one.[17]

Vulnerable statuses, as determined by the IRB-Social and Behavioral Sciences Department of the University of Virginia, are important in the context of our expanded status. There are 8 categories: *"Cognitive and Communication; Institutional; Deferential; Medical; Economic; Social; Legal; and Study."* When examining all categories of vulnerability, it is evident that they will result in harm to a significant number of vulnerable species.[18]

1. The Interactions of Vulnerable Subjects with AI

The most significant concern is the way in which vulnerable subjects will interact with AI and benefit from the right to human oversight. If AI is not specifically designed to facilitate vulnerable subjects' interactions or ensure continuity subsequent to interactions, it will result in inequalities and discrimination. This situation, which will lead to the development of innovation against vulnerable subjects who cannot benefit

[15] MA Fineman, 'Beyond Identities: The Limits of an Antidiscrimination Approach to Equality' (2012) 92(6) Boston University Law Review 1750.
[16] R Calo, 'Privacy, Vulnerability, and Affordance' (2017) 66(2) DePaul Law Review 592-593.
[17] J Chen, S Piasecki, 'Complying with the GDPR when vulnerable people use smart devices' (2022) 12(2) International Data Privacy Law 113-131.
[18] 'Vulnerable Participants', <https://research.virginia.edu/irb-sbs/vulnerable-participants> accessed 10 December 2023.

from AI on an equal basis, will trigger an increase in the gap in society.[19] Let us examine the IMPACT and HireVue programs utilized in Washington DC schools. It is conceivable that a vulnerable subject may engage with these programs, resulting in the termination of a teacher's duties as they are unable to effectively utilize AI programs.[20] COMPAS, an AI system used in the United States, is a program developed to predict whether a criminal will commit a crime in the future. If the vulnerable person is not aware of how to interact with this program, they are likely to be discriminated against. These issues have occurred numerous times.[21]

Interaction is frequently described as a mutual relationship, a mutual feeding, and a receiving-giving relationship.[22] The human element is absolutely necessary for effective human-machine interaction. While it is still debatable how humans interact with machines, the cognitive part in particular outweighs the interaction. In this context, it is observed that the user-to-user definitions of a machine as being "social" or "intelligent" undergo alterations. The process that commenced with the interaction with voice assistants such as Alexa and Siri enabled humans to construct a mutual existence through the incorporation of anthropomorphized technology. The important says is: "*What is happening is that on the one hand there is a tendency to subjectivise natural elements, and there is a strong movement for the recognition of a new legal status for the animal, but at the same time there is a strong tendency towards the subjectivisation of humanoid robots or in general of embodied algorithms that represent expressions of the highest levels of artificial intelligence, anthropomorphic machines that represent the pinnacle of anthropocentrism not only legal but philosophical and ontological in general, in the sense that everything that resembles man is supposed to become a legal subject*."[23]

Empirical research has indicated that*: "Humanlike engagement patterns with technological entities also depend on humans themselves. The patterns depend, for example:*

- *on humans' own gender or ethnicity, and how it matches technology's embodied gender or ethnicity (e.g., (Eyssel et al., 2012; Kamide et al., 2013));*
- *on humans' personalities (e.g., (Walters et al., 2008));*
- *on humans' socio-cultural backgrounds (e.g., (Evers et al., 2008));*
- *on humans' prior experience with technology (e.g., (Goudey & Bonnin, 2016));*
- *on humans' needs and feelings for sociality or control (e.g., (Wang, 2017; Eyssel & Reich, 2013; Kim & McGill, 2011; Waytz et al., 2010a));*
- *on humans' age (e.g., (Kamide et al., 2013)).*"[24]

---

[19] E Hermann, G Yalcin Williams, S Puntoni, 'Deploying artificial intelligence in services to AID vulnerable consumers' (2023) Journal of the Academy of Marketing Science < https://doi.org/10.1007/s11747-023-00986-8> accessed 10 December 2023.

[20] F Fossa, G Tamburrini, V Schiaffonati, Automi e Persone - Introduzione Al'Etica Dell'Intelligenza Artificiale e Della Robotica (Studi Superiori 2021).

[21] European Economic and Social Committe, 'A guide to Artificial Intelligence at the workplace' (2022) <https://www.eesc.europa.eu/sites/default/files/files/qe-03-21-505-en-n.pdf> accessed 10 December 2023.

[22] G Assadi, J Seifert, O Friedrich, S Gerlek, S Schleidgen, 'The concept of "interaction" in debates on human–machine interaction' (2023) 10 Humanities and Social Sciences Communications 551.

[23] Gatt L, 'Legal anthropocentrism between nature and technology: the new vulnerability of human beings' (2022) 1 EJPLT, 15-26.

[24] I Siraj, J Festerling, 'Anthropomorphizing Technology: A Conceptual Review of Anthropomorphism Research and How it Relates to Children's Engagements with Digital Voice Assistants' (2021) 56 Integrative Psychological and Behavioral Science 709–738.

We need to emphasize that agents are essential to interacting with AI. Any vulnerability in the human ability to survive is also fed by various agents. This will also be the case when they interact with AI. The feeling that when interacting with an AI, the control is entirely in itself or that when control comes out of itself, the feeling of being distracted from it is the ability of the human sense of agent.[25] Understanding and minimizing the damage that will happen to a human is an ability that nourishes and flexes the agent's feelings. Individuals who interact with AI, such as lawyers or non-governmental organizations that defend the rights of vulnerable subjects, share or hand over their agent characteristics to AI. It is also possible to say that the vulnerability will be shared with AI agents.[26] In the case of the Voodoo doll, the flexibility of people's agent emotions and their efforts to mitigate vulnerability through this flexibility reveal their ability to adapt to the object in their communication with the object.[27] In the interaction of vulnerable subjects with AI, the assistance provided by both AI agents and human agents to these agent elements enables the precise determination of the duration of human oversight and the elimination of meaningless human oversight from the system.[28]

The most influential theory for developing a sense of agent is the Comparator Theory. Human interaction begins with actions and intentions, and motor control continues with the system's understanding.[29] At this point, agent ability can be felt on the system as a result of matching system outputs with human senses.[30] This requires that AI agents and human agents work together to ensure that vulnerable people feel the agent's capabilities when they are left vulnerable by the system. That demonstrates the necessity of interaction.

It is necessary to keep situations of necessity created by power, such as a public or worker-employer relationship, separate from each other to interact with vulnerable subjects and interact with artificially intelligent systems. Apart from the phenomena such as discrimination and inequality caused by the damage to AI from its own nature, we are trying to convey the design of the system for vulnerable subjects to balance the vulnerability with various examples.[31] If the system is specifically designed to serve vulnerable subjects, it would be a timely approach to the advancement of the in-system agents by assessing the impact of the vulnerable subjects and concluding the training of the agents in this direction.[32]

In the interaction of a vulnerable subject with AI, the human-like characteristics of the latter, which distinguish it from humans, certain features that influence the level of social interaction, and the degree of anthropomorphism increase or decrease, have a

[25] H Katsui, M Heikkilä, M Mustaniemi-Laakso, 'Vulnerability, Disability, and Agency: Exploring Structures for Inclusive Decision-Making and Participation in a Responsive State' (2023) 36 International Journal for the Semiotics of Law 1581–1609.

[26] ibid.

[27] JW Moore, 'What Is the Sense of Agency and Why Does it Matter?' (2016) 7(1272) Front Psychol <doi: 10.3389/fpsyg.2016.01272> accessed 10 December 2023.

[28] BERTOLINI, Andrea; CARLI, Rachele. Human-robot interaction and user manipulation. In: International Conference on Persuasive Technology. Cham: Springer International Publishing, 2022. p. 43-57.

[29] CD Frith, DM Wolpert, SJ Blakemore, 'Abnormalities in the awareness and control of action' (2000) 355(1404) Philos. Trans. R. Soc. Lond. B Biol. Sci 1771–1788.

[30] Moore (n 26).

[31] Gatt, Lucilla; Cristina Gaeta, Maria; Aulino, Livia; Grimaldi, Paola, The vulnerability of the human being in a technological environment: the need for protective regulation, 2021, 11

[32] The concept of human vulnerability in the technological environment has been analysed during the PROTECH-Jean Monnet Chair project activity: https://www.protech-jeanmonnet.eu/.

direct impact on the degree of vulnerability in the vulnerable subjects. Naturally, an individual will possess the ability to communicate more effectively with an individual possessing similar traits or to instill confidence in an individual possessing these traits.[33] When we enhance the anthropomorphization of AI in a manner that allows vulnerable subjects to interact with it, we also enhance the interaction of vulnerable subjects with AI.[34].

2. The Key Element: Human Oversight

Despite the significant advances in AI and machine learning technologies, AI systems still require human oversight as an essential factor[35]. Hence, AI systems may potentially result in discrimination, exacerbate existing discrimination, and impede fundamental human rights and freedoms, or produce outcomes that are contrary to the law[36]. AI systems may exhibit an inability to comprehend certain aspects of human nature, such as empathy, subjective experiences, feelings, or conscience. They may also fail to make decisions based on moral principles or make incorrect decisions as a result of technical errors such as training errors, incorrect data sets, or application errors, among other factors. Moreover, the outcomes of AI systems may exhibit inconsistencies, inaccuracies, manipulative behavior, unsatisfactory results, or ambiguity[37].

AI systems, however, possess considerable potential, and their utilization can yield significant advantages in numerous domains[38]. It is not advisable to disregard the

[33] B Wrede, F Binkofski, F Hegel, G Sagerer, S Krach, 'Can machines think? Interaction and perspective taking with robots investigated via fMRI' (2008) 3(7) PLoS One <doi: 10.1371/journal.pone.0002597> accessed 10 December 2023, e2597.

[34] Gatt L, 'Legal anthropocentrism between nature and technology: the new vulnerability of human beings' (2022) 1 EJPLT, 15-26.

[35] B Green, 'The flaws of policies requiring human oversight of government algorithms' (2022) 45 Computer Law & Security Review 105681, 3-4; European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (Communication) COM (2019) 168 final, 2.

[36] Council of Europe, 'Algorithms and Human Rights, Study on the human rights dimensions of automated data processing techniques and possible regulatory implications' (2018) DGI (2017)12, 10-33; European Commission, 'White Paper on Artificial Intelligence – A European approach to excellence and trust' COM (2020) 65 final, 11; J Constantino, 'Exploring the Article 14 of the EU AI Proposal: Human in the Loop Challenges When Overseeing High-Risk AI Systems in Public Service Organisations' (2022) 14(3) Amsterdam Law Forum 8-10.

[37] Ad Hoc Committee on Artificial Intelligence (CAHAI), 'Towards Regulation of AI Systems' (2020) DGI (2020) 16, 23-32; A Thurai, J Mckendrick, 'AI Isn't Ready to Make Unsupervised Decisions' (Harvard Business Review, 15 September 2022) <https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions> accessed 8 December 2023; Commissioner for Human Rights, Council of Europe, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights' (Recommendation) (2019) <https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64> accessed 8 December 2023, 5; Committee of Ministers, 'Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes' Decl(13/02/2019)1; R Koulu, 'Proceduralizing control and discretion: Human oversight in artificial intelligence policy' (2020) 27(6) Maastricht Journal of European and Comparative Law 724.

[38] High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 8 December 2023, 4-9; J Taylor, M Ross, 'Managing AI Decision-Making Tools' (Harvard Business Review, 10 November 2021) <https://hbr.org/2021/11/managing-ai-decision-making-tools> accessed 8 December 2023.

benefits of AI-based decision-making entirely or to completely disregard the decisions of AI systems (a phenomenon commonly referred to as *"algorithm aversion"*)[39]. Human oversight is the key to utilizing these systems with minimal risk and maximum benefit. In this manner, the most effective outcomes can be achieved through collaboration between humans and AI.

In numerous literary works, human oversight in the life cycle of AI is characterized by various terminologies. In this paper, the definition of human oversight combines surveillance, monitoring, observation, governance, and intervention and provides a more comprehensive understanding. In the section of the paper that examines the existing regulations, the original terms used in these regulations are provided.

2.1. What is "human oversight"?

AI systems possess the ability to identify optimal solutions or the optimal decision on paper; however, in the real world, the optimal outcomes of AI systems may not prove to be particularly useful in every scenario. The most significant aspect is that, in certain instances, optimal solutions may not be appropriate for the human realm, thereby eliminating human value, autonomy, and significance[40]. We are still unable to arrive at a conclusive conclusion that AI systems possess the capability to comprehend concepts of ethics or moral considerations or precisely analyze languages. AI systems possess data-driven insights and render fact-based decisions; however, certain decisions, which are made by humans, are subjective. These individualized decisions are typically based on our experiences, common sense, creativity, intuitive reasoning, or emotions, such as empathy, sympathy, anger, or excitement. Moreover, AI systems lack the capability to comprehend real-world scenarios, unpredictable situations, or the complete context[41]. Besides, the fact-based decisions made by AI systems derive from their inputs. When their inputs are not sufficiently impartial, their outputs will also exhibit bias. In various scenarios, AI systems are incapable of elucidating the reasoning. Lack of reasoning has the potential to infringe upon fundamental rights and freedoms[42]. This is where the oversight of humans comes into play. A process influenced by human oversight is enhanced by principles of ethics, conscience, experience, logic, reasoning, transparency, accountability, holism, and authenticity. Human oversight is of utmost importance for the well-being of humanity and the safeguarding of society's dynamics[43].

Human oversight is crucial in ensuring that an AI system does not compromise human autonomy, infringe upon fundamental human rights, or result in any other adverse outcomes. AI systems ought to uphold human autonomy, refrain from interfering with fundamental rights and freedoms, refrain from manipulating, misguiding, herding, or condescending individuals, provide meaningful opportunities for human choices, prevent harm, and exhibit fairness and explainability. AI systems are intended to raise individuals' standards and enhance their skills, rather than replace

[39] A Chouldechova, M De-Artiga, R Fogliato, 'A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores' (2020) Proceedings of the 2020 CHI Conference on Human Factors Paper 509 <https://dx.doi.org/10.2139/ssrn.4050125> accessed 8 December 2023, 2-9.
[40] Council of Europe (n 31) 30-32.
[41] Koulu (n 32) 734; Thurai, Mckendrick (n 32).
[42] C Reed, 'How should we regulate artificial intelligence?' (2018) 376(2128) Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 3-4.
[43] European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (n 30) 3-6; Green (n 30) 2; Thurai, Mckendrick (n 32).

them[44]. This leads us to believe that it is inherent in this industry to involve humans in the evolution of AI. The mentioned involvement can only be accomplished by incorporating human oversight into the AI's life cycle[45].

Human oversight refers to the process of intervening in the design cycle of AI systems, monitoring the decision processes of the system, and intervening in the decision cycle. The evaluation of the legal, economic, ethical, and social ramifications of the implementation of AI systems, the selection of appropriate stages of the process where human discretion can be effectively employed, and the ability to challenge and override system decisions are also within the purview of human oversight[46]. In AI systems, human oversight is imperative in establishing resilience to cyberattacks and security, evaluating the quality and nature of data sets, minimizing, or compensating for risks or damage, and ensuring accuracy and reliability that can be anticipated within the scope of technical robustness and safety. With human oversight, it is feasible to regulate the consistency and integrity of the decision mechanisms of AI systems. In order to guarantee that AI systems are developed in accordance with universal design principles and are equally accessible to all segments of society, particularly vulnerable subjects, it is imperative that human oversight be made an integral component of the process[47].

It is evident that AI systems are undergoing rapid development. Human oversight is a highly effective approach to enhancing the precision of AI-based decisions, preventing risks by detecting them early, providing guidance through feedback, offering suggestions for system enhancement, and assisting AI systems in making decisions based on real-world experience. For instance, complex data sets on which AI systems are trained can be labeled and tested by human agents and provided with feedback. This allows AI systems to achieve more meaningful, efficient, and accurate results. In other words, human oversight is a highly effective means of augmenting the efficiency of AI systems. This perspective, as well as a form of human oversight wherein humans are incorporated into the life cycle of AI, is commonly referred to as *"Human-in-the-Loop"*[48]. By incorporating human beings into the decision-making process, this perspective ensures that the decision is not solely influenced by AI-based systems but also involves human oversight[49].

The degree and type of human oversight vary in AI-based decision-making processes[50]. In certain decision-making processes, solely human oversight may suffice; however, in other instances, it may be imperative for humans to assume a more proactive and efficient role and act as the sole decision-maker. Every decision-making process based on AI possesses distinct characteristics. The European Commission has provided various types and levels of human oversight in its AI White Paper, although it is not exhaustive: the validity of AI systems' outputs cannot be established before being

---

[44] European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (n 30) 2-4.
[45] High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 11-13, 16.
[46] Green (n 30) 3.
[47] High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 16-19.
[48] Ge Wang, 'Humans in the Loop: The Design of Interactive AI Systems' (Stanford University Human-Centered Artificial Intelligence, 20 October 2019) <https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems> accessed 8 December 2023.
For more information about "human in the loop", see D. Alonso-Ríos, E. Hernández-Pereira, E. Mosqueira-Rey et al., 'Human-in-the-loop machine learning: a state of the art' (2023) 56(4) Artificial Intelligence Review 3005–3054.
[49] B Wagner, 'Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems' (2019) 11(1) Policy & Internet 108-109.
[50] For a categorized approach to the different levels of human oversight, see High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 16; Ross, Taylor (n 33).

reviewed and verified by a human *(Human-in-the-Loop)*; AI systems' outputs are immediately effective but require human intervention to ensure their validity *(Human-out-of-the-Loop)*; a human monitors the AI system in operation, and this human has the authority to intervene and deactivate the system in real time *(Human-on-the-Loop)*; a human is involved in the design process and establishes certain operational lines within the design *(Human-Back-in-Control)*[51].

When contemplating the degree of human oversight, it is imperative to consider the risks, the impact of the decision on fundamental rights and freedoms, the target group, the complexity of the decision, the intended application of the system, and the involvement of the AI system in the aforementioned decision. This assessment ought to be performed on a case-by-case basis. Otherwise, inadequate human oversight may lead to irreversible harm to vulnerable subjects, while excessive levels of human oversight may hinder the potential utilization or complete utilization of AI systems potential. Every decision-making process based on AI should be monitored; however, certain aspects may require active human oversight[52]. Although human oversight may be limited to assisting the decision-making mechanisms of low-risk AI systems, it will be significantly more extensive and efficacious for high-risk AI systems.

When it comes to automated decision-making and profiling, the absence of human oversight may prove disastrous, as the creation of profiles and automated decisions have a significant impact on individuals' fundamental rights and freedoms. Individuals may not be aware that they are profiled or subject to an automated decision because AI systems have the potential to be opaque. When the system lacks sufficient transparency, we cannot expect meaningful human oversight. These systems can lead to discrimination, make inaccurate predictions, or undermine individuals' freedom of choice[53]. The absence of human oversight results in damages that are significantly greater than the norm. Perhaps these damages can be mitigated or averted through human oversight; nevertheless, the occurrence of irreversible harm with AI systems that are devoid of human oversight is inevitable in numerous scenarios[54].

Human oversight must prevail throughout the life cycle of AI systems. Not only when adverse outcomes arise, but also when measures are taken to mitigate such effects, minimize the risk of AI system utilization, and advance the systems. All stakeholders, including manufacturers, developers, distributors, users, and others who possess the potential to interact with the system throughout the life cycle of an AI system, constitute distinct dimensions of human oversight. We hold the belief that enhancing the understanding of the fundamental principles, risks, and advantages of AI systems among all stakeholders will result in meaningful human oversight. If the stakeholders are cognizant, trained, experienced, qualified, and have a high level of awareness, their influence on the process will increase[55]. Consequently, the presence of human oversight may not always be a definitive solution to issues[56]. The overreliance on algorithmic decisions, commonly referred to as *"automation bias,"* can diminish or, in certain

[51] European Commission, 'White Paper on Artificial Intelligence – A European approach to excellence and trust' (n 31) 21; G Lazcoz, P De Hert, 'Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities' (2023) 50 Computer & Law Security Review 105833, 7-8.
[52] High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 12-13; Ross, Taylor (n 33).
[53] Article 29 Data Protection Working Party, 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679' (2018) 17/EN, WP251rev.01, 5-6.
[54] Thurai, Mckendrick (n 32).
[55] Constantino (n 31) 4-5, 14-15.
[56] Green (n 30) 7-8.

instances, eliminate the impact of human oversight. This is since individuals tend not to question the decisions made by automated decision systems, despite having the opportunity to intervene[57]. They hold the belief that these decisions are innately correct and consider them to be *"credence good"*. This circumstance has a detrimental impact on the capacity to make judicious and deliberate decisions[58].

Furthermore, the insufficient time allocated for human oversight, the lack of skilled workers, or the inadequate provision of authority, autonomy, and competence may render human oversight meaningless or insufficient. The necessity to make decisions within a limited time frame gives rise to individuals' overreliance on AI-based decision-making. Meaningful human oversight is not possible at a level where human oversight consists solely of approving or rejecting the system's decisions. In numerous instances, human oversight is incorporated into the system as an ineffective but symbolic component solely to display the existence of human oversight, to avoid responsibilities, or to assert that legal requirements have been met[59].

The assumption of meaningful human oversight may legitimize the unlawful use of AI systems, provide a false sense of security, harm accountability, or infringe upon individuals' fundamental rights and freedoms[60]. In other words, meaningless human oversight can result in greater harm than good due to the false perception of trust it creates. The requirements stipulated in legal regulations for meaningful human oversight will be discussed in the following sections of the article

2.2. Reflections of Human Oversight in Legal Regulations

2.2.1. Reflections in the GDPR

Data is the foundation of the information society. The magnitude and variety of data held by these organizations have the potential to impact individuals' free will and the decisions they make. Another risk is that individuals may be subjected to the dictatorship of algorithmic decisions as a result of decisions made based on algorithms.[61] In order to safeguard individuals whose personal data are processed and thus considered data subjects from the potential hazards of AI systems manipulating them, jeopardizing their individual autonomy, imposing conditioning, or deceiving them, the regulation contemplated under the GDPR stipulates the right to refrain from being subjected to a decision solely based on automated processing, including profiling, when it produces legal effects on individuals or has a significant impact on them. This right prevents the use of solely automated decision-making, provides the right to request a human inspection, considers altering the decision, and is an important instrument to challenge the negative impacts of AI decision-making[62].

---

[57] Chouldechova, De-Artiga, Fogliato (n 34) 2-10; Constantino (n 31) 16-17.

[58] For interesting studies measuring the automation bias that individuals develop against algorithms, see. Chouldechova, De-Artiga, Fogliato (n 34); J Biermann, J Walter, JJ Horton, 'Algorithmic Advice as a Credence Good' (2022) ZEW – Centre for European Economic Research Discussion Paper No. 22-071 <https://dx.doi.org/10.2139/ssrn.4326911> accessed 8 December 2023.

[59] Constantino (n 31) 11-17; Koulu (n 32) 734-735; Wagner (n 44) 110-114.

[60] Chouldechova, De-Artiga, Fogliato (n 34) 9-10; Green (n 30) 9-11.

[61] Council of Europe (n 31) 12-14; European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (n 30) 5; European Parliament, 'The impact of General Data Protection Regulation (GDPR) on artificial intelligence' (2020) PE 641.530, 75.

[62] De Hert, Lazcoz (n 46) 3; Green (n 30) 5; E Küzeci, Kişisel Verilerin Korunması (4th edn, On İki Levha Yayıncılık 2020) 263-264; F Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making'

Article 22 of the GDPR is solely applicable to automated procedures that do not necessitate human intervention.[63] At this juncture, it is imperative to assess the minimal level of human intervention required to avert a decision-making process that is solely automated. It is imperative that human intervention be meaningful and executed by an individual who possesses the authority and expertise to alter the decision. If there is a presence of symbolic human intervention instead of meaningful human intervention in a system, it will become impossible to assert that the system is solely automated in its decision-making process[64].

According to the GDPR, automated decision-making, including profiling, should have serious consequences for individuals, such as legal or similar significant effects. In other words, Article 22 will only cover serious impactful effects. It is deemed that decisions that result in discrimination or disadvantage in accessing education or health services have serious consequences[65].

Exceptions to this right are accounted for in the GDPR[66]. Certain conditions apply to the processing of special category data within the scope of Article 22 of the GDPR[67]. It is imperative to consider the special category data that can be obtained through profiling. This is because profiling enables the acquisition of special category data by analyzing a group of personal data and combining it with other data. By implementing this approach, the acquisition of numerous special category data, such as health conditions, political or religious beliefs, race, and ethnicity, will enhance the vulnerability of individuals who are already vulnerable[68]. When exceptions are in effect, it is imperative to implement measures to safeguard the rights, freedoms, and legitimate interests of the data subject[69].

The controller should take appropriate measures to safeguard the data subjects' rights, freedoms, and legitimate interests in automated decision-making, including profiling. The primary component of this provision is human intervention. The data controller ought to provide at least the right to obtain human intervention to express their point of view, obtain an explanation, and contest the decision[70]. This intervention is expected to have significance, which implies that it should be executed by an individual who possesses the appropriate authority and capability to evaluate and, if necessary, modify the decision. The individual who is authorized must possess the necessary qualifications, expertise, and training in automated decision-making systems and their methodologies; possess a comprehensive understanding of both the inputs and

---

(Council of Europe Study, Directorate General of Democracy, Strasbourg, 2018) 40-44; Gabriela Zanfir-Fortuna, Sebastião Barros Vale, 'Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities' (2022) Future of Privacy Forum <https://fpf.org/wp-content/uploads/2022/05/FPF-ADM-Report-R2-singles.pdf> accessed 8 December 2023, 7; GDPR, art. 22; High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 16; HC Aksoy, 'Kişisel Verilerin Korunması Yönüyle Algoritmik Karar Verme' (2022) 4(2) Kişisel Verileri Koruma Dergisi 77-78.

[63] WP 29, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 17/EN WP251rev.01, adopted on 3 October 2017, as last revised and adopted on 6 February 2018. Available online at https://ec.europa.eu/newsroom/article29/items/612053

[64] Article 29 Data Protection Working Party (n 48) 6-8, 20-21; Barros Vale, Zanfir-Fortuna (n 57) 28-34; De Hert, Lazcoz (n 46) 11-12; GDPR Recital 71; Wagner (n 44) 117.

[65] Article 29 Data Protection Working Party (n 48) 21-22; Barros Vale, Zanfir-Fortuna (n 57) 35-38; Council of Europe (n 31) 15.

[66] Barros Vale, Zanfir-Fortuna (n 57) 9-12.

[67] Küzeci (n 57) 264.

[68] Council of Europe (n 31) 15-16; European Parliament (n 56) 61-62.

[69] Article 29 Data Protection Working Party (n 48) 12-15, 19-20, 23-24; De Hert, Lazcoz (n 46) 4; GDPR, art. 22/2-4.

[70] GDPR Recital 71.

outputs of the systems; possess the ability to access all relevant information; modify or reject the decisions; and possess sufficient time to conduct an evaluation[71]. The objective of human intervention is to establish data subjects' control over their data and the decisions made regarding them[72].

Individuals require an explanation in automated decision-making. The rationale may pertain to *system functionality*, including but not limited to the logic and model of the system, references, anticipated consequences, specifications, and basic structure criteria, or to *specific decisions*, including but not limited to motives, justification, allocation of attributes, decision guidelines, reference groups, or individual circumstances. The explanation of *system functionality* is an *ex ante* explanation. Nonetheless, *ex post* explanations may encompass both *system functionality* and *specific decisions*. The notification obligation stipulated in Articles 13 and 14 of the GDPR and the right of access provision in Article 15 of the GDPR may provide an *ex ante* explanation. The notification obligation and the right of access, as outlined in the pertinent articles, entail articulating the decision mechanism, logic, purpose of processing, scope, intention, and predictable outcomes of the system prior to the automatic decision-making process. Nonetheless, there is no explicit provision that encompasses the *ex post* right to explanation. Despite the explicit mention of this statement in Recital 71 of the GDPR, it is imperative to note that it is not legally binding but rather aids in the interpretation of the GDPR provisions[73].

According to Article 22 of the GDPR, the potential of human intervention is an *ex post* tool. The possibility of requesting a human intervention is available to individuals after the automated decision has been made[74], as this provision only applies when there are serious consequences for them following this decision-making procedure[75]. At this point, it remains uncertain whether human intervention provides a sufficient explanation for automated decision-making, whether vulnerable subjects will be able to accurately assess the consequences of the decision, and whether they will be able to effectively exercise this right to do so.[76] In our perspective, the sole method of providing an explicit *ex post* explanation is through meaningful human intervention.[77]

In general, profiling and automated decision-making procedures are not transparent to individuals. It is in contradiction to the requirements of transparency and accountability stipulated in the GDPR. Individuals are unaware of the means of processing or the results. A lack of transparency and accountability makes it impossible to anticipate fairness, purpose or storage limitation, data minimization, or accuracy. The absence of these components in the profiling or automated decision-making process can increase individuals' vulnerability[78]. Another aspect of transparency pertains to individuals being made aware of the automated decision-making or profiling activities

[71] Wagner (n 44) 114-116.
[72] Aksoy (n 57) 79-82; Article 29 Data Protection Working Party (n 48) 27-29.
[73] B Mittelstadt, L Floridi, S Watcher, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) International Data Privacy Law 78-90.
[74] WP 29, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, 17/EN WP251rev.01, adopted on 3 October 2017, as last revised and adopted on 6 February 2018. Available online at https://ec.europa.eu/newsroom/article29/items/612053
[75] A Lukács, S Váradi, 'GDPR-compliant AI-based automated decision-making in the world of work' (2023) 50 Computer Law & Security Review 105848, 13.
[76] AI e GDPR: l'Automated Decision Making, la protezione dei dati e il diritto alla 'intellegibilità' dell'algoritmo Emiliano Troisi https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1027
[77] Hard law and soft law on data protection: what a DPO should know to better perform his or her tasks Maria Cristina Gaeta https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1069
[78] Article 29 Data Protection Working Party (n 48). 9-12; Council of Europe (n 31) 15-16.

taking place. Data subjects must be informed about the intended use and the current and anticipated consequences of the processing. Nevertheless, it can be challenging to provide meaningful and sufficient information to data subjects in the context of automated decision-making and profiling due to a lack of transparency. As a consequence, it may become impracticable for the data subjects to exercise the rights granted to them under the GDPR[79].

The GDPR provides for human intervention as an organizational measure to control risks in data processing. The implementation of Article 22 of the GDPR and its human intervention mechanisms may serve as a testament to the principle of accountability. The requirement for data protection impact assessment (DPIA) as stipulated in Article 35 of the GDPR is a fundamental component of this framework, as it enables the controller to demonstrate its compliance with human intervention mechanisms. The principle of accountability requires that logical explanations be documented and recorded. This encompasses elucidating whether the controller's decision-making procedure necessitates human intervention, scrutinizing the interaction between automated and human elements in this process, examining the impact of this interaction on the overall outcomes, determining and recording the level of human intervention in the decision-making process[80], and at what stage this occurs. The existence and continuity of meaningful human intervention can be evaluated within the scope of the DPIA. The evaluation of meaningful human intervention is regarded as one of the procedures that can serve not only to prevent error, bias, and discrimination but also to legitimize a system and even to safeguard the dignity of an individual, safeguarding his fundamental rights and freedoms within this system[81].

2.2.2. Reflections in the AI Act

In addition to the GDPR, another document that is pertinent to our discussion is the European Commission's proposal for a legal regulatory framework on AI (AI Act)[82]. The new legal framework aims to provide a unified set of regulations that align with the existing human rights instruments of the European Union (EU). The provisions of the GDPR regarding human oversight concern the implementation phase of AI, whereas the AI Act employs a more comprehensive approach that encompasses both the development and implementation phases. Simultaneously, the AI Act is compatible with the GDPR. It contains related provisions, and these related provisions strengthen the connection between these two regulations[83].

The AI Act adopts a risk-based approach. The idea of establishing regulations that are proportionate and measured to the extent of the risk posed by an AI system has been adopted. The AI Act has incorporated four distinct risk categories *(unacceptable, high,*

[79] Article 29 Data Protection Working Party (n 48) 12-13, 24-27; C Kuner, C Millard, DJB Svantesson, FH Cate, O Lynskey, 'Machine learning with personal data: is data protection law smart enough to meet the challenge?' (2017) 7(1) International Data Privacy Law 2.
[80] Gatt L, Caggiano I, 'Consumers and Digital Environments as a Structural Vulnerability Relationship ' (2022) 2 EJPLT, 8-16.
[81] De Hert, Lazcoz (n 46) 15-17.
[82] European Commission, 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, COM (2021) 206 final, 2021/0106 (COD).
[83] De Hert, Lazcoz (n 46) 6-7.

limited, and minimal)[84]; however, in the context of our study, high-risk systems that involve human intervention will be included.

AI systems that possess a significant likelihood of infringing fundamental rights and freedoms and jeopardizing the security of individuals are categorized as high-risk[85]. This is since these systems make decisions that may result in discrimination, violations of fundamental rights and freedoms, harm to personality, and serious victimization of individuals[86]. When evaluating the potential for infringement, factors such as the intended use of the AI system, the audience it affects, the temporal and geographical prevalence of use, the possibility of reversing the decision made by the system, the possibility of compensating for this decision, and the effectiveness of compensation mechanisms for the resulting damages are considered. In accordance with the text adopted by the European Parliament on June 14, 2023, which is "the adopted version of the AI Act"[87], it was recommended that the evaluation of systems that pose a similar or greater risk to the high-risk AI systems identified in Annex III also take into account the extent of human oversight and the likelihood of a human intervention to override a decision or recommendation that could potentially cause harm[88]. According to the AI Act, several precautions must be applied to high-risk systems. These include the quality of data sets, technical documentation and record keeping, transparency and the provision of information to users, human oversight, robustness, accuracy, and cybersecurity[89].

Human oversight is described under Article 14 of the AI Act. Pursuant to the relevant Article, the design and development of high-risk AI systems[90] shall enable them to *be effectively overseen by natural persons for the duration of their use*. Human oversight is intended to prevent or mitigate risks to health, safety, and fundamental rights that may arise from the intended use or foreseeable misuse of high-risk AI systems within their reasonable use, particularly if the other requirements outlined in the AI Act are fulfilled. Natural persons should be able to evaluate the outputs of systems, detect anomalies, take action against the outputs, and override or disregard the systems decisions[91]. In accordance with the the adopted version of the AI Act, it was recommended that human overseers, who are accountable for human oversight, possess a sufficient level of AI literacy and the necessary support and authority to enable a thorough investigation throughout the lifespan of the AI system and subsequent to any incidents[92].

In order to ensure human oversight, the AI Act provides two different measures. These measures may be implemented either independently or in a combined manner. Initially, in cases where it is technically feasible, it is imperative that human oversight be identified and built by the provider before the introduction of high-risk AI systems to

[84] F Zuiderveen Borgesius, M Veale, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22(4) Computer Law Review International 98.
[85] For a list of AI systems considered high risk, see. AI Act, Annex III.
[86] AI Act, Recital 27-63.
[87] European Parliament, Artificial Intelligence Act - Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM (2021)0206 – C9-0146/2021 – 2021/0106(COD)), Texts Adopted, P9_TA(2023)0236.
[88] The Adopted Version of the AI Act, Amendment 254.
[89] AI Act, art. 6-15.
[90] AI Act, art. 6.
[91] AI Act, art. 14/1-2.
[92] The Adopted Version of the AI Act, Amendment 314.

the market or into service. Second, the provider should identify human oversight before high-risk AI systems are placed on the market or put into service and then implemented by the user. For meaningful human oversight to be effective, human overseers must meet certain requirements. It is imperative that human overseers possess the ability to effectively monitor high-risk AI systems by acquiring a full understanding of their capabilities and limitations, identifying anomalies, and accurately interpreting their outputs[93]. With the adopted version of the AI Act, the requirement that human overseers *"fully understand the capabilities and limitations"* of the high-risk AI systems has also been changed to *"be aware of and sufficiently understand the relevant capacities and limitations"*[94]. This modification demonstrates a reasonable expectation or acceptance that it may not always be feasible or necessary for users to fully comprehend an AI system[95].

The adopted version of the AI Act recommended that human agency and oversight be taken into account in AI systems at all risk levels in order to promote a coherent human-centric approach and develop ethical and trustworthy AI[96]. Human oversight is a reflection of the EU's human-centric AI approach. The human-centric AI approach represents an understanding that prioritizes humans and human values throughout the life cycle of AI systems. Human agency and oversight are defined as *"AI systems shall be developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans"*[97]. Human oversight is a human-centric AI approach tool that is intended to safeguard individuals from the adverse effects of AI systems[98].

Human oversight and other precautions foreseen for high-risk AI systems mentioned above have a complementary effect. It is necessary to increase the level of human oversight in cases where other precautions are or may remain inadequate or inefficient to mitigate the risk. Discretionary opportunity in decision-making and intervention, fairness, ethic values of overseers, transparency, accountability of oversight, and other obligations such as technical documentation, record keeping, and robustness are essential for meaningful human oversight. The absence of any one of these components may result in the violation by human overseers of the fundamental rights and freedoms of all individuals, particularly those who are vulnerable[99].

Especially in the case of high-risk AI systems used to provide information or advice for human decisions, it is imperative that human overseers are aware of the potential tendency to automatically rely or over-rely on the output produced. If deemed necessary, human overseers ought to possess the capability to defer the utilization of the high-risk AI system, disregard its output, override it, invalidate it, produce an adverse outcome, or intervene in its operation[100]. This requires that human overseers be provided with sufficient transparency within the system. In order for human overseers to understand, interpret, and take appropriate action, sufficient information must be

[93] AI Act, art. 14/4-a; AI Act, Recital 48.
[94] The Adopted Version of the AI Act, Amendment 318.
[95] L Enqvist, ''Human oversight' in the EU artificial intelligence act: what, when and by whom?' (2023) 15(2) Law, Innovation and Technology 519.
[96] The Adopted Version of the AI Act, Amendment 213.
[97] L Enqvist (n 85) 510-512.
[98] Constantino (n 31) 10.
[99] Constantino (n 31) 6-8, 11-12; Enqvist (n 85) 517-518.
[100] AI Act, art. 14/3-4.

provided. Article 12 of the AI Act requires the automatic recording of events (logs) while high-risk AI systems are operating. Furthermore, Article 13 stipulates that the system operations must be designed and developed in a sufficiently transparent manner and that the system outputs must be comprehended and interpreted by users. In accordance with the AI Act, transparency requirements include disclosure of human oversight measures, including technical measures that make it easier for users to make sense of AI outputs[101]. The AI Act's article regulating the obligations of providers of high-risk AI systems has been updated with the adapted version of the AI Act to ensure that those who provide human oversight are specifically informed of the risk of automation or confirmation bias[102]. These obligations help ensure that sufficient information is provided to the human overseer. It is imperative for providers to ensure that their high-risk AI systems are equipped to accommodate human interference[103].

Obligations for human oversight are generally delegated to the provider. Users are obligated to utilize the systems solely in accordance with the established instructions of the provider. The user has discretion to organize its resources and activities while applying human oversight measures determined by the provider. The user is also obliged to monitor the relevant AI system throughout the period of use and to inform the provider or distributor in the event of any risks, malfunctions, or incidents[104]. This approach is the outcome of a proactive perspective[105]. In the adopted version of the AI Act, the European Parliament calls on users (called deployers in the adopted version) to ensure that "*the natural persons assigned to ensure human oversight of the high-risk AI systems are competent, properly qualified, and trained, and have the necessary resources in order to ensure the effective supervision of the AI system in accordance with Article 14"*[106]. It is anticipated that this amendment will enhance the compatibility and connection between the user and the provider, thereby elevating the standard of human oversight and enhancing its significance[107].

As per the AI Act, the design choices made by providers have a significant impact on the time when human oversight is implemented[108]. Nonetheless, the formulation of Article 14 largely fails to provide a precise indication of when human oversight ought to be incorporated into a system. It is uncertain when human oversight is necessary, the manner in which it should be implemented, the scope and limitations of the responsibilities of human overseers, and the standards of human oversight. Furthermore, there is a lack of guidance that provides clarification on the matter. This is likely to diminish the impact of the AI Act on the implementation and effectiveness of human oversight. The ambiguity of the provisions and the absence of adequate guidance may result in ineffectiveness. This situation has the potential to adversely impact an individual's fundamental rights and freedoms[109].

According to the adopted version of the AI Act, a fundamental rights impact assessment is envisaged for high-risk AI systems. Prior to deploying an AI system with a high risk, it is imperative that the deployers conduct an evaluation of its impact within

[101] AI Act, art. 13/3-d.
[102] The Adopted Version of the AI Act, Amendment 334.
[103] Enqvist (n 85) 521-524.
[104] AI Act, art. 29.
[105] Enqvist (n 85) 518.
[106] The Adopted Version of the AI Act, Amendment 401.
[107] Enqvist (n 85) 529-530.
[108] Enqvist (n 85) 525-528.
[109] Constantino (n 31) 3-4, 12-14.

the specific context of its utilization, except for AI systems intended for use in Area 2 of Annex III. The evaluation also includes measures taken in relation to human oversight[110]. Furthermore, the obligation to provide and keep up-to-date information regarding human oversight and governance is envisaged for high-risk AI systems, for which registration is required in accordance with Article 51 of the AI Act[111].

Although the protections contemplated by the AI Act and the provisions it envisions remain innovative and commendable, it is noteworthy that AI systems that do not fall under the high-risk category but may pose a significant threat to vulnerable subjects are not included within the purview of the human oversight provisions of the AI Act. We believe it will weaken the protection it aims for. Indeed, the failure to implement concrete measures regarding human oversight and the failure to clearly define boundaries may diminish the efficacy of the protection envisioned by this provision. It is inevitable that vulnerable subjects will be the most adversely affected segment of society in this circumstance

3. The Phenomena That Best Define the Right: Transparency and Accountability.

The correlation between accountability, transparency, and human oversight is bidirectional. While human oversight facilitates or makes it possible to achieve transparency and accountability, transparency and accountability also allow for human oversight and make it more effective[112].

Transparency entails that the life cycle of an AI should be comprehensible and readily accessible, that the language employed should be plain and unambiguous, that care should be taken in the language employed in the case of vulnerable subjects and made comprehensible to them, that individuals should be notified through written or other effective means, and that any information to be provided and the process of fulfilling the transparency requirements should be provided without any cost. It is imperative that the capabilities and limitations of AI systems are adequately explained to individuals. The information provided must be clear, intelligible, easily accessible, accurate, and updated. Individuals must be able to comprehend that they are communicating with an AI system and its results. Individuals should be given the option of requesting not to interact with AI systems and instead opting for a human interaction[113].

Systems where the reasons for their decisions cannot be traced and it is not possible to explain how or in what way they make decisions are called *"black box"* systems[114]. As algorithms become more complex, it becomes increasingly difficult to understand and explain how the AI system works[115]. Although the decision-making mechanism of the AI system can be explained, it remains a black box for users because it is still not understandable to them[116]. The issue of opacity in these systems is in contradiction to the principle of transparency.

---

[110] The Adopted Version of the AI Act, Amendment 413.
[111] The Adopted Version of the AI Act, Amendment 770.
[112] Z Öğretmen Kotil, Kişisel Verilerin Korunması Çerçevesinde Yapay Zekâ (On İki Levha Yayıncılık 2022) 152-153.
[113] High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 18.
[114] High-Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines' (n 8) 5.
[115] N Bostrom, Süper Zekâ: Yapay Zekâ Uygulamaları, Tehlikeler ve Stratejiler (Ferit Burak Aydar tr, 2nd edn, Koç Üniversitesi Yayınları 2020) 167.
[116] Reed (n 37) 7.

Transparency is especially important when it comes to vulnerable subjects. The violations of rights that these subjects may encounter cannot be addressed without transparency[117]. The utilization of AI systems, which are not transparent, has the potential to cause irreversible harm to individuals without recourse. Especially when it comes to vulnerable subjects, a system that fails to meet transparency standards should not be used[118]. Hence, the aforementioned notion of transparency does not imply the comprehensive exposition of a system. When determining the requirements and standards for effective transparency, it is essential to consider the individual who is in contact with the AI system. The requirements for transparency are contingent upon age, educational level, potential harm, vulnerability, and other factors. Every AI system has a different risk level, and this level should be considered when determining transparency requirements[119]. Considering the risk level of the system, it is important to assess the appropriate stage at which transparency (*ex ante* or *ex post*) should be sought[120].

Transparency is essential in order to ensure effective and efficient remedy mechanisms for the aggrieved individual in the event of a violation of rights[121]. Human oversight is a crucial element in safeguarding individuals against the adverse effects of AI systems, and transparency is a crucial component in ensuring its proper implementation. The absence of transparency results in a distortion of the power balance between AI systems and individuals, ultimately leading to the disadvantage of individuals[122]. Human oversight is also an important mechanism for AI systems to meet transparency requirements. Individuals can request explanations from human overseers regarding decisions; the functioning of the system can be interpreted by human overseers and explained when necessary.

Transparency is also imperative in ascertaining and verifying the technical reliability, safety, and robustness of the system. False positives and false negatives have the potential to cause significant discrimination in the lives of individuals. Moreover, in the event that the system possesses a direct discriminatory component, it can be identified through the principle of transparency. It is imperative to identify systems that are susceptible to discriminatory outcomes to prevent individuals from being victimized by algorithmic decisions. It is extremely difficult to detect and eliminate discriminatory decisions made by opaque systems[123].

The principle of accountability serves as a significant instrument for determining the content and degree of transparency, and transparency serves as a means to reinforce this principle. Accountability involves the ability to identify those responsible and to explain, document, and record the decision-making processes of the system. The development of AI systems ought to be founded on the principle of accountability, which facilitates the identification of risks and the implementation of appropriate

[117] Ad Hoc Committee on Artificial Intelligence (CAHAI) (n 32) 30, 32, 79-80, 89.
[118] Commissioner for Human Rights, Council of Europe (n 32) 10.
[119] Council of Europe (n 31) 38; High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 13; Reed (n 37) 7, 9-10.
[120] A Murray, J Black, 'Regulating AI and Machine Learning: Setting the Regulatory Agenda' (2019) 10(2) European Journal of Law and Technology 15-16.
[121] Commissioner for Human Rights, Council of Europe (n 32) 14.
[122] European Union Agency for Fundamental Rights, 'Getting the Future Right – Artificial Intelligence and Fundamentral Rights' (2020) <https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf> accessed 8 December 2023, 66.
[123] European Parliament (n 56) 44-45; Zuiderveen Borgesius (n 57) 38.

measures to mitigate them[124].

As per the accountability principle, it is imperative that AI systems exhibit transparency and explanation. An AI system that can be explained should possess the capability to articulate the primary factors that influence a decision, the consequences of altering a particular factor on the decision, the rationale behind different decisions on the same issue, or the rationale behind the same decision on different issues. Thus, it is possible to assess the system's consistency, integrity, and predictability. The notion of explainability does not necessitate a comprehensive examination of the system's functional aspects. Rather, it is sufficient to comprehend the rationale behind a decision and its underlying circumstances. Explainability is a method of determining responsibility and guaranteeing accountability[125]. It is imperative that a sufficient level of explainability be provided for diverse user groups. It is important that the explanation of the decision system be customizable for different users[126].

Accountability also enables a process that is sustainable, dynamic, and flexible enough to adapt to new developments. It is important that the stakeholders involved in the life cycle of the AI system are able to keep up with the dynamism of the process. The procedure ought to be capable of adapting to alterations in risk factors, and a risk-based approach ought to be implemented. Additionally, internal control and auditing are integral parts of accountability. Aiming for accountability reduces the workload of supervisory authorities, optimizes stakeholder processes, enables auditability, and provides a wider and safer margin of action for all stakeholders[127].

Potential adverse consequences may be identified, evaluated, documented, and mitigated through the implementation of accountability measures[128]. As will be evident, accountability is a highly efficient means of addressing the power disparity between vulnerable subjects and AI systems[129].

The human-centric AI approach and its reflection, human oversight, are effective in ensuring transparency and accountability. When humans are involved in the process, it becomes more feasible to design the system in a manner that is comprehensible to individuals, explicate the decision mechanism, maintain records, identify crucial steps and responsibilities, and incorporate human judgment and human reasoning into the process.[130] Due to the fact that human oversight brings together two distinct actors, namely humans and AI, on a single platform.

4. Would it be possible to regulate the right of vulnerable subjects to request human

[124] Council of Europe (n 31) 39-40; D Leslie, 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector' (The Alan Turing Institute 2019) <https://doi.org/10.5281/zenodo.3240529> accessed 8 December 2023, 23-24; High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 19-20.

[125] F Doshi-Velez, M Kortz, 'Accountability of AI Under the Law: The Role of Explanation' (2017) Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper <https://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584> accessed 8 December 2023, 3-10; Öğretmen Kotil (n 102) 178-180.

[126] AN Yates, CA Hahn, DA Broniatowski, K Greene, MA Przybocki, PC Fontana, PJ Phillips, 'Four Principles of Explainable Artificial Intelligence' (2021) NISTIR 8312 <https://doi.org/10.6028/NIST.IR.8312> accessed 8 December 2023, 2-8.

[127] MB Kaya, 'Kişisel Verilerin Korunmasında Yeni Paradigma: Hesap Verilebilirlik İlkesi' (2020) 78(4) İstanbul Hukuk Mecmuası 1885-1895.

[128] European Commission, 'Building Trust in Human-Centric Artificial Intelligence' (n 30) 6; High-Level Expert Group on Artificial Intelligence, 'Ethic Guidelines for Trustworthy AI' (n 33) 15-16.

[129] Kaya (n 117) 1894.

[130] Borelli, D. and Gatt, L., 2019. Privacy by Design and by Default and Data Protection Impact Assessment (DPIA). Eur. J. Privacy L. & Tech., p.100.

oversight?

4.1. Analysis of the Current Situation

Regulations are supposed to liberate people and close the security gap.[131] Following Habermas' conception of the social welfare state is also important for understanding the current regulations. Although there is a broad pattern of vulnerability, there is no specific or adequate regulation for vulnerable subjects.[132]

The preamble to the GDPR states that everyone, regardless of their nationality and domicile, has the right to have their personal data protected, which is the origin of the GDPR in relation to vulnerable data subjects.[133] When we look at the justification of the GDPR, we observe that children are identified as vulnerable subjects, and there is a clear indication in this regard, and two criteria are set for other vulnerabilities. The first is what kind of measures will be taken when consent cannot be given, and the other is in cases where there is an inequality between the data controller and the data subject even if consent is given. For instance, the GDPR emphasized a striking vulnerability in cases where the counterparty is the state, that the consent obtained from the data subject should not be considered valid in any case, and that a situation superior to consent should be sought. Indeed, criteria such as the degree of employee-employer relations and the position of the person interacting with the technology show that vulnerability is not limited to adjectives.[134] The GDPR, which binds the data controller's data processing due to legitimate interest to the condition that the fundamental rights and freedoms of the data subject are not outweighed, has tried to balance the concept of vulnerability with the criterion of "not outweighing the fundamental rights and freedoms" this time. On the other hand, there are uncertainties due to the lack of regulation on the processing of data on vulnerable subjects, especially in criminal investigation, immigration, money laundering, and data processing for the public interest. For instance, it is noteworthy that there are uncertainties for subjects who are vulnerable regarding the legitimacy of consent granted in cases where the public interest prevails, particularly in the context of processing for genetic or scientific research[135].

Although all these explanations indicate that the GDPR includes many regulations on taking vulnerable subjects seriously and not completely surrendering them to the mercy of someone else, it is observed that it is not sufficient when evaluated together with the AI Act that followed. In particular, it does not make sense with the analysis of the current state of many points such as profiling, discrimination, lack of effective utilization and learning from AI, failure to fully ensure the right to human oversight, and failure to strengthen the human agency side, especially regarding data processed in AI.

---

[131] S Rossi, 'Forme Della Vulnerabilita e Attuazione Del Programma Costituzionale' [2017] RIVISTA AIC.
[132] Chriss (n 14).
[133] GDPR Recital 75.
[134] Federico Cabitza, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger, Andreas Holzinger, Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI,Expert Systems with Applications,Volume 213, Part A,2023.
[135] GDPR Recital 38: "*Children merit specific protection with regard to their personal data, as they may be less aware of the risks, consequences and safeguards concerned and their rights in relation to the processing of personal data. [2]Such specific protection should, in particular, apply to the use of personal data of children for the purposes of marketing or creating personality or user profiles and the collection of personal data with regard to children when using services offered directly to a child. [3]The consent of the holder of parental responsibility should not be necessary in the context of preventive or counselling services offered directly to a child.*"

The AI Act centered on the hazards associated with AI and aimed to safeguard vulnerable subjects by defining risk categories. In this context, it is appropriate to mention a procedure that progresses from a moderate level of vulnerability to a level of vulnerability that is unacceptable.

The justification[136] states that the regulation is based on certain vulnerabilities caused by vulnerability and vulnerabilities that lead to vulnerability, and that the degree of risk of AI-based programs on issues such as discrimination and gender equality is determined. It is imperative to reconsider these statements in light of the increasing vulnerability caused by AI. As we explained in the previous section, if an AI that develops applications that violate human rights cannot gain the trust of vulnerable subjects, it will cease to be a valid product under the AI Act. Trust is also a vulnerable asset, and according to current regulations, this trust has not yet been achieved. The list of prohibited AI practices laid down in Article 5 of the AI Act includes programs that manipulate the minds of vulnerable subjects, especially children and persons with disabilities[137], influence their behavior, and distort their behavior by taking advantage of their psychological and physical vulnerability. As previously mentioned, practices that render genuine individuals, who are subject to the will of natural persons, who are subjected to oppression by public authorities or their employers under the balance of power, even more vulnerable to AI are also categorized under the high-risk category, despite not being included in the prohibition list in the AI Act.[138]

In the most recent revision to the AI Act, the general AI models, commonly referred to as "Foundation Models," which were discovered through OpenAI applications, were placed in the high-risk category. The Foundation Model, which finds definition in many high-risk models, poses a serious risk because it can work with many technologies of the future and be articulated with them. In the report prepared by the Human-Centered AI Foundation Model Research Center of Stanford University, it has been stated that we will never have access to which programs these models will act together and cause irreversible damage, and it will be difficult for us to detect the damage itself.[139]

It is imperative to emphasize that the AI Act does not directly affect vulnerable subjects, but rather, systems assess risks when leaving individuals vulnerable. The AI Act articulates the notions of human oversight and human vulnerability, since AI systems do not eliminate human dignity. Humans can make erroneous decisions, but this is their characteristic, and it is not acceptable for this to be sacrificed to a flawless algorithm. High-risk AI systems, which will be designed in such a way that natural persons can control their functioning, should also include a system that will allow

---

[136] AI Act, Explanatory Memorandum 1.2: "*Consistency is also ensured with the EU Charter of Fundamental Rights and the existing secondary Union legislation on data protection, consumer protection, non-discrimination and gender equality. The proposal is without prejudice and complements the General Data Protection Regulation (Regulation (EU) 2016/679) and the Law Enforcement Directive (Directive (EU) 2016/680) with a set of harmonised rules applicable to the design, development and use of certain high-risk AI systems and restrictions on certain uses of remote biometric identification systems. Furthermore, the proposal complements existing Union law on non-discrimination with specific requirements that aim to minimise the risk of algorithmic discrimination, in particular in relation to the design and the quality of data sets used for the development of AI systems complemented with obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle*."

[137] Minors and new technologies: From parental responsibility to parental control in balancing with the child's right to personality Livia Aulino https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1071

[138] L Edwards, 'The EU AI Act: A Summary of Its Significance and Scope' (2022) Ada Lovelace Institue <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf > accessed 10 December 2023.

[139] R Bommasani et al., 'On the Opportunities and Risks of Foundation Models' (2021) Standford Human-Centered Artificial Intelligence Report <https://crfm.stanford.edu/report.html> accessed 10 December 2023.

human oversight. Article 14 of the AI Act perceives the human oversight phenomenon as an attempt to minimize the risks of high-risk AI systems.

The lack of human oversight would leave a wider audience vulnerable. It is imperative to acknowledge the right to request human oversight for vulnerable subjects in high-risk AI systems or to prohibit them from being used. Nonetheless, such a framework is not currently in place. The primary issue is the damage that vulnerable people will suffer when faced with these systems without taking advantage of a guardianship authority. Subjects who are deemed vulnerable and do not effectively possess the entitlement to request human oversight or control the system will be unable to determine where to halt the system or discontinue communication. Article 14 of the AI Act does not provide any guidance on how to take measures when human oversight is misused by vulnerable subjects or when the oversight is misaligned with the AI.[140] To guarantee that all the consequences of this right against vulnerable subjects are thoroughly examined, not only as a form but also as a basis. If the right cannot be examined, it must be subjected to an alternative system. Additionally, it is imperative to eliminate all the consequences of automated decisions made by the algorithm. An unfair, error-prone or unlawful decision made by the algorithm could subsequently damage the fundamental rights of the vulnerable. To summarize, the entitlement to human oversight is solely granted at a level in the AI Act that legitimizes algorithmic error.[141]

4.2. A Fundamental Rights Advocacy on Vulnerability: From Human Agent to Human Oversight

As a result of the automatic and self-decision-making capabilities of AI and its increasingly autonomous nature, we are increasingly concerned about vulnerable subjects in contemporary times. It is imperative that a more stringent regulation mandate that one wears an ethical shirt and washes it with transparency. We can say that AI is more democratic, transparent and ethical to the extent that we can explain it to those who do not understand the system. Our main point was that the current regulations were aimed at protecting vulnerable subjects, but the rate of progress of AI was not willing to do so. What kind of regulation are we now moving towards? In our opinion, two elements are essential to the future of regulation.

- Building a proactive system with various phenomena. (Privacy by Design, Transparency by Design, Vulnerable Ethics Impact Assessment)
- Building a system of meaningful human oversight with the human agent and AI agent to eliminate the vulnerabilities of vulnerable subjects.

We have already mentioned in detail the transparency criteria and how human oversight will take action with these criteria. Nonetheless, all the narratives of the second significant substance are as urgent as daylight, rendering them more pertinent for Generative AI systems. Generative AI models may bring individuals face-to-face with risks with their interactive learning ability and ability to establish meaningful relationships with different systems and big data. It is plausible that among the high and unacceptable risk categories outlined in the AI Act, a system that combines human agents and AI agents may provide a controlled transition, provided that researchers

[140] Enqvist (n 85) 508-535.
[141] Enqvist (n 85) 508-535

attempt to mitigate the rigidity of these systems through the use of "Foundation Models."[142]

When speaking with the GPT-J-backed “Eliza” chatbot, the Belgian man's suicide reveals the need to take the regulation of the interaction of vulnerable subjects with AI even more seriously. The ex-wife of the man who asked the chatbot about climate change points to an important issue when he says that his wife would still be alive if it were not a chatbot. As we mentioned, the concept of anthropomorphic technology and the necessity of taking measures to protect the person interacting with it are more important than knowing that it is not human.[143]

We follow the innovative work of the closest autonomous vehicles, which are moving towards complete autonomy by imitating human behavior. Any variation of human oversight will also be tried in autonomous vehicles. Although studies continue on how much the human will exist in the cycle during the operation and cessation of autonomous vehicles that can be accepted from the high-risk category, the work done so far is as close as the human pressing a stop button. The time to press the stop button is not sufficiently perceived, and the human or autonomous responsibility for the defects that occur since the system is not adequately disclosed can lead to meaningful conclusions within the legal dialectic.[144] It has been stated that these studies will be completed by 2030 and that autonomous vehicles will have complete autonomy. TESLA recently announced that the vehicles possess autonomy; however, they were unable to proceed due to their inability to experience the system in the regulatory sandbox. It is noteworthy to mention that, regarding autonomous systems, human oversight will not commence at the moment the vehicles stop.[145] During the informing of the driver or passenger about basic situations, such as traffic rules, we encounter human oversight or human phenomena.[146] The incomplete realization of the information phase will be against human oversight or the concept of humans in the cycle. This implies that the number of people or actors in the cycle is not limited. In every hallway, the person who understands, interprets, or oversees is in the loop.[147]

Article 5 of the AI Act, which prohibits the implementation of particular AI systems, also calls for the oversight of these prohibited practices by the European AI Board (EUAIB). These applications of the group of human oversight experts created by the EUAIB should be made possible to be audited *ex ante* and *ex post*.[148] Although the

---

[142] H Felzmann, E Fosch-Villaronga et al., ‘Towards Transparency by Design for Artificial Intelligence’ (2020) 26 Science and Engineering Ethics 3333–3361.

[143] ‘”Suicida dopo chat con un’intelligenza artificiale”: il caso Eliza scuote il mondo dell’AI’ (2023) Europa Today <https://europa.today.it/attualita/parla-bot-suicida-moglie-colpa-intelligenza-artificiale.html> accessed 10 December 2023.

[144] A Saini, A Sciuto, J Forlizzi, JI Hong, ‘Hey Alexa, What’s Up?: A Mixed-Methods Studies of In-Home Conversational Agent Usage’ (2018) In Proceedings of the 2018 Designing Interactive Systems Conference <https://doi.org/10.1145/3196709.3196772> accessed 10 December 2023, 857–868.

[145] A Biswas, HC Wang, ‘Autonomous Vehicles Enabled by the Integration of IoT, Edge Intelligence, 5G, and Blockchain’ (2023) 23(4) Sensors 1963.

[146] Gaeta M. C. (2016), Automazione e responsabilità civile automobilistica. RESPONSABILITÀ CIVILE E PREVIDENZA, V, pp. 1718-1750, ISSN: 0391-187X.

[147] DA Abbink, ER Boer, M Mulder, ‘Sharing control with haptics: seamless driver support from manual to automatic control’ (2012) 54(5) Hum Factors 786–798.

[148] See: Information and Privacy Commissioner of Ontario, ‘Privacy by Design’ (2018) <https://www.ipc.on.ca/wp-content/uploads/2018/01/pbd.pdf> accessed 10 December 2023; M Hammerschmidt, N Mozafari, WH Weiger, ‘AI and Vulnerable’ (2022) Pacific Asia Conference on Information Systems <https://www.uni-goettingen.de/de/document/download/50ad96559eda96f5d43f97c0c83cab87.pdf/Mozafari-etal-2022_PACIS_AI-and-the-Vulnerable.pdf> accessed 13 December 2023.

methodologies to be employed are not based on the guidelines outlined in the AI Act, following the identification of prohibited practices, these inadequacies can be rectified through the recommendations and decisions of the EUAIB. Best practice guidelines published after the GDPR came into force and the European Data Protection Board's guidelines and explanations filled major gaps in implementation and resolved question marks.[149] Similarly, it is reasonable to expect that the EUAIB's decisions and reviews will shape the field. The flexible structure of the regulation and the lack of regulation are due to the rapid and surprisingly multidimensional progression of developments. Sanctions regarding developments can also be determined through commercial contracts and mutual commitments under the Incoterms approach. It is also crucial to consider the structure of the EUAIB and the actors who will have a say in it.[150]

Proactive by design means that the requirements are met to ensure human oversight of the system. The proactive by design, which can work with the transparency criterion, is to transfer the information embedded in the system to the other user and to prevent progress other than the purpose of the system when the system is self-learning.[151] Transparency also emphasizes that in automated decision-making systems and high-risk systems, it is imperative for the system to incorporate transparency within the cell as a virtuous behavior. It ensures that proactive behavior is conscious and that system phenomena, such as explainability, interpretability, and data protection, are effectively controlled.[152] It is important to have a conscious oversight of the system and fulfill human oversight as a conscious action to maintain transparency. In 1913, Brandeis said, "Sunlight was the best disinfectant." Transparency from this word is like sunlight for understanding and interpreting AI systems for the time being.[153]

Moreover, the inclusion in systems is also crucial for understanding that system and realizing the principles of democratic participation. Only a system that humans understand can be interpreted. But is one willing to trust a transparent system that humans can also interpret? Is trust a phenomenon that comes with transparency?[154] This issue remains controversial. Hirshman's sound and output theory is a meaningful example in this context. Increasing trust and loyalty in a system reduces separation from it. It is among the opinions that a more accurate and reliable system will be established with less transparency.[155] On the contrary, complete transparency may result in a violation of privacy. When a loss occurs, the developing risk approach becomes concerned with designing the system in a way that does not cause much harm before the

---

The "seven foundational principles" are: 1. Proactive not Reactive, Preventative not Remedial; 2. Privacy as the Default Setting; Privacy Embedded into Design; 4. Full Functionality — Positive-Sum, not Zero-Sum; 5. End-to-End Security — Full Lifecycle Protection; 6. Visibility and Transparency — Keep it Open; 7. Respect for User Privacy — Keep it User-Centric.

[149] HD Seval, SS Yilmaz, 'Mind Over Matter: Examining the Implications of Machine Brain Interfaces on Privacy and Data Protection Under the GDPR' (2022) 2 EJPLT 347-361.

[150] EDPS, 'EDPS'Final Recommendations on the AI Act' (2023) <https://edps.europa.eu/system/files/2023-10/EDPS-2023-12-Own-Initiative-Opinion-AI-ACT_EN.pdf> accessed 13 December 2023.

[151] Zalnieriute, Monika and Bennett Moses, Lyria and Williams, George, The Rule of Law 'By Design'? (2021). Tulane Law Review, Vol. 95(5), 2021, pp.1063-1101, Available at SSRN: https://ssrn.com/abstract=3805479

[152] J Cobbe, J Singh, 'Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges' (2021) 42 Computer Law & Security Review 105573.

[153] A Berger, 'Brandeis and The History of Transparency' (2009) <https://sunlightfoundation.com/2009/05/26/brandeis-and-the-history-of-transparency/> accessed 10 December 2023.

[154] E Fosch-Villaronga, H Felzmann et al., 'Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns' (2019) 6(1) Big Data & Society <https://doi.org/10.1177/2053951719860542> accessed 10 December 2023.

[155] AO Hirschman, Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States (Harvard University Press 1972).

damage has occurred. This brings with it a two-stage protection mechanism. Although the system is designed to protect fundamental rights and human dignity, the risk approach will come into play.[156]

Conclusion

In the event that the vulnerability of vulnerable individuals will increase in the face of AI, it is imperative that regulators and innovators prioritize their agenda first. In our study, we have attempted to present the right to request human oversight to eliminate vulnerability and our suggestions on how this right can be strengthened against GDPR and the AI Act. Legislation and court decisions aimed at eliminating the vulnerability of vulnerable subjects in public and private spaces and preventing discrimination in these areas are inadequate in the face of the vulnerability created by AI. We cannot simply address the damage with a risk-based approach, compensation, insurance, or standard development. Our experience indicates that generative AI models will result in a greater number of vulnerabilities than anticipated, and numerous segments that have yet to be identified will be susceptible to this development.

In the context of anthropomorphized technology, individuals will impose this qualification on technology. It is imperative that technology be fed with more human and AI agents. The basis of this study is to highlight the coexistence of the concept of human agent with the concept of AI agent, which will remove the barriers of vulnerability and allow meaningful human oversight. It is not easy to achieve. It is difficult for many disciplines to come together to come up with a model without understanding the behavior patterns of vulnerable subjects and investigating how they interact with AI. Despite our progress in human comprehension, we are still struggling to ensure that AI with human characteristics does not harm vulnerable subjects twice. The entitlement to request human oversight in the law of AI is not a matter of transparency, and the direct and continuous application of this intervention may impede innovation.

It is difficult for vulnerable subjects to decide when they can benefit from the right to request human oversight when they interact with AI. Individuals with disabilities, women who have been the victims of violence, immigrants, and children may encounter difficulties engaging with AI as it interacts with them. It is possible for some epilepsy patients to have epileptic seizures when AI is functioning rapidly. In our article, we mentioned a Belgian man who committed suicide. Although he was a man, he was not yet prepared for the information given to him by AI.

For effective use of the right to request human oversight, it is necessary to strengthen the human agent and anthropomorphize AI agents. The purpose of this is not to try to liken AI to humans, but to strengthen the right to request human oversight to protect against damage and to ensure the right of expression in the structure with which the vulnerable subject will interact. An AI agent or agent who does not understand basic aspects such as human behavior sciences, reactions, and emotions cannot provide human oversight to vulnerable subjects. During a trial, these agents may fail to report issues such as discrimination and violations of fundamental rights suffered by

[156] M Almada, 'Regulation by Design and the Governance of Technological Futures' (2023) European Journal of Risk Regulation 1-13.

vulnerable people. As previously mentioned, it is imperative that system outputs be interpreted to establish transparency. The way in which these outputs can be interpreted by individuals who are vulnerable is through the utilization of comprehensible language by AI agents. This understandable language will ensure a perfect association between human agents and AI agents. Then, AI will be able to become embodied for humans.

Throughout our study, the question we had in mind was how vulnerable people would have a presence in the system of an AI. It is evident that humans will be more involved than expected in the development of AI. When we predict that mechanisms to be embodied together will find meaning through interaction at the same time, the anthropomorphized structure needs to design agents correctly. So, we used the concept of vulnerable by design, and we understand the importance of studying the vulnerable impact assessment phenomenon.

With the awareness that we will protect everyone when we protect vulnerable subjects from the negative effects of AI, designing the involvement of more human agents and AI agents with historical, ethical and philosophical arguments will be one of the best solutions. There is little in most of them.