Unacceptable Risks in Human-AI Collaboration: Legal Prohibitions in Light of Cognition, Trust and Harm

Rostam J. Neuwirth [1], Sara Migliorini [2]

[1] *University of Macau, Avenida da Universidade, Taipa, Macau SAR, PRC*
[2] *University of Macau, Avenida da Universidade, Taipa, Macau SAR, PRC*

Abstract

Recent advances in technologies broadly referred to as "artificial intelligence" (AI) have increased the awareness about serious ethical concerns related to the impact of AI on societies and individuals alike. These concerns also aroused the attention of regulators and lawmakers around the world, which are leading to proposals to ban certain AI systems or practices. As a concrete example, the present paper discusses the prohibited AI practices listed in Artificial Intelligence Act (AIA) proposed by the European Union (EU). Given their cross-cutting nature, these AI practices raise a number of complex and transdisciplinary questions that will be addressed by reference to their impact on human trust and cognition, and the types of harm they may cause.

Keywords

EU AI Act, Prohibited AI Systems, AI Harm

1. Introduction

The global hype about the rapid development of technologies broadly referred to as artificial intelligence (AI), has recently been complemented by growing serious ethical concerns related to their use. The emerging global consensus about the risks and dangers inherent in AI were recognized in November 2021, when the 193 members United Nations Educational, Scientific and Cultural Organization (UNESCO) adopted the *Recommendation on the Ethics of AI* [UNESCO, 2022]. The said recommendation explicitly recognizes "the profound and dynamic positive and negative impacts of artificial intelligence (AI) on societies, environment, ecosystems and human lives, *including the human mind*".

The European Union (EU) has already moved beyond the stage of ethical recommendations when it issued a proposal in April 2021 for an Artificial Intelligence Act (AIA) as a comprehensive legally binding instrument with the goal of guaranteeing "a secure, trustworthy and ethical artificial intelligence" [European Commission, 2021]. The proposed AIA uses a risk-based approach and sets out to prohibit those kind of AI systems that pose unacceptable risks, in the sense that they contravene Union values by violating fundamental rights. Similarly, the Cyberspace Administration of the People's Republic of China (CAC) also released draft Administrative Measures for Generative Artificial Intelligence Services (Draft Measures) in April 2023, which complement earlier ethical recommendations and proceed to the formulation of binding rules to regulate AI.

Given the accelerating pace of the development of new AI technologies, their instant global cross-border availability, their combination with other technologies and especially their impact on the human mind, give rise to fundamental questions about the ways how human-AI interaction is best organized for the future to secure a trustworthy and secure use of these technologies. For these and other reasons,

Ethics and Trust in Human-AI Collaboration: Socio-Technical Approaches, August 21, 2023, MACAO, China
EMAIL: rjn@um.edu.mo; saramigliorini@um.edu.mo.
ORCID: 0000-0002-0641-5261 (A. 1); 0000-0002-1441-3442 (A. 2).

every jurisdiction is currently urgently called upon to ponder the optimal forms for legislative and regulatory actions to be taken. Such assessment needs to proceed from a broad, inclusive and transdisciplinary debate that focuses on a variety of issues. To this end, the present paper proceeds from the category of prohibited AI practices formulated by the EU AIA that are deemed to pose unacceptable risks.

The planned prohibition aimed to address those unacceptable risks particularly raises a set of complex transdisciplinary questions regarding the possibilities of different AI-supported technologies to cause serious harms to individuals, societies and humanity as a whole. To explain these unacceptable risks, Section II first briefly outlines each of the four legal categories of prohibited AI practices listed in the AIA, which entail subliminal and exploitative AI systems, social scoring systems and real-time' remote biometric identification systems. It also extracts some of their cross-cutting nature and intrinsic mutual connections, which raise a number of transdisciplinary issues. To clarify these issues, Section III discusses the concept of "harm" in relation to AI, and specifically to what extent certain AI-systems can impact individuals and society as a whole at a very fundamental level, undermining or rendering impossible the creation and maintenance of any trust bond, unless fully prohibited. Last, the article concludes by a brief assessment of the current state of the regulatory debate and brief outlook for future actions to be taken.

2. Prohibited AI Systems and the Transliminal Manipulation of the Mind

2.1. From Ethical Concerns to the Legal Prohibition of Certain AI Practices

The use of principles or ethical recommendations for the governance of AI has already been noted and the time has come for the adoption of binding rules [Munn, 2022]. In this respect, the EU's AIA represents the first attempt at a horizontal or comprehensive regulatory approach to AI, which is one adopting rules for all kinds of AI, rather than a vertical approach that focuses only on one specific aspect of AI sectorally. This poses difficult challenges due to the cross-cutting nature of AI, which means that they bear complex and multidimensional characteristics that call for innovative, cross-sectoral, and transdisciplinary policy responses. This approach bears significant problems, because it requires not only to guarantee the internal consistency of the relevant act, but also to ensure its coherence with other existing and future acts or laws. On the other hand, this approach offers important advantages in terms of the goal of future-proofing the regulatory framework in light of the rapid pace of development of AI and their mutual convergence with other technologies.

In order to cover all kinds of AI systems but also contain the complex and dynamic nature of AI, the AIA has opted for a broad and, to some degree, flexible definition of AI combined with a risk-based approach, which categorizes AI into three levels of risks, namely 1) unacceptable risk, 2) high risk, and 3) low or minimal risk. The category of unacceptable risks is to be understood to comprise all those AI systems the use of which is considered unacceptable as contravening Union values, for instance, by violating fundamental rights. The AIA plans to eliminate these unacceptable risks by prohibiting four categories of AI practices listed in Article 5 AIA. These categories include a) an AI system that deploys subliminal techniques beyond a person's consciousness, b) AI system that exploits any of the vulnerabilities of a specific group of persons, c) AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons (social scoring system), and d) 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement.

Overall, each of these four categories of prohibited AI practices is well-defined as it is made subject to further requirements [Neuwirth, 2023b]]. However, the division in four categories also gives rise to multiple problems. First and foremost, it is possible that an AI system will be disassociated from other

technologies it is combined with in its use, which may result in a situation where a certain AI system will not be captured by the letter any of the four categories, but still violate the spirit of all of them combined. Second and no less important, each of these categories raises important transdisciplinary questions of notably a legal, psychological, neurological and technological nature, which also require a sound scientific and transdisciplinary inquiry into the underlying issues.

2.2. Legal, Cognitive and Technological Aspects of Prohibited AI Practices

The serious and complex regulatory difficulties related to these prohibited AI practices are best exemplified by the first category, the one of so-called "subliminal AI systems". The initial proposal of the AI does not define these subliminal AI systems but merely prohibits the placing on the market of "an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm" (Art. 5(1) lit. a) AIA). This single sentence alone gives rise to multiple complex and transdisciplinary questions [Neuwirth, 2023a].

First, the provision requires a sound understanding of human perception and cognition, namely whether there exists an absolute threshold of perception that can be drawn between cognitive functions being carried out in a subliminal or a supraliminal manner, meaning that they are either below or above the threshold of a person's awareness. This is important to be able to respond to doubts regarding the effectiveness of subliminal techniques to influence a person's thoughts and actions. Even if it were possible to draw such a line, it further requires to answer the question of whether it needs to be drawn for all of the human senses or just one. This in turn also demands a scientific answer to the open questions of how many human senses there are and how the known (and unknown ones) may interact and influence each other.

Additionally, these questions regarding human cognition must be evaluated in close connection with technological issues, given that these technologies themselves are developed using scientific insights from psychological and neurological studies. In this regard, it is also necessary to analyse both existing and future technologies that might qualify as subliminal AI systems. For instance, several experiments with a technology referred to as "spyware" have been carried out and found to be able to extract sensitive private data from a person's mind, such as banking information, PIN codes, addresses or birth data, from the brains of users of Brain-Computer Interfaces (BCIs) by showing them different (subliminal or supraliminal) stimuli and using machine learning models [Martinovic et al, 2021; Steinhagen and Kettani, 2020]. More recently, as study was published that offers non-invasive ways using functional magnetic resonance imaging (fMRI) to decode brain stimuli in a way to reconstruct continuous natural language, which practically promises to make machine-assisted "mind reading" a technological reality [Tang et al., 2023].

Similar problems also persists for the remaining three categories of prohibited AI practices. For instance, for AI systems that exploit any of the vulnerabilities of a specific group of persons, open questions concern the matter of what constitutes a person's vulnerability and when as well how a technology might be capable of exploiting it. In this context, it was noted that every person potentially or actually has a weak point, and displays some vulnerabilities at certain times during their life, and even during a single day, which makes everyone overall particularly vulnerable. Used in combination with other technologies, such as wearable health devices, emotion recognition technology, or big data, it will be easier to detect and exploit vulnerabilities using, for instance, targeted advertising or dark patterns [Neuwirth, 2023b].

Various open questions pertaining to human nature of a scientific nature also emerge in the context of the third and fourth categories of prohibited AI practices. For social scoring systems, the bundled use of technologies, such as big data, artificial intelligence, cybernetics and behavioural economics, give rise to the need to evaluate how they "are shaping our society—for better or worse" [Helbing et al, 2019]. Last but not least, the fourth category of "'real-time' remote biometric identification systems" similarly gives rise to many open scientific, technological and legal questions. To give but one example, it is important to note that the definition of biometrics adopted by Art. 3 (33) AIA not only covers

personal data resulting from specific technical processing relating to the physical, physiological but also behavioural characteristics.

Cutting a very long story short, the proposed AIA is in many instances based on factual assumptions related to the human being and its modes of perception and cognition that are incomplete or deficient. As regards human cognition, the current state of science can be understood to mean that the majority of cognitive functions are carried out below the threshold of awareness [Mlodinow, 2012], but that the same threshold of awareness varies not only inter- but also intrapersonally [Smith and McCulloch, 2012]. For this reason, the use of the phrase "subliminal" should actually be replaced by "transliminal" in the final version of Art. 5 AIA. Equally, it is not yet sufficiently known how many human senses there are and how they interact and influence each other but in any case a multisensory or synaesthetic approach should be taken by regulators and judges in this regard [Neuwirth, 2023a].

Most of all, particular attention must be paid to the rapid advances in these technologies and the possibilities for their complex mutual combinations. The reason is that it is in their creative combination that lies the greatest potential to enhance their effectiveness in manipulating thoughts and behaviour. In this regard, it was found that the overall efficiency of subliminal stimuli can be expected to be enhanced by the combination of the collection of big data, profiling, targeted advertisements and other "immersive" and neuromarketing-based applications [Reitberger et al., 2011].

To briefly restate, there are serious difficulties encountered in efforts to regulate AI, which are caused by their rapid pace of innovation, their cross-cutting and cross-boundary as well as transdisciplinary nature. In sum, these factors also require a holistic understanding of the legal system as a whole and the analysis of specific legal concepts and instruments that are essential to regulate the technology. Since the European lawmaker has proposed to downright prohibit certain AI systems and practices, because they pose an acceptable risk, it is necessary to discuss what the materialization of such risks, in the form of harm could be, before going back to how the legal systems shall protect society from such harms and assess whether the proposed framework in Art. 5 AIA is appropriate.

3. New Types of Harms, Ethics and Trust

The notion of harm, or injury, is one of the building blocks of the law of torts, which, in general terms, is the system of legal remedies available to the victim of a wrongfully inflicted harm. Approaches to tort varies grandly across jurisdictions. However, one common, key aspect is that only some individual and collective harms are recognized as "actionable". Harm is actionable when the interest that has been harmed is considered worthy of being protected by the legal system, which therefore offers the victim a legal remedy. One classic example of actionable harm is personal injury, which protects people's physical integrity. Other types of harms are also traditionally recognized, such as the psychological consequences of an accident.

As mentioned in the previous section, the advent of AI systems, such as the ones capable of transliminal manipulation, has prompted discussion about the risks they pose to society. But a proper conceptualization of the harms that can be the potential materialization of such risks has not yet been carried out.

To conceptualize harm, there are two background trends to consider. Firstly, historically, tort law has evolved precisely as a response to technological innovation, and the inherently accrued risks of industrial societies [Priest, 1985]. Entire branches of law have emerged from tort law as a response to industrialization, such as insurance law and product liability. It is therefore to be expected that the rise of new types of machines will shake the foundations of tort law as well, in particular of the notion of actionable damage. Secondly, over the past 60 years, many systems of tort law have evolved under the influence of human rights reasoning. Human rights, or fundamental rights, are a list of individual rights and interests that are recognized and protected by the legal system at the highest level in the hierarchy of norms. The influence of these rights across jurisdictions have led to the recognition of specific harms, notably in the field of privacy [UKHL, 2004]. It would be logical to expect that fundamental rights should play a crucial role in steering the legal systems towards correctly identifying the type of unacceptable harms related to the deployment of certain AI-systems.

Against the backdrop of these two trends, it is important to look, first, at the harms that can be connected to the prohibited practices in human-AI collaboration and, secondly, whether preventing such harms requires a legal prohibition, or if a consent-based system could be sufficient.

3.1. Individual and Collective Harms in Human-AI Collaboration

Covering all negative consequences arising from each of the prohibited AI practices is difficult in view of their scope and rapid development. What is most relevant here, in light of cognition, harm and trust, is to define broadly two aspects of to the new harm that arises from these practices.

The first aspect concerns how all the prohibited practices affect the fundamental rights to dignity, personhood and freedom of expression. In short, common to all the practices prohibited in Article 5(1) AIA, is an aspect of surveillance and external control, which infringes upon some basic human needs and fundamental rights. Transliminal practices interfere with the very essence of our thought processes. Biometrics use our unchangeable and unique bodily and behavioral characteristics, from fingerprints to gait, to put an identity on, or categorize humans. In this process, no respect is paid to accuracy of identification, or to the views of the identified person regarding the identity attributed to them. Social scoring places individuals in a constant state of observation and judgement, which directly affects their ability to participate fully in society.

In so doing, all these systems infringe upon the right of each individual to self-represent themselves in their social relations, and freely express their personality [Schwartz, 1968]. One key aspect of self-definition and representation in social relationships is the possibility to step out of all our public identities, and enjoy unobserved time. The deployment of the prohibited AI systems directly infringes upon this need, which is protected by fundamental rights, at the highest level of the hierarchy of norms.

The second aspect, that is also common to all the prohibited practices, is their harmful effect on society as a whole. This type of harm is newer for the law of torts, but it is not unheard of for the legal system, for example when a collective interest, such as the right to future generations to a healthy environment, is legally recognized [UNHRC, 2021]. In a similar fashion, it is essential to acknowledge that unobserved time enables humans "to develop intellectually and emotionally, by giving us breathing room to embrace risks and make mistakes without the stigma of being forever associated with failures and fads [Cohen, 2013]. To the contrary, research has shown that societies where surveillance levels are high are less creative and encounter greater feelings of anxiety [Solove and Citron, 2017]. In addition, these negative effects are bound to affect individuals pertaining to certain groups in a more significant way, contributing to exacerbate the inequalities rooted in our society [Allen, 1988]. In so doing, the prohibited practices affect freedom of expression of individuals, but also harm society in its entirety, precisely because the vast majority of individuals in it are or feel observed and judged, and thereby are restricted in the exercise of their fundamental right to express their personality.

Because of these important effects on individuals and their fundamental rights, as well as society as a whole, we submit that the prohibited AI systems must be considered as leading to harms, which must be actionable if the legal system has to maintain its coherence and efficiency. Indeed, if a practice infringes upon the exercise of a fundamental rights, it must be considered that it is harmful to humans. However, in many jurisdictions, the system of tort law seems to remain anchored to the old categorizations, rooted in personal injury and psychological consequences of an personal injury as the main and practically only types of harms that can lead to legal action and compensation. Instead, we belive that interdisciplinary research should sustain a review of the concept of harms when related to the use of AI systems that affect people's fundamental rights.

3.2 The Ethical Limits of Consenting to Harm and Human-AI Trust

An additional issue, from the point of view of the legal system, is the relationship between harm and consent. In field of privacy and data protection, legal systems have generally embraced consent as their cornerstone, in order to justify restrictions of such fundamental rights. However, the EU lawmaker has proposed to outright prohibit the mentioned AI systems. In order to maintain proportionality and to

allow the deployment of AI systems, could a different regulatory choice other than prohibition, such as a consent-based system, be appropriate? After all, legal systems usually associate a person's consent with values that are considered fundamental, such as personhood, autonomy and individual self-determination, which are exactly those that the prohibited AI systems seem to affect. And, traditionally, tolerance for privacy intrusions rests on the liberal premise that individuals should be able to determine autonomously the degree of intrusion in their private life that they are willing to tolerate, including through granting access to personal data.

Indeed, the fetishization of consent as the ultimate key to freedom and individual self-determination has led to a very restrictive view of the extent to which the commercial exploitation of a person's own personal data, including sensitive ones, can be restricted. This has been true even in cases in which fundamental ethical issues were at stake.

However, the advent of big data, the development of machine learning techniques and the necessary infrastructure, have rendered this system based on consent obsolete, because, in the context of this new data-based economy, consent is incapable of achieving the very thing that it was designed to do: protect human dignity and freedom to self-determinate.

As a matter of fact, invasions of privacy are routinely allowed by individuals for trivial reasons, such as convenience, while control on information about oneself is in fact, a primary or foundational good, a fundamental need, "on which access to many other goods rests" [Allen, 2011]. In a same way, facial recognition technology and other biometrics cause an effect of mass surveillance, and the attached harm, that derive from many individuals consenting to it [Selinger and Hartzog, 2020].As a consequence, in our digital society, we put forward the idea that privacy and personal data, in particular, sensitive ones, should not be consentable, i.e. it should not be possible to "sell" one's data or privacy in exchange for a service, such as access to a social media.

Protection of human dignity and self-determination cannot therefore derive from a system based on consent. In the case of Art. 5(1) AIA, the lawmaker has chosen to prohibit a series of harmful systems and uses of AI in a view to avoid significant harms on individual and society. Putting aside for a moment the shortcoming of the drafting of the AIA, it seems that this regulatory choice is a sound one.

If we accept that human-AI collaboration can yield advantages in many sectors of society, and may even lead humans towards a cognitive evolution, such collaboration can only be fruitful if it is based on a bond of trust. Such bond of trust cannot exist if manipulative and other types of invasive techniques of surveillance are routinely allowed, under pretext that the human has "consented" to them. True freedom to self-determinate and respect for human dignity can only exist in a world that does not tolerate practices that pose unacceptable risks that can lead to unacceptable harms, including upon an individual's own consent.

4. Conclusion

Following the emergence of a global consensus regarding the ethical concerns related to AI as formulated in the 2021 UNESCO Recommendation on the Ethics of AI, legislators around the world are pondering the different regulatory ways to ensure a safe, trustworthy and sustainable use of AI. The present article briefly presented the serious dangers connected to the "prohibited AI practices" in the Artificial Intelligence Act proposed by the EU. These AI practices include the four categories of subliminal and exploitative AI systems, social scoring systems and remote biometric identification systems, which underscore the cross-cutting nature of AI due to their growing combination with other technologies, such as different neurotechnologies or biometric identification systems. Based on the convergence and intrinsic connections between different technologies, the article highlighted the need to closely link the global regulatory debate with a sound transdisciplinary inquiry into the dangers posed by existing as well as future technologies. The reason is that they pose dangers that also require a critical re-evaluation of the foundations of each legal system as much as of specific legal concepts, such as notably harm. It also requires a full reconsideration of the rules on causation and remoteness, through which the law of torts attributes harm to a certain behavior, and decides at which point it cannot be legally considered that there is a causal link between the two. In this respect, transdisciplinarity and the

latest scientific findings need to find a proper venue within the rules on causation, when considering harms related to AI systems, and in particular the prohibited practices.

5. References

[Allen, 1988] Anita Allen. *Uneasy access: Privacy for women in a free society*. Totowa: Rowman & Littlefield, 1988.

[Allen, 2011] Anita Allen. *Unpopular privacy: What must we hide?*. Oxford University Press, 2011.

[Cohen, 2013] Julie E Cohen. What privacy is for. *Harvard Law Review* 126: 1904-1933, 2013.

[European Commission, 2021] European Commission. *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. COM (2021) 206 final, 21 April 2021.

[Helbing *et al*., 2019] Dirk Helbing *et al*. Will Democracy Survive Big Data and Artificial Intelligence? In *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*, pages 73-98, Springer, Cham, 2019.

[Martinovic *et al*., 2021] Ivan Martinovic *et al*. On the Feasibility of Side-Channel Attacks with Brain-Computer Interfaces. *Proceedings of the 21st USENIX Security Symposium*, 143–158, 2021; https://www.usenix.org/conference/usenixsecurity12/technical-sessions/presentation/martinovic.

[Mlodinow, 2012] Leonard Mlodinow. *Subliminal: How Your Unconscious Mind Rules Your Behaviour*. Vintage Books, New York, 2012.

[Munn, 2022] Luke Munn. The Uselessness of AI Ethics. *AI Ethics*, 1–9, 2022; https://doi.org/10.1007/s43681-022-00209-w.

[Neuwirth, 2023a] Rostam J. Neuwirth. *The EU Artificial Intelligence Act: Regulating Subliminal AI Systems*. Routledge, New York, 2023.

[Neuwirth, 2023b] Rostam J. Neuwirth. Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act (AIA). *Computer Law & Security Review* 48:105798, 2023; https://doi.org/10.1016/j.clsr.2023.105798.

[Priest, 1985]G. L. Priest, 'The Invention of Enterprise Liability: A Critical History of the Intellectual Foundations of Modern Tort Law' (1985) Journ. Leg. Stud. 461

[Reitberger *et al*., 2011] Wolfgang Reitberger *et al*. Ambient Persuasion in the Shopping Context. In *Pervasive Advertising*, pages 309-323, Springer, London, 2011.

[Schwartz, 1968] Barry Schwartz. The social psychology of privacy. *American Journal of Sociology* 73:741-752, 1968.

[Selinger and Hartzog, 2020] Evan Selinger and Woodrow Hartzog. *The inconsentability of facial surveillance*. Loyola Law Review 66:33, 2020.

[Smith and McCulloch, 2012] Pamela K. Smith and Kathleen McCulloch. Subliminal Perception. In *Encyclopedia of Human Behavior*, vol. 1, 2nd ed., pages 551-557, Academic Press, London, 2012.

[Solove and Citron, 2017] Daniel J Solove and Danielle Keats Citron. Risk and anxiety: A theory of data-breach harms. *Texas Law Review*. 96:737, 2017.

[Steinhagen and Kettani, 2020] Dustin Steinhagen and Houssain Kettani. An Inventory of Existing Neuroprivacy Controls. *ICISDM 2020: Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, 77–83, 2020; https://doi.org/10.1145/3404663.3404664.

[Tang *et al*., 2023] Jerry Tang *et al*. Semantic reconstruction of continuous language from non-invasive brain recordings. bioRxiv, 2023; https://doi.org/10.1101/2022.09.29.509744.

[UKHL, 2004] *Campbell v Mirror Group Newspapers Ltd* [2004] UKHL 22

[UNESCO, 2022] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. UNESCO, Paris 2022.

[UNHRC, 2021] Resolution of 8 October 2021, the UN Human Rights Council (UNHRC), https://documents-dds-ny.un.org/doc/UNDOC/LTD/G21/270/15/PDF/G2127015.pdf?OpenElement.