**BALTIC JOURNAL OF LAW & POLITICS**
A Journal of Vytautas Magnus University
VOLUME 117, NUMBER 2 (2024)
ISSN 2029-0454

Cit.: *Baltic Journal of Law & Politics* 17:2 (2024): 116–141
https://content.sciendo.com/view/journals/bjlp/bjlp-overview.xml
DOI: 10.2478/bjlp-2024-00018

# ARTIFICIAL INTELLIGENCE IN MANIPULATION: THE SIGNIFICANCE AND STRATEGIES FOR PREVENTION

**Eglė Kavoliūnaitė-Ragauskienė**

**Doctor, Researcher**
**Law Institute of Lithuanian Centre for Social Sciences, Lithuania**

**Contact information**
Address: A. Goštauto g. 12, LT-01108 Vilnius, Lithuania
Phone: +370 61830599
E-mail address: egle.kavoliunaite@teise.org

Received: August 17, 2024; reviews: 2; accepted: November 13, 2024

## ABSTRACT

Manipulation has always been a persistent aspect of society, but the rise of artificial intelligence has significantly amplified its scope and impact. AI's ability to access and process vast amounts of personal data has revolutionized the way individuals can be targeted with specific information. This technology enables the rapid delivery of highly personalized messages, making it easier to exploit psychological vulnerabilities and influence decision-making without the individual's awareness.

The use of AI for manipulation may pose serious threats to personal autonomy, as individuals are increasingly nudged towards choices they might not have made independently. Beyond personal implications, this also may endanger human rights and the foundations of democracy. By subtly shaping public opinion and electoral behaviour, AI-driven manipulation can undermine the integrity of democratic processes, leading to outcomes that do not truly reflect the will of the people. Acknowledging that not all uses of AI in manipulation are necessarily unacceptable, this article examines the possible approaches to tackle AI's undesirable impacts in democratic societies rather than manipulation *per se*. It explores various tools and methods of AI employed for manipulation, identifying key areas where this technology is most used for unacceptable and harmful purpose in democratic society, such as in advertising, political campaigns, and the spread of misinformation. It also analyses the legislative approach of the EU to limit the usage of undesired manipulative AI and explores additional strategies, including ethical guidelines, public education,

and technological solutions to mitigate the risks and prevent the misuse of AI for manipulative purposes.

**KEYWORDS**

Artificial Intelligence (AI), Manipulation, Personal autonomy, Democracy.

## INTRODUCTION

The use of artificial intelligence (which is often called just AI) in public life is expanding both in terms of areas and methods being used. Nowadays it is common to see AI used in ensuring public safety, conducting traffic surveillance, collecting data for various public sector reports, marketing purposes, dating portals, etc. Artificial intelligence has already become an important part of our lives. As one of the European Union's first documents on artificial intelligence[1] points out, it is no longer science fiction but a reality, from a virtual personal assistant to organize the work day, to traveling in a self-driving vehicle and phones that suggest songs or restaurants we might like. Artificial intelligence not only makes our lives easier, but also helps solve some of the world's biggest challenges: from treating chronic diseases to reducing traffic accident deaths.

However, the rapid development of AI has not only brought positive outcomes, but also raises concerns about the adequate protection of human rights. The fast and barely unregulated development of artificial intelligence poses various challenges related to mental integrity of the individual, non-discrimination, impartiality, the threat of curtailing the freedoms of assembly and expression, changing the preferences and opinions of persons by insisting that they buy certain goods or services or change their political, religious views or beliefs[2]. Moreover, AI based manipulation may alter persons' political views and preferences and thus can erode trust in institutions, media and bring threat to democracy[3].

This undesired effect of the use of the AI is also officially recognised in the EU legislation. In particular, the EU Artificial Intelligence Act[4] states that "aside from the many beneficial uses of artificial intelligence, that technology can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices".

In this context, on the other side, it is important that European Union has set a goal to lead the world in the development and use of artificial intelligence in various processes of public and private life. From the policy documents adopted at the level of the European Union[5], it is clear that the European Union intends to compete with other regions of the world in the development and use of artificial intelligence systems and is making significant intellectual, economic and strategic efforts for this. No doubt that other countries have set similar goals and are ready to compete with the European

---

1 *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe* (COM/2018/237 final), part 1, par. 1.

2 Viktorija Morozovaite, "Hypernudging in the changing European regulatory landscape for digital markets," *Policy Internet* 15 (2022): 86 // DOI: 10.1002/poi3.329

3 Sefano Faraoni, "Persuasive technology and computational manipulation: Hypernudging out of mental self-determination," *Frontiers in Artificial Intelligence* 6 (2023) // DOI: 10.3389/frai.2023.1216340

4 *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)* (OJ L, 2024/1689), recital 28.

5 E.g., see *European Commission White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*, Brussels, 19.2.2020 (COM (2020) 65 final).

Union in this race. Therefore, the limitations of the development and the use of AI should be set with a precise purpose and clarity and while respecting the human rights not to impede the development of technologies, including AI.

This article aims at determining the appropriate approach towards the restrictions of the deployment of AI for manipulative purposes, as well as suggesting the possible means of limiting such practices. To this end the article first will provide an overview of concepts of manipulation and its main elements with a view to provide the criteria to determine the types of manipulation which should be restricted in democratic societies. Later on, it will provide a recap on the variety of AI tools and techniques used for manipulation in general and having a malicious intent in particular. Further, the article will present an overview of the fields of social life where AI-based manipulation in used mostly and in what ways it is done. After this it will analyse the threats posed to human rights and the democracies by the manipulation executed using AI tools. Next, it will systematize and introduce the main legal instruments intended to reduce or mitigate the negative impact of manipulation using AI. Finally, it will describe non-legal possible directions which need more discovery to prevent the likely infringements of human rights and personal autonomy brought by the use of AI tools. These tasks are reflected in the respective six chapters of the article.

The article provides theoretical insights which are based on analysis of secondary sources (previous research works, political (mostly European Union level) and non-governmental institutions documents and insights, as well as of the primary sources of the European Union legislation and policy documents. The theoretical resources and legislative/policy documents were selected in accordance with two criteria: the relevance to the topic and potential to provide value to the research. The analysis of the collected data deploys different methods used in theoretical analysis: historical method was used to understand and describe the development of analysed concepts and practices, critical analysis was used to assess the level of importance, credibility of theories and approaches, systemic analysis method helped grouping, separating and interrelating the relevant information in the research data, deductive method was used to investigate the actual and possible outcomes of implementation of certain ideas, whereas inductive method was invoked to draw conclusions from the researched data. Standard legal analysis was used in studying primary sources of law, invoking the legal dogmatic, content analysis and descriptive methods.

There are different theories to approach the limitations of certain practices as posing certain threats. One part of theories concentrates on limitations on values, others – on limitations on the effect[6]. The article will discuss and assess which approach would be more appropriate to limit the undesired effect of usage of AI in manipulation.

As the article provides theoretical analysis, it will not involve any empirical data usage, so the scope of the findings will be limited to theoretical assumptions. Also taking into account the aim of the article to assess the possibilities to restrict the negative effect of the usage of AI based manipulation, the biggest attention will be given to analysis of the research data which is closely connected to this questions. Therefore, for example, the article will not discuss in detail other possible ontologies on manipulation

[6] E.g. see distinction between the design for values and conceptual engineering and design for non- manipulation in Michael Klenk, "Ethics of Generative AI and Manipulation: a Design-Oriented Research Agenda," *Ethics and Information Technology* (2024) 26: 9 // DOI: 10.1007/s10676-024-09745-x. This theoretical distinction is used in other fields of public life – another example might be *pro-life* and *pro-choice* approaches used in the discussion on the beginning of human life (see. e.g. Eglė Kavoliūnaitė, "Concepts of the beginning of legal protection of human life," *Teisės problemos* 2 (48) (2005): 8.

using AI[7], the different available definitions and concepts of manipulation and elements thereof which are important for different contexts, also, the technical side of AI capabilities for manipulating the data will be presented only to particularly address the analysed issue.

The article also has values related limitations as it is based on the assumption that the limitations of certain practices should seek to protect human rights and other values of a democratic society. Therefore, the criteria in determining the acceptable and unacceptable manipulation practices are already predefined on the basis of democratic values.

## 1. WHAT IS MANIPULATION?

As regards the concept of manipulation, "AI manipulation" or "AI based manipulation" may be understood as any influence exerted through the use of AI technology that is intentionally designed to bypass reason and to produce an asymmetry of outcome between the manipulator and the target. However, as will be discussed, the notion and elements of "manipulation", its impact and harm may be understood differently by invoking different theoretical approaches.

The term "manipulation" has a long history and has been analysed and used in many contexts – politics, medicine, social life and many more. As will be seen from the further analysis, the tools used for manipulation change with ages, however, the idea and the common features of this phenomenon stay similar.

For example, in Ancient Greece, manipulation was mainly seen as the use of clever and misleading arguments to convince others, usually to achieve a personal goal. The sophists, who were skilled teachers of rhetoric, were often criticized for using manipulation to win arguments and gain power over people[8]. Even earlier, in Ancient Egypt there were examples of purposive distortion of reality when it was attempted to erase people from history, later known as *damnatio memoriae*[9]. Although nowadays in public discussions we more often use a word "persuasion" rather than manipulation, and these two terms may seem similarly used to express forms of influence on others, there is a clear distinction between them. Whereas "persuasion" involves presenting arguments that are truthful and fair, and which aim to convince others through reason rather than deception, "manipulation" is typically understood as the use of clever or deceptive arguments to persuade others, often with the goal of achieving some self-interested end[10].

In clinical psychology, manipulation refers to behaviours aimed at controlling or influencing others for personal gain, often disregarding their well-being[11]. These behaviours can include lying, guilt-tripping, gaslighting, or other forms of emotional manipulation. It's considered a significant issue in relationships, especially when one

---

7 This topic is researched by other authors, e.g. see Micah Carroll, Alan Chan, Henry Ashton, and David Krueger, "Characterizing manipulation from AI systems": 1–13; in: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (2023).

8 Marcelo Ienca, "On Artificial Intelligence and Manipulation", *Topoi* 42 (2023): 834 // DOI: 10.1007/s11245-023-09940-3

9 Noémi Bontridderp and Yves Poullet, "The Role of Artificial Intelligence in Disinformation," *Data & Policy* 3 (2021): e32-2 // DOI: 10.1017/dap.2021.20

10 Marcelo Ienca, *supra* note 8, 836.

11 Ahona Guha, "Understanding and Managing Psychological Manipulation," *Psychology Today* // https://www.psychologytoday.com/us/blog/prisons-and-pathos/202104/understanding-and-managing-psychological-manipulation; Marcelo Ienca, *supra* note 8, 834.

person seeks to exploit or control another. Manipulation as a general rule involves power imbalance.

The Cambridge online dictionary offers the following definition of manipulation – it is "controlling someone or something to your own advantage, often unfairly or dishonestly"[12]. In many cases manipulation is considered ethically wrong because it involves influencing someone's behaviour or beliefs in a non-transparent way that (i) undermines their autonomy, freedom, or dignity, (ii) promotes the personal gain of the manipulator at the expense of the manipulated, and (iii) may result in direct or indirect harm for the manipulated[13].

Researchers often point at manipulation's harmful impact on autonomy[14].

Other authors empathise other features of manipulation, however, these are very similar to the already described ones. For example, according to Susser, Roessler and Nissenbaum, manipulation is an attempt to subvert conscious decision-making through hidden influence and it involves bypassing reflective and deliberative capacities through covertness. Thus, one of the important features of manipulation is that a person does not notice that he does not take a decision not independently, but upon someone's intension[15]. On the other hand, it might be cases when the forms of manipulation are relatively less negative. For example, nudging which in general in society is seen as positive and desirable practice, may also regarded as a form of manipulation which infringes human autonomy[16]. Also, as Klenk points out, it may be that the positive impact of some public health communication driven by manipulative generative AI may outweigh the negative value that accrues from the manipulative nature of the influence. Similarly, some users may willingly adopt personal health assistants that use effective but manipulative influence tactics[17].

As will be discussed later in this article, it is important that manipulation often involves the exploitation of cognitive, emotional, or other decision-making vulnerabilities of people. Manipulation affects emotions, as rationality and critical thinking fade into background. At the same time, it is always aimed at a specific vulnerability of a person in the process of interpersonal communication. In the science of psychology these emotional triggers and values that are affected by the manipulator, are called the manipulation target[18]. Therefore, in order to resist manipulation, besides other means, one of the core actions is to properly limit the possibilities of a manipulator to get to know the manipulatory target.

To summarise, the manipulation is a term used to describe a usually negative exertion of influence upon a person which displays features of intention which is usually

---

12 Cambridge online dictionary, *Manipulation* // https://dictionary.cambridge.org/dictionary/english/manipulation

13 Marcelo Ienca, *supra* note 8, 835.

14 Marietjie Botes, "Autonomy and the social dilemma of online manipulative behavior," *AI Ethics* 3 (2023): 315–323 // DOI 1007/s43681-022-00157-5; Daniel Susser, Beate Roessler and Helen Nissenbaum, "Technology, Autonomy, and Manipulation," *Internet Policy Review* 8(2)-1 (2019): 10–11 // DOI 10.14763/2019.2.1410; Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, "Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges," (2023): 57 // DOI: 10.1007/978-3-031-17040-9; European Commission Independent High Level Expert Group on Artificial Intelligence, *Ethics guidelines for trustworthy AI,* (2019): 17 // https://ec.europa.eu/newsroom/dae/document.cfm? doc_id=60419

15 Daniel Susser, Beate Roessler and Helen Nissenbaum, *supra* note 14, 11; Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, *supra* note 14, 57.

16 T. Martin Wilkinson, "Nudging and Manipulation," *Political Studies* 61, 2 (2013): 344.

17 Mikhael Klenk, *supra* note 6, 10.

18 Nikita Nepryakhin, "Classification of vulnerability factors in the process of psychological manipulation," Conference *International Conference on Advanced Research in Social Sciences* (2019) // DOI: 10.33422/icarss.2019.03.93

aimed the benefit of the manipulator, and is related to the violation of autonomy by using vulnerabilities of a particular individual. Depending on a concept, manipulation not in all cases is harmful, as a person might know that she/he is being influenced, agrees with this for the sake of the final result she/he is seeking (e.g. health or sports results). Thus although there are certain infringements of persons' autonomy (e.g. in case of nudging), or persons may not know whether what they see was performed genuinely or was created using AI (e.g. a movie scene), the important criteria in defining whether the AI based manipulation is undesired and should be restricted in a democratic society is whether it has a negative impact, which usually is regarded as harm, loss or damage to an individual or the public.

## 2. AI AS A TOOL FOR MANIPULATION

AI is a constantly evolving technological arrangement. It evolved through different stages – first it was related to systems' capabilities to perform tasks intelligently in specific areas without allowing these areas to be united to produce a more comprehensive intelligence, later interrelating capabilities were developed enabling to solve problems and do automated planning[19] which evolved into generative AI and the start of general AI. According to OECD, General AI refers to the idea that general human intelligence could be matched or even surpassed by machines. Some research uses the term "Generative AI", "Artificial Super Intelligence", or "superintelligence", to refer to systems that could far surpass the abilities of humans[20]. In exerting manipulation, the main difference between the generative AI and the general AI is the subject which stands behind the manipulation act. In case of generative AI, the AI is just a tool to achieve the result as ordered by a human manipulator, however, in case of General AI (which is understood as a superpower beyond human capabilities) the manipulator would be the AI itself without humans knowing the aim and method of manipulation. What is important here, if an AI system operates with a different framework of understanding than humans do, it can be challenging to interpret its behaviour in terms of incentives, secrecy, or intent. For instance, the AI's approach to planning might not resemble the familiar process of attempting to influence human thought, even if its actions ultimately have that effect[21]. Thus the ontological approach towards the manipulation, its aim, planning, methods, effect, etc. used by the AI would be different from the ontology used by humans. Therefore, the analysis of this article is limited to the scope of simpler forms of AI, in particular, the AI is seemed as a tool for manipulation performed by humans on other humans rather than the generator and initiator of a manipulative act.

Researchers assert that the use of AI is always related to bias. AI can be biased both at the system and the data or input level and this occurs either intentionally or unintentionally. Bias at the system level involves developers building their own personal biases into the parameters they consider or the labels they define. This happens either when developers allow systems to conflate correlation with causation, or when

---

[19] OECD Working Papers on Public Governance. *Hello, World. Artificial intelligence and its use in the public sector,* (2019): 56 // https://www.oecd-ilibrary.org/docserver/726fd39d-en.pdf?expires=1666023983&id=id&accname=guest&checksum=2B0678AFDEB937C7A5B42575C7C96F44; European Commission/ Deloitte *Opportunities and Challenges for the Use of Artificial Intelligence in Border Control, Migration and Security*, Vol. 2: Addendum (2018): 15–16 // https://op.europa.eu/en/publication-detail/-/publication/69f33ff7-a156-11ea-9d2d-01aa75ed71a1/language-en

[20] OECD Working Papers on Public Governance, *supra* note 19, 15.

[21] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger, *supra* note 7, 13.

developers choose to include parameters that are proxies for known bias. Bias at the data or input level occurs in a number of ways. First, the use of historical data that is biased. Because AI systems use an existing body of data to identify patterns, any bias in that data is naturally reproduced. Selection bias appear when the input data are not representative of the target population, when the input data are poorly selected or when the data are incomplete, incorrect, or outdated[22]. As regards the use of AI as a tool for manipulation, it is always connected with a purposeful bias.

There are several AI techniques which are widely exploited for manipulation.

The use of algorithms to mimic human behaviour dates back to early computing, with social bots being used for manipulation since at least 2010[23]. However, here the focus lies on the new ability to create realistic content (text and images) to support such campaigns. Manipulative content generation can be viewed in three stages: early crude bots, more sophisticated deepfakes, and now generative AI.

Crudefakes involved simple bots producing synthetic, automated text content. These bots lacked meaningful interaction capabilities, making them easy to detect.

Deepfakes, a word combined of two words "deep learning" and "fake", are created using two AI algorithms within a Generative Adversarial Network (GAN). GANs generate new data by analysing existing datasets. For instance, a GAN could study thousands of pictures of a politician and produce a new, similar image that isn't an exact replica of any single one. This technology applies to various types of content, including images, videos, sound, and text[24]. Deepfakes are related to more advanced bots with humanlike features, such as scraping the internet for content, using natural posting cycles, and limited user interaction. Deepfakes are already widely used advances in audiovisual manipulation that generate more accessible and realistic falsifications of events and of people's faces, words and actions[25]. AI improved both the content and the messenger, making detection harder. In 2017, a software developer nicknamed deepfakes on Reddit online platform posted his creations that he swapped the Hollywood celebrities' faces onto the faces of porn artists. After the creations spread rapidly, deepfakes became a new trend[26]. Quite a lot of deepfakes are being generated for entertainment purpose and are available on Youtube and other social platforms. The most common manipulation is switching the main actors with other persons. For example, a short film named Home Stallone is a reference to the 1990 comedy movie Home Alone. It features an 8 year-old AI generated Sylvester Stallone instead of the real actor, Macaulay Culkin. However, deepfakes are quite often used for political purposes, for example, in 2020 Legislative Assembly elections in Delhi; also in 2020 The environmental movement Extinction Rebellion used deepfake technology to alter a nearly 5 min speech of Belgian premier Sophie Wilmès, as if she was speaking of the COVID-19 pandemic as a consequence of environmental destruction, etc.[27].

---

[22] AccessNow, *Human Rights in the Age of Artificial Intelligence*, (2018): 12 // https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf

[23] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini, "The Rise of Social Bots," *Communications of the ACM* Vol. 59, No. 7 (2016).

[24] Noémi Bontridderp and Yves Poullet, *supra* note 9, 3.

[25] Sam Gregory, "Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism," *Journalism* Vol. 0(0) (2021): 3 // DOI 10.1177/14648849211060644

[26] Vikrant Sopan Yadav, "AI and Human Rights: A Critical Ethico-Legal Overview," *Agathos; Iasi* Vol. 14, Issue. 1 (2023): 264, 29; Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková and Daniel Thilo Schroeder, "Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes," *Front. Commun*. 6: 632317 (2021): 5 // DOI: 10.3389/fcomm.2021.632317

[27] Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková and Daniel Thilo Schroeder, *supra* note 26, 8.

Nowadays manipulators use generative AI, a technological leap forward that blurs the line in terms of what is detectable as real versus synthetic content, by humans in particular but also through machine means[28]. To this end the social media is a perfect medium for the wide use of manipulative AI tools. Social media platforms are designed with various interactive features, like creating profiles, publishing posts, sharing with others, commenting, and reacting to content. They serve as hubs for communication, social connections, and networking among users. Social media is used for a big variety of purposes – entertainment, information dissemination, political engagement, and business promotion and advertising, etc[29]. Manipulation in social media is being alleviated by using filter bubbles (advanced AI algorithms allows controlling the content users see, allowing advertisers to target specific demographics with customized ads); fake accounts and bots (which can manipulate social media by spreading misinformation, amplifying certain voices, or creating the illusion of widespread support for specific ideas); content moderation (systems used to censor certain voices or ideas); amplification of extreme or sensational content; addictive potential (e.g. endless scrolling, notifications, and autoplay). These features increase user exposure to advertisements, generate more data, and ultimately boost platform profits[30].

The variety of methods how to reach users and to make use of their vulnerability empowers the manipulators to individualise the content delivered or the way of delivery of the content to target users. This technique is called microtargeting. It enables the manipulators to deliver personalized messages to specific individuals or groups based on their demographic, behavioural, or psychographic characteristics. This involves collecting and analysing large amounts of data from sources like social media, search engines, and third-party data brokers to create highly customized campaigns. Information and communication platforms developed tracking methods using AI technologies to enable the targeting of advertisements to specific consumers for the sake of efficiency. Using cookies and other means like advertising networks, social media widgets, and website analytics scripts allows different parties to re-identify users across different websites and build a browsing history[31]. The goal is to present messages that resonate deeply with individual users, increasing the chances they will engage with the ad or take a desired action, such as making a purchase or sharing the content[32].

The development of generative AI brought a new perspective in the manipulation activities as it increased credibility of the messenger rather than the message itself. Generative AI can indeed produce higher-quality false or misleading messages, but this is only a minor improvement. The truly revolutionary change is the emergence of extensive bot networks that appear human, creating text, images, video and audio, to enhance the perceived authenticity of the messenger. For example, a person might be interacting with someone online who is sharing the same interests: music, hobbies, sports, they might be speaking the same language variation, share posts about their camping with a family, a dog lying on the rug in a living room, family celebrations,

---

28 William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0. Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI,* (2003): 4 // https://www.rand.org/pubs/perspectives/PEA2679-1.html; Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini, "The Rise of Social Bots," *Communications of the ACM* Vol. 59, No. 7 (July 2016): 99 // DOI 10.1145/2818717

29 Marcelo Ienca, *supra* note 8, 838.

30 *Ibid*., 838–839.

31 Noémi Bontridderp and Yves Poullet, *supra* note 9, 4.

32 Marcelo Ienca, *supra* note 8, 839.

parties with friends, etc. and receive responses from similar other users. However, from time to time they share their political views. And there might be tens, hundreds or thousands of similar online people who in reality are AI generated persons created on a deliberate attempt to influence public opinion through social media manipulation.

The use of AI in internet, with the intention to manipulate other users, is constantly increasing[33]. The development of AI makes the large scale manipulation even more accessible to a wide range of actors, as social bots appear more and more similar to humans making this difficult to apprehend; the use of AI for mass manipulation requires less and less resources, generative AI scales well and at the same time the detection thereof becomes more and more complex[34].

Although AI is only a tool enabling or alleviating the manipulation, it already changed and has a potential to further change the manipulative practices used at a very large scale. Cohen pointed out three aspects how AI strengthens the power of manipulation. The first aspect is the ability to systematically find weaknesses in how individuals or small groups make decisions and to adjust appeals based on that. By using large datasets and machine-learning techniques, manipulators can personalize their approaches for each person and further refine these interactions using feedback, all while the person remains unaware of how this is happening. In comparison, when manipulation does not involve machine learning, a human manipulator might rely on knowledge gained from their relationship with a target about that person's compulsions or addictions, or they may draw on general research about cognitive biases to steer the individual toward specific choices. While a human may have the advantage of personal familiarity with a potential target or insights from behaviour studies, a machine learning system can create detailed and dynamically evolving profiles of behaviour that identify and exploit weaknesses in the target's decision-making processes. These abilities arise not just from machine learning techniques but also from various features of digital environments. Factors like the portability of devices, the widespread use of sensors and tracking technologies, the compatibility of online services, and advancements in data storage have all contributed to the gathering of large amounts of constantly updated data. Machine learning can quickly transform this data into predictions about vulnerabilities and can reveal hidden patterns, associations, or correlations while continuously learning from new information over time without needing explicit instructions. In fast-changing digital environments, machine learning algorithms can help speed up the process of customizing and targeting specific groups of persons or particular individuals. The second aspect is the massive volume and rapid pace of distributing personalized information made possible by machine learning in digital spaces. For example, automated systems can buy and place ads, matching them with millions of consumers in less than a second. This allows for customized, dynamic messages to be delivered at the right moment and on a large scale. Major digital platforms use machine learning in their ad delivery systems to estimate the likelihood that a user will take a desired action after seeing an ad, allowing them to quickly learn and adjust those predictions based on responses like clicks, views, or purchases. The appearance of these tailored and targeted messages on digital platforms helps ensure that the intent to change

[33] See, e.g. EUROPOL, *Facing reality? Law enforcement and the challenge of deepfakes* (2022): 5–6 // https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf

[34] William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *supra* note 28, 5.

behaviour is hidden from the individual being targeted. The third characteristic, which gives machines a significant advantage over humans, is the relentless nature of AI systems. Unlike the human nature, the processing capabilities of an AI system do not get tired or are affected by any other human feature[35].

There is a very thin line between the issues described above, where the initial commander of the manipulation act is human (oriented at a certain result rather than selection of methods) and the previously discussed possibility of AI performing manipulation on humans. Although a human would expect that AI performs what is commanded using individually selected manipulative actions, as AI would have the capabilities to make certain decisions on its own, there is a risk that a human would not capture the manipulative AI behaviour against him, as AI might start acting on its own ontology.

## 3. WHERE IS AI-BASED MANIPULATION MOST USED?

The possibility to individualise the content and the conveyer of the message to a particular user opened broad possibilities for manipulation with the help of AI. There are several fields where the possibilities of manipulative AI is used at a very broad scale, especially making use of the social networks, as described above.

First, the manipulation using AI is widely used in boosting business through individualised advertising. AI is used for intended profiling when an organisation seeks to identify potential customers to target with advertising campaigns[36]. As already explained before, the AI tools created possibilities for the data analysis social media companies use to create profiles of users for targeted advertising[37] and thus the collecting of data from social media enables to use manipulation for increasing corporate profits[38].

Initially the microtargeting tools were largely developed by corporations like Google and Facebook, who identified advertising as a means of using the of personal information they collect about internet users for commercial purpose[39]. As Ryan Calo put it, "the digitization of commerce dramatically alters the capacity of firms to influence consumers at a personal level. A specific set of emerging technologies and techniques will empower corporations to discover and exploit the limits of each individual consumer's ability to pursue his or her own self-interest"[40]. This results in placing a person to an "information bubble" where the purchases are determined not by the person autonomously, but they are suggested by the surrounding information and advises. Moreover, as human emotions and behaviours fluctuate, with people experiencing highs and lows, periods of greater resilience, and times of increased vulnerability, this may also be targeted by manipulative AI tools. A 2013 marketing study indicates that there are predictable moments when individuals are more likely to feel vulnerable. For instance, women in the US from various demographic backgrounds

35 Tegan Cohen, "Regulating Manipulative Artificial Intelligence," *SCRIPTed* 20, 1 (2003): 216–218 // DOI: 10.2966/scrip.200123.203
36 Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, *supra* note 14, 26.
37 AccessNow, *supra* note 22, 16.
38 Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, *supra* note 14, 56.
39 Daniel Susser, Beate Roessler and Helen Nissenbaum, *supra* note 14, 2; Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization," *Journal of Information Technology* 30(1) (2015): 77–78 // DOI: 10.1057/jit.2015.5
40 Daniel Susser, Beate Roessler and Helen Nissenbaum, *supra* note 14, 2; Ryan Calo, "Digital Market Manipulation," *The George Washington Law Review* 82(4) (2014): 998 // https://digitalcommons.law.uw.edu/faculty-articles/25

tend to feel least attractive on Mondays, making them potentially more receptive to purchasing beauty products[41].

Manipulation in advertising is used not only in promoting sales of regular goods, but also of certain important services, e.g. healthcare. For example, an AI system could be designed to recommend different treatments depending on the insurance status of the patient or how much they are able to pay, potentially denying lifesaving care to someone because of their socioeconomic status, harming marginalized groups who already suffer from insufficient access to quality healthcare[42].

It is also claimed that as the AI gains more control over information and shapes human behaviour to fit business models, they are likely to reinforce and expand power imbalances among citizens, businesses, and governments. This control allows corporations to exploit their positions, potentially imposing the negative consequences of faulty data on consumers and workers. This even can deepen societal inequalities and increase the influence and authority of businesses and governments over individuals, further entrenching power disparities[43].

The second field where the use of microtargeting is extensively used is political campaigns, and, in particular, elections. Following concerns raised in 2016 and 2017 about using information technology to influence global elections, it became clear that the threat of targeted advertising extends beyond the commercial realm. Political campaigns can leverage ad targeting platforms like Facebook, YouTube, and other social media services to significantly influence voter behaviour and decisions. The Cambridge Analytica scandal, which revealed the company's use of voter profiling and targeted ads to exploit psychological vulnerabilities in the United States, United Kingdom, France, Germany, and beyond, sparked global outrage and heightened public awareness of these issues[44]. Moreover, in certain countries manipulative AI tools are used not only during the election campaigns, but also as a propaganda tool in national (e.g. in China AI is the main tool in the party-state propaganda apparatus and is one of the means to strengthen the military[45]) and foreign policy (China uses Generative AI broadly in relations with Taiwan through polarisation of the population and manipulations in elections)[46].

Manipulative AI is also used in other important areas of society life thus making the ordinary sources of evidence and news non-reliable. For example, deep fakes may be used to distort evidence in court proceedings thus jeopardising the use of photos, video and sound recording as reliable evidence.

Also, deepfake videos massively target female celebrities in sports, entertainment, and fashion. Many of these videos are instances of involuntary pornography, where the target's face is superimposed onto a pornographic video[47]. Initially, creating such deepfakes required numerous images of the target, limiting the focus to

---

41 PHD Media, *New beauty study reveals days, times and occasions when U.S. women feel least attractive,* (October 2012) // https://www.prnewswire.com/news-releases/new-beauty-study-reveals-days-times-and-occasions-when-us-women-feel-least-attractive-226131921.html

42 AccessNow, *supra* note 22, 27.

43 Onur Bakiner, "The promises and challenges of addressing artificial intelligence with human rights," *Big Data & Society* (July–December 2023): 9 // DOI: 10.1177/20539517231205476

44 Daniel Susser, Beate Roessler and Helen Nissenbaum, *supra* note 14, 2; also see AccessNow, *supra* note 22, 25; Sam Gregory, *supra* note 27, 9; Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, *supra* note 14, 53; William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *supra* note 28, 1.

45 William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *supra* note 28, 14.

46 *Ibid*., 20.

47 Sensity, *The world's first detection platform for deepfakes*, (2020) // https://sensity.ai/

celebrities. However, advancements in deepfake technology using generative adversarial networks now allow for the targeting of individuals with only a few available images. Deepfake is also used to impersonate individuals. For example, there was a case where criminals used audio manipulation to mimic a CEO's voice during a phone call, convincing an employee to transfer a large sum of money to a private account[48]. Deep fake technology has numerous potential criminal applications, including its use by sexual predators to impersonate minors.

In all these fields, AI's ability to exploit individual and group vulnerabilities poses significant ethical concerns. It not only influences personal choices but also threatens the integrity of public discourse and democratic processes, making it a powerful tool for manipulation on a scale never seen before.

As described in the previous chapter, the different reality-distorting methods used by AI, which sometimes are clearly manipulative, do not intent to bring any harm and are being created just for entertainment purpose or are being accepted in the society (age reducing photo filters, nudging apps, etc.). However, as described in this chapter, there are several areas of public life, where AI based manipulation is used very broadly and due to its effect unsupported in democratic societies, might raise legal and ethical concerns. The particular harm of manipulation performed using AI is discussed in the next chapter.

## 4. HARM OF AI-BASED MANIPULATION

As described previously, although manipulation in certain circumstances or depending on a particular description may be acceptable in the society, there are certain fields where manipulation brings harm to the values of democratic societies. In many case the harm to individuals, at larger scale, has a negative effect on the whole society.

Encouraging individuals to shop or gamble against their own interests can result in financial and psychological harm. On a larger scale, these practices can lead to market inefficiencies by prompting excessive consumption of harmful goods or causing people to spend time and money[49] where they would not want initially thus threatening the welfare of people, thus endangering social rights of individuals.

At an individual level, the manipulative AI infringes individuals' cognitive liberty, which is described as the individuals' right to access and control their own thoughts, emotions, and perceptions, free from interference or coercion by others. According to a number of researchers cognitive liberty is an essential component of individual autonomy and human dignity, and a necessary prerequisite of several other freedoms such as freedom of speech and freedom of expression[50]. It was also mentioned before that AI tools facilitating manipulation threaten autonomy of persons. European Court of Human Rights in the Pretty case stated that the notion of personal autonomy underlies the right to privacy[51]. The right to privacy may be violated as the AI algorithms make use of personal data for profiling purpose, i.e. for determining the content to be

[48] Johannes Langguth, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková and Daniel Thilo Schroeder, *supra* note 26, 8.

[49] Tegan Cohen, *supra* note 35, 213.

[50] Marcelo Ienca, *supra* note 8, 837; Nita A. Farahany, "The Costs of Changing Our Minds", *69 Emory L. J. 75* (2019)// https://scholarlycommons.law.emory.edu/elj/vol69/iss1/2; Jan Christoph Bublitz and Reinhard Merkel "Crimes against minds: On mental manipulations, harms and a human right to mental self-determination," *Crim Law Philos* 8(1) (2014): 51–77; Marcello Ienca and Effy Vayena, "Digital Nudging Exploring the Ethical Boundaries": in: Veliz (ed.), *The Oxford Handbook of Digital Ethics* (Oxford University Press, Oxford, 2021).

[51] *Pretty v. United Kingdom*, European Court of Human Rights (2002, Application no. 2346/02), para 61.

displayed individually and distorting individual's capacity to take a decision without any influence[52].

However, consequences of the manipulative use of AI are evident at a broader scale rather than individual level. In the beginning social media, micro-blogging, and video-blogging possibilities were praised for promoting civil–political rights, in particular the freedom of expression[53]. However, the social media combined with the vast spread of usage of AI makes it far more easier to manipulate not only individuals, but even a great part of the society thus threatening democracy. AI has important repercussions for democracy, and people's right to a private life and dignity. For instance, when as described in previous chapter AI can be used to determine people's political beliefs, then individuals in our society might become susceptible to manipulation. Political strategists could use this information to identify which voters are likely to be persuaded to change party affiliation, or to increase or decrease their probability of turning out to vote, and then to apply resources to persuade them to do so. Manipulating someone to vote or abstain from voting for a particular candidate strips them of their voting autonomy and infringes on their democratic rights. Widespread voter manipulation undermines the democratic process's legitimacy, which relies on free and informed voting. Even minor voter manipulation can erode trust in democratic institutions. Overall, the impact of manipulation varies depending on the context, affecting individuals, communities, and society differently[54]. Citizens' ability to make free and informed decisions is compromised by the digital ecosystem, which also impacts their participation in democratic discourse. When individuals' capacity to engage fully in public debate is hindered, including through the manipulation of their opinions and voting choices, the integrity of democracy is seriously jeopardized[55]. Committee of Ministers of the Council of Europe noted that

> fine grained, sub-conscious and personalised levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions. These effects remain underexplored but cannot be underestimated. Not only may they weaken the exercise and enjoyment of individual human rights, but they may lead to the corrosion of the very foundation of the Council of Europe. Its central pillars of human rights, democracy and the rule of law are grounded on the fundamental belief in the equality and dignity of all humans as independent moral agents[56].

Sometimes it seems that deepfakes may have only an entertainment purpose and effect. However, even such use of deepfakes in the media can have harmful societal impacts. They may not only contribute to the spread of false information but also undermine the credibility of legitimate sources, fostering scepticism about all information, even from traditional press or governmental agencies. In the realm of disinformation, this allows manipulators to deceive individuals in two powerful ways: presenting fake content as real and dismissing authentic information as fake[57]. As noted in the Access

---

52 Yves Poullet, *Ethique et droits de l'Homme dans notre société du numérique* (Brussels: Académie Royale de Belgique, 2020), 114.
53 Onur Bakiner, *supra* note 43, 2.
54 Tegan Cohen, *supra* note 35, 213–214.
55 Noémi Bontridderp and Yves Poullet, *supra* note 9, 6.
56 Council of Europe, *Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes*, Para. 9 (Decl (13/02/2019)1).
57 Noémi Bontridderp and Yves Poullet, *supra* note 9, 4.

Now study[58], the AI behind deep fakes continues to advance, and there is "potential for sowing chaos, instigating conflict, and further causing a crisis of truth". Therefore, the media society is deeply concerned about the discreditation of media as it has become very difficult for an average user of media to distinguish deep fakes and other manipulative AI content from genuine information content[59].

Thus, as stated in the Access Now study, manipulative AI "created new forms of oppression, and in many cases disproportionately affects the most powerless and vulnerable. The concept of human rights addresses power differentials and provides individuals, and the organizations that represent them, with the language and procedures to contest the actions of more powerful actors, such as states and corporations.[60]" This leads to a situation described by Hannah Arendt back in 1978:

> … A people that no longer can believe anything cannot make up its own mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please[61].

## 5. LEGAL INSTRUMENTS TO COPE WITH THE USE OF AI FOR MANIPULATIVE PURPOSES

United Nations International Covenant on Economic, Social and Cultural Rights adopted back in 1966 recognize the right of everyone to enjoy the benefits of scientific progress and its applications[62]. Paradoxically, now concerns are growing how to ensure that the use of emerging technologies do not jeopardize human rights.

There are different initiatives globally seeking to reduce the negative impact of possible manipulations, and this trend is not a new one. Previously, similar initiatives emerged seeking control the mass media powers in the society. The emergence of mass media in the early 20th century raised new ethical issues concerning the use of advertising to sway consumer behaviour. Critics contended that advertising was frequently misleading, manipulative, and exploitative, breaching principles of autonomy, honesty, and respect for individuals. These concerns led to the establishment of ethical standards and regulations for advertising[63]. However, the previous legal attempts to fight manipulation were rather isolated.

The possible infringement of the right to autonomy, along with multiple other related rights by using AI generated manipulation is mentioned in a number of international and regional documents[64]. Nowadays one of the pioneering legal documents in

---

58 AccessNow, *supra* note 22, 16.

59 See e. g. Sam Gregory, "Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI," *Journal of Human Rights Practice* 15 (2023): 702–714 // DOI: 10.1093/jhuman/huad035; Sam Gregory, *supra* note 25; Sam Gregory, "Synthetic media forces us to understand how media gets made," *NiemanLab Predictions for Journalism* (2023) // https://www.niemanlab.org/2022/12/synthetic-media-forces-us-to-understand-how-media-gets-made/; Sam Gregory and Eric French, "How do we work together to detect AI-manipulated media?" *Witness Media Lab* (2019) // https://lab.witness.org/projects/osint-digital-forensics/

60 AccessNow, *supra* note 22, 17.

61 Hannah Arendt, "Hannah Arendt: From an Interview" (Interview with Roger Errera in 1974), *The New York Review of Books* (1978).

62 *United Nations International Covenant on Economic, Social and Cultural Rights* (1966), art. 15, sec. 1(b) // https://www.ohchr.org/sites/default/files/cescr.pdf

63 Marcelo Ienca, *supra* note 8, 835.

64 E. g., see OHCHR, *Report on the Right to Privacy in the Digital Age* (UN doc A/HRC/39/29, Geneva, 2018); UNGA, Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General (UN Doc A/73/348, New York City, NY: UNGA, 2018); UNDESA, *Resource Guide on Artificial Intelligence (AI) Strategies* (New York City, NY: UNDESA, 2021); OECD, *Dark Commercial Patterns* (OECD Digital Economy Papers n. 366. DSTI/CP (2021)12/FINAL., Paris, 2022), etc.

democratic societies to tackle the manipulation enabled by the use of AI is expected to be the EU AI Act, adopted in July 2024[65].

The AI Act reflects a risk-based approach, setting out a four-tiered regime of rules and oversight. AI systems will attract different levels of oversight and rules depending on the system's risk profile. The AI act puts the manipulative AI to the group of unacceptable risks as it prohibits the "subliminal techniques" in respect of all persons, and the exploitation of vulnerabilities of persons. In particular, it prohibits the „placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm" and „the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm[66]. As explained in the AI Act Explanatory Memorandum, the rationale for seeking to prohibit AI-enabled subliminal techniques and the exploitation of vulnerable groups is based on the 'significant potential' of such practices to manipulate individuals[67]. As seen from the wording of the AI Act, the manipulative practices using AI are prohibited only when they cause "significant harm" to person or persons. Thus the AI act excludes from prohibition manipulative practices which are acceptable in the society and / or used for positive impact (such as nudging, or similar).

The second aspect of manipulative practices using AI regulated by the AI Act is the transparency requirement. Initially, the idea of the AI act was to require the disclosure of deep fakes in all cases[68], however, Article 50(4) of the AI Act requires deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake to disclose that the content has been artificially generated or manipulated (except the cases of investigating a criminal activity). However, it is stated that „where the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work". As regards texts which are published with the purpose of informing the public on matters of public interest, it is required to disclose that the text has been artificially generated or manipulated (except the cases of criminal investigation or where the AI-generated content has undergone a process of human review or editorial control and where a natural or legal person holds editorial responsibility for the publication of the content).

---

65 *Artificial Intelligence Act*, *supra* note 4.

66 *Artificial Intelligence Act, supra* note 4, art. 5 sec. 1(a).

67 *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), Explanatory Memorandum, par. 5.2.2.

68 Article 52(3) of the *Proposal to the AI Act* (*supra* note 67) stated: "Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ("deep fake"), shall disclose that the content has been artificially generated or manipulated."

The actual implementation of this provision will be seen in the near future, however, it is already clear that although the requirement to disclose the AI generative nature of the information related the public interest only when it appears as a text, and possibility not to disclose it when the information is in other format (e.g. a photo or a video) intended for entertainment, there might be endeavours of extortion of public information on particular events or persons (e.g. politics) under the guise of entertainment.

Just before the adoption of the AI Act a number of stakeholders raised concerns about the threats of manipulation using AI and the need to control it in the AI act[69], but these voices were in fact left unheard. For example, it is stated that although the AI Act addresses two key aspects of manipulation, i.e. hidden influence and exploiting vulnerabilities, it still lacks sufficiency in some crucial areas. Firstly, the Act focuses on hidden content, such as subliminal techniques, which are not clearly perceptible. While these are indeed forms of hidden influence, the real power of manipulative AI lies in its ability to identify and exploit individual vulnerabilities in perception, attention, and decision-making. Secondly, although the AI Act prohibits AI systems that exploit vulnerabilities, it only considers group-level vulnerability, assuming that certain groups, like those defined by age, are susceptible to manipulation while others are not[70]. It is argued that manipulation is related to making use of vulnerabilities of all persons, not only the officially determined "vulnerable groups" which as a general rule includes children and persons with disabilities. As vulnerabilities are intrinsic to all people (and without making use of vulnerabilities there would be no manipulation as according to the definitions described in the previous sections of this article, the targeting of vulnerabilities is one of the element of manipulation) it would be reasonable to change the standard used in the AI Act for prohibition of malicious use of AI used in respect of "vulnerable groups", at least to "average consumer", as provided for in the Unfair Commercial Practices Directive[71]. Finally, definition of manipulation and its harms is lacking in this document. Thus, according to the critics, the Act's prohibitions only partly address the issue of manipulative AI, which increasingly targets individual weaknesses to exert hidden influence[72].

Necessary to mention that the Explanatory Memorandum to the AI Act also explains that other manipulative or exploitative practices affecting adults that might be facilitated by AI systems could be covered by the existing data protection, consumer protection and digital service legislation that guarantee that natural persons are properly informed and have free choice not to be subject to profiling or other practices that might affect their behaviour.

Indeed, the General Data Protection Regulation[73] and Directive on privacy and electronic communications[74] restrict the collection, accessibility and storing of data

---

[69] European Commission, *Public Consultation on the AI White Paper. Final report*, (November 2020) // https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68462

[70] Tegan Cohen, *supra* note 35, 206–207.

[71] *Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No. 2006/2004 of the European Parliament and of the Council ("Unfair Commercial Practices Directive")* (OJ L 149, 2005).

[72] Tegan Cohen, *supra* note 35, 206–207.

[73] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* (OJ L 119, 2016).

[74] *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)* (OJ L 201, 2002).

which later can be used for profiling. However, even strict requirements do not always work as expected. It is claimed that for many individuals, the vast array of data protection policies and cookie banners they encounter while navigating digital services – now essential to their social, economic, and political lives – offer limited practical benefit. These services often require users to share data about their online behaviour as a condition of access. Machine learning models typically rely on a representative sample to make inferences about a broader population, even when only limited data is available for some individuals who share similar observable traits. As a result, even the most cautious users who carefully read privacy policies, adjust their cookie settings, and choose to withhold certain information may still be impacted by others who freely share their data. This creates an environment where trying to control the flow of behavioural data that fuels manipulative AI becomes increasingly challenging, and sometimes, futile[75]. The complexity of the threat posed by user profiling was highlighted in the case where the European Commission conducted a thorough investigation into Google's acquisition of Fitbit. The concerns centred around advertising, particularly that the acquisition would further expand Google's already vast data pool, enhancing its ability to personalize ads. This raised fears that competitors would struggle to match Google's dominance in online search advertising, ultimately creating higher entry barriers and limiting growth opportunities for competitors. As a result, advertisers could face higher costs and fewer options[76].

Recently enforced Digital Services Act supplements the AI Act provisions as relates prohibition of deployment of "subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour" as it prohibits the providers of online platforms from designing, organising or operating their online interfaces in a way that deceives or manipulates the recipients of their service or in a way that otherwise materially distorts or impairs the ability of the recipients of their service to make free and informed decisions[77]. As this act entered into force in February 2024, its real effect is not clear yet, however, it is highly expected that the manipulation in the online platforms will be reduced.

The challenge of proper regulation of AI malicious practices is compounded by the fact that AI's capabilities evolve rapidly, outpacing the development of legal tools designed to control its misuse.

The existing legal instruments, including the GDPR and the AI Act, are important steps forward but are not sufficient to fully address the complexities and potential harms of AI-based manipulation. As AI continues to permeate various aspects of life, the legal system should evolve to keep pace with these technological advancements and the new forms of manipulation they enable and should be supported by other educational and soft law initiatives.

## 6. OTHER MEANS TO TACKLE MANIPULATION USING AI

As described above, there are some legal regulations emerging aimed at tackling possible malicious use of AI. However, one of the most often used critics towards legal regulation is that law only intends to rectify the undesired practices while failing to address

[75] Tegan Cohen, *supra* note 35, 231–232.
[76] Bernd Carsten Stahl, Doris Schroeder and Rowena Rodrigues, *supra* note 14, 42.
[77] *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)*, art. 25(1) (OJ L 277, 2022).

the social–structural causes underlying them. On the other hand, the too strict rules or overregulation might lead to unjustified restrictions of technological development, the right of expression and other related rights. Therefore, as in case of the previously mentioned soft law regulation, law usually only provides recommendations rather than strict requirements in the field of use of results of technological advancements. For example, the Council of Europe in its Declaration on the manipulative capabilities of algorithmic processes encourages member States to assume their responsibility to address the "threats to the right of human beings to form opinions and take decisions independently of automated systems, which emanates from advanced digital technologies", and especially "to their capacity to use personal and non-personal data to sort and micro-target people, to identify individual vulnerabilities and exploit accurate predictive knowledge, and to reconfigure social environments in order to meet specific goals and vested interests"[78].

Assuming the limited effect of the legal regulation to tackle the undesired use of AI, this chapter is devoted to discuss soft law and deliver recommendations on the other possible directions to restrict the manipulative use of AI.

### 6.1. Ethical rules

First of all, in order to supplement the strict legal regulation, ethical principles come into play. For example, High-Level Expert Group on Artificial Intelligence established by the European Commission adopted Ethics Guidelines for Trustworthy AI[79] which set out seven requirements for trustworthy AI, based on four ethical principles.

Beyond the EU, there are several other important documents and initiatives worldwide which address the threats of manipulation using AI, emphasizing the need for ethical standards, transparency, and accountability to mitigate these risks. OECD AI Principles[80] highlight the importance of transparency, fairness, and accountability in AI systems. They call for measures to ensure AI does not perpetuate bias or manipulate users and emphasize protecting human rights and democratic values. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[81] provides guidelines for the ethical design of AI systems, including considerations to prevent manipulation and ensure user autonomy. It stresses the importance of transparency, accountability, and respecting user consent. NIST AI Risk Management Framework[82] includes guidelines to ensure AI is trustworthy and transparent. It addresses the potential for AI to manipulate users and emphasizes the need for ethical safeguards. UNESCO Recommendation on the Ethics of Artificial Intelligence explicitly addresses the risks of manipulation using AI. It emphasizes the need for ethical principles to govern AI, including protecting human rights, ensuring transparency, and preventing the misuse of AI for manipulative purposes. Each of these documents highlights the importance of ethical guidelines, transparency, and accountability to safeguard against the manipulation of users through AI systems. They collectively aim to ensure that AI technologies are developed and deployed in ways that respect individual autonomy and democratic values.

---

78 Council of Europe. *Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes* (2019) (Decl (13/02/2019)1).

79 European Commission Independent High-Level Group on Artificial Intelligence, *supra* note 14.

80 OECD, *AI principles*, (2024) // https://www.oecd.org/en/topics/sub-issues/ai-principles.html

81 IEEE Standards Association, *The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems*, (2023) // https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/

82 NIST, *AI Risk Management Framework*, (2024) // https://www.nist.gov/itl/ai-risk-management-framework

However, being more ethical guidelines rather than mandatory documents, they have a limited impact to the factual usage of manipulative AI.

### 6.2. Education and digital literacy

When users are informed about the risks and harms of digital manipulation, they can better recognize and resist such attempts. This might involve sharing information on common manipulative tactics, like fake news and phishing scams, and encouraging the development of critical thinking skills[83]. Education might be adapted to different stakeholders – users or daily information in social media targeted at them as buyers of particular goods or services (e.g., explaining how to recognize hidden advertising), also as participants in the elections (how dirty political schemes are working and how AI may be used there). Such in-depth education is relevant to journalists in order to allow to recognise reliable and not reliable information to build their articles and announcements to the general public in the media. Also special education programmes may be prepared to politics or military heads. These types of trainings may include very particular aspects, such as tips how to recognise an AI generated photo (for example although such images may be photorealistic, they often have visual anomalies (e.g., too many fingers) that can be detected with a close eye[84]).

### 6.3. Technical restrictions of using AI

This direction of managing AI posed threats is being investigated and even enforced (e.g. strict limitations to use facial recognition technologies), but still in general the technical restrictions on AI raise many discussions. As already mentioned, the boost of the use of manipulation is directly related with the development of technology. However, manipulatory practices are created using a wide range of AI tools, therefore, the restrictions on technology rather than content/channels of distribution would mean the restrictions on the AI itself. Even speaking about limitations to deep fakes alone would threaten to make different kind of satire, humour, internet memes, and possibly even movies illegal. This might lead to censorship which would be unacceptable in democratic societies. This was also taken into account in regulating the deepfake technologies in the AI Act. Besides, the European Court of Human Rights in the case *Handyside v. United Kingdom* stated that freedom of expression "is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference, but also to those that offend, shock or disturb the State or any sector of the population. Such are the demands of that pluralism, tolerance and broadmindedness without which there is no "democratic society." This means, amongst other things that every "formality", "condition", "restriction" or "penalty" imposed in this sphere must be proportionate to the legitimate aim pursued."[85] Therefore, restrictions on AI to cope with manipulatory practices should be considered in the light of possible infringement of the mentioned human rights.

On the other hand, technical restrictions may be used to control the manipulative content. To this end, different kind of technology solutions were created, including filtering of content, blocking of content, disabling and suspension of accounts, etc. To discover deep fakes, it is possible to use algorithms similar to those which generated the deep fake. For example, it was observed in 2018 that actors in AI-generated

[83] Marcelo Ienca, *supra* note 8, 840–841.
[84] William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *supra* note 28, 27.
[85] *Handyside v. United Kingdom*, European Court of Human Rights, para 49 (1976, Application no. 5493/72).

videos did not blink due to the lack of faces with closed eyes in most training datasets. Technologies looking for abnormal patterns of eyelid movement were thus developed. However, once this method became public, new deepfake videos were adapted by featuring blinking people[86]. Catching deepfakes with AI is something of a cat-and-mouse game. A detector algorithm can be trained to spot deepfakes, but then an algorithm that generates fakes can potentially be trained to evade detection[87]. Or, for example, to detect articles containing false information, it is possible with machine learning to train end-to-end models using labelled data, namely articles containing false information and articles containing accurate information. The system is then able to directly differentiate between these two types of articles[88]. Thus, the development of technology in order to detect deepfakes or misleading information online resembles the education of people and should also be brought forward together with the advancement of the manipulative AI.

Also, for US based platforms a domestic, however, very complex policy approach considered[89] requiring platforms to verify the identities of users behind accounts. This would create a consistent standard for social media platforms. Such a requirement, similar to the identity checks used in banking, could reduce the large-scale manipulation advantage of generative AI on social media. However, it might also stifle free speech, as the anonymity provided by social media can protect those who voice unpopular opinions, such as criticism of the government.

Although the article is limited to the analysis of the usage of AI as a tool for manipulation, it was already mentioned that together with the rapid development of technologies the capabilities of the AI are expanding, and the decision making powers of AI are gaining new forms. In this aspect, the technical restrictions on the AI might help limiting or postponing, if not escaping, the possibilities of the AI to manipulate people rather than to serve a tool for this.

### 6.4. Need for more research for the use of AI involving different stakeholders

While data protection, transparency, and accountability measures are very important directions addressing AI related concerns, it is necessary to investigate further the use of AI especially when it threatens human rights and personal autonomy. It is crucial to focus on identifying potential threats and developing effective response strategies to minimize negative impacts. These efforts should be inclusive and involve multiple stakeholders to ensure that all risks are considered and that solutions do not favour any particular group or silence undesired voices[90].

The need for surveillance over AI development comes not only from the governments or regional organisations, but civil societies and other organisations alone or in partnerships with businesses. For example, the Partnership on AI[91] is led by a group of AI researchers representing six of the world's largest technology companies: Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft. In 2017 the addition of six not-for-profit Board members expanded the Partnership into a multi-stakeholder organization—which now represents a community of more than 50 member organizations.

---

86 Noémi Bontridderp and Yves Poullet, *supra* note 9, 8.
87 WIRED, *Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them,* (2020) // https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/
88 Noémi Bontridderp and Yves Poullet, *supra* note 9, 8–9.
89 William Marcellino, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith, *supra* note 28, 26.
90 AccessNow, *supra* note 22, 35.
91 *Partnership on AI* // https://partnershiponai.org/

This organisation is aimed at better understand AI's impacts. The Partnership was established to study and formulate best practices on AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society. Another prominent example of research of AI and fighting it's malicious effect is the strong investigative journalism initiatives in the EU, e.g. the organisation WITNESS[92] aimed at creating tools and practices for detecting misinformation and promoting reliable media.

In summary, while the law is an important tool in regulating AI, due to different reasons (inflexibility, generalised nature, etc.) it is alone not a sufficient instrument to tackle the undesired usage of AI in manipulation that may cause harm to persons. A combination of ethical guidelines, public education, technological safeguards, and collaborative efforts is necessary to effectively address the challenges posed by AI-driven manipulation.

## CONCLUSIONS

1. Manipulation as a phenomenon has always been a part of human interaction, a way for some to influence others' choices by exploiting their vulnerabilities, often without the victim's awareness. Although this behavioural pattern is deeply embedded in the society, the rise of AI has significantly amplified these manipulation techniques and has brough significant threats to human rights and principles of democracy.

2. Although manipulation is a model of behaviour embedded in the society, there are different concepts and definitions of it reflecting the different elements of the phenomenon. While in most cases manipulation is undesired behaviour and there are initiatives to restrict it, in certain environments the society accepts and approves manipulation (e.g., entertainment, nudging, etc.). Therefore, in order to introduce restrictions on particular activities, especially the ones related to technological advancement, it is needed to assess in which particular cases restrictions are really necessary – i.e. to determine when the manipulation is posing threats to individuals and principles of democracy.

3. AI-driven manipulation pose threat to individuals and principles of democracy in several key areas, primarily in advertising, political campaigns, and the spread of hate speech. By leveraging insights into individual vulnerabilities, AI makes it easier to tailor personalized content that influences consumer behaviour, political opinions, and social dynamics. In advertising, businesses use AI to identify the specific weaknesses or desires of consumers, allowing them to craft highly targeted and individualized proposals. This can lead consumers to purchase goods and services they might not have otherwise considered. In field political campaigns, for example, voters are being approached by bots in a way that feels relatable and authentic, gradually introducing and reinforcing specific political views that the target audience begins to adopt as their own. In the business world, deepfake technologies are employed to defame competitors or discredit their products and services, creating an environment of distrust and damaging reputations.

4. AI-based manipulation negative consequences present a relatively new challenge to the legal landscape, one that current regulations struggle to fully address. One of the pioneering efforts in this area is the EU's AI Act, which aims to regulate AI-driven

[92] E. g. see Sam Gregory, *supra* note 59.

manipulation through two main strategies. First, it prohibits AI practices that "deploy subliminal techniques" and target and exploit vulnerable individuals where it causes "significant harm". Second, subject to certain conditions, the AI Act mandates transparency in the use of deepfake technologies, requiring clear disclosure when content has been artificially generated or altered, to prevent deception and protect the integrity of information. Thus, EU has chosen the strategy to limit the negative consequences of AI related manipulation in two ways – first, restrict the AI base manipulative activities resulting in significant harm, and second – requiring transparency, which in essence is removal of one of the essential element of covertness in manipulation. Other EU documents, such as General Data Protection Regulation or Digital Services Act supplement this legislative framework by setting the requirements related to transparency, purpose and disclosure.

5. Although law in is a fundamental instrument to protect human rights and principles of democracy, due to vast development of technologies and variety of new possibilities to generate manipulation using AI, law may not be the only means to limit the undesired practices of usage of AI. To this end additional resources are invoked – ethical principles, awareness raising, technological solutions detecting malpractices, partnerships between different sectors, etc. The variety of forms of cooperation can create a multi-layered defence that is more adaptable and responsive to the evolving threats posed by manipulative usage of AI than law alone.

6. From the legislative framework, first of all, the EU AI Act, it may be inferred that at EU level it was chosen to base the regulation of AI on the concept of limitations on the effect rather than limitations on the values. Indeed, this approach is more suitable to address the nuances and the variety of concepts of manipulation, as banning certain types of technological solutions (e.g. deepfakes) might restrict certain technological advancements which are accepted or even beneficial to individuals or the society (e.g. health *apps* using nudging, entertainment technologies, etc.).

7. However, the previously presented approach of limitations on the effect is applicable when AI is regarded as a tool for manipulation, and not an initiator of it. Having in mind the vast development of AI, the capabilities of AI to take high complexity decisions without human intervention may lead to a situation where the AI would start manipulating the humans. As this would involve a different ontological discussion, this threat should be investigated separately.

## BIBLIOGRAPHY

### Legal references

1. *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe* (COM/2018/237 final).
2. Council of Europe, *Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes*, Para. 9. Decl (13/02/2019)1.
3. Council of Europe. *Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes* (2019). Decl (13/02/2019)1.
4. *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in*

*the electronic communications sector (Directive on privacy and electronic communications)*. OJ L 201, 2002.

5. *Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ("Unfair Commercial Practices Directive")*. OJ L 149, 2005.
6. *European Commission White Paper on Artificial Intelligence – A European Approach to Excellence and Trust*. Brussels, 19.2.2020 COM (2020) 65 final.
7. *Handyside v. United Kingdom*. European Court of Human Rights (1976, Application no. 5493/72).
8. *Pretty v. United Kingdom*. European Court of Human Rights (2002, Application no. 2346/02).
9. *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.
10. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. OJ L 119, 2016.
11. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).* OJ L 277, 2022.
12. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. OJ L, 2024/1689.
13. *United Nations International Covenant on Economic, Social and Cultural Rights.* (1966) // https://www.ohchr.org/sites/default/files/cescr.pdf

**Other resources**

1. AccessNow. *Human Rights in the Age of Artificial Intelligence*. (2018) // https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdfArendt, Hannah. *Hannah Arendt: From an Interview* (Interview with Roger Errera in 1974). The New York Review of Books, 1978.
2. Bakiner, Onur. "The promises and challenges of addressing artificial intelligence with human rights." *Big Data & Society* (July–December 2023): 1–13 // DOI: 10.1177/20539517231205476
3. Bontridderp, Noémi and Yves Poullet. "The Role of Artificial Intelligence in Disinformation." *Data & Policy* 3 (2021): e32-2 // DOI: 10.1017/dap.2021.20, Botes, Marietjie. "Autonomy and the social dilemma of online manipulative behavior." *AI Ethics* 3 (2023): 315–323 // DOI: 1007/s43681-022-00157-5, Bublitz, Jan Christoph and Reinhard Merkel. "Crimes against minds: On mental manipulations, harms and a human right to mental self-determination." *Crim Law Philos* 8(1) (2014): 51–77.

4. Calo, Ryan. "Digital Market Manipulation." *The George Washington Law Review* 82(4) (2014): 995-1051 // https://digitalcommons.law.uw.edu/faculty-articles/25
5. Cambridge online dictionary. *Manipulation* // https://dictionary.cambridge.org/dictionary/english/manipulationCarroll, Micah, Alan Chan, Henry Ashton, and David Krueger. "Characterizing manipulation from AI systems": 1–13. In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (2023).
6. Cohen, Tegan. "Regulating Manipulative Artificial Intelligence." *SCRIPTed* Vol. 20, No. 1 (2003): 203–242 // DOI: 10.2966/scrip.200123.203, European Commission Independent High Level Expert Group on Artificial Intelligence. *Ethics guidelines for trustworthy AI.* (2019) // https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419European Commission. *Public Consultation on the AI White Paper. Final report*. (November 2020) // https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68462European Commission/Deloitte. *Opportunities and Challenges for the Use of Artificial Intelligence in Border Control, Migration and Security*. Vol. 2: Addendum (2018) // https://op.europa.eu/en/publication-detail/-/publication/69f33ff7-a156-11ea-9d2d-01aa75ed71a1/language-enEUROPOL. *Facing reality? Law enforcement and the challenge of deepfakes*. (2022) // https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdfFarahany, Nita A. "The Costs of Changing Our Minds." *69 Emory L. J. 75* (2019) // https://scholarlycommons.law.emory.edu/elj/vol69/iss1/2Faraoni, Stefano. "Persuasive technology and computational manipulation: Hypernudging out of mental self-determination." *Frontiers in Artificial Intelligence* 6 (2023) // DOI: 10. 3389/frai.2023.1216340Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. "The Rise of Social Bots." *Communications of the ACM* Vol. 59, No. 7 (July 2016): 96–104 // DOI 10.1145/2818717, Gregory, Sam and Eric French. "How do we work together to detect AI-manipulated media?" *Witness Media Lab* (2019) // https://lab.witness.org/projects/osint-digital-forensics/Gregory, Sam. "Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism." *Journalism* Vol. 0(0) (2021): 1–22 // DOI 10.1177/14648849211060644
7. Gregory, Sam. "Synthetic media forces us to understand how media gets made". *NiemanLab Predictions for Journalism* (2023) // https://www.niemanlab.org/2022/12/synthetic-media-forces-us-to-understand-how-media-gets-made/
8. Gregory, Sam. "Fortify the Truth: How to Defend Human Rights in an Age of Deepfakes and Generative AI." *Journal of Human Rights Practice* 15 (2023): 702–714 // DOI: 10.1093/jhuman/huad035, Guha, Ahona. "Understanding and Managing Psychological Manipulation." *Psychology Today* // https://www.psychologytoday.com/us/blog/prisons-and-pathos/202104/understanding-and-managing-psychological-manipulationIEEE Standards Association. *The IEEE Global Initiative 2.0 on Ethics of Autonomous and Intelligent Systems.* (2023) // https://standards.ieee.org/industry-connections/activities/ieee-global-initiative/Ienca, Marcello and Effy Vayena. "Digital Nudging Exploring the Ethical Boundaries." In: Veliz (ed.), *The Oxford Handbook of Digital Ethics.* Oxford University Press, Oxford, 2021.

9. Ienca, Marcelo. "On Artificial Intelligence and Manipulation." *Topoi* 42 (2023): 833–842 // DOI: 10.1007/s11245-023-09940-3
10. Kavoliūnaitė, Eglė. "Concepts of the beginning of legal protection of human life." *Teisės problemos* 2 (48) (2005): 5–26.
11. Klenk, Michael. "Ethics of Generative AI and Manipulation: a Design-Oriented Research Agenda." *Ethics and Information Technology* 26 (2024): 9 // DOI: 10.1007/s10676-024-09745-x, Langguth, Johannes, Konstantin Pogorelov, Stefan Brenner, Petra Filkuková and Daniel Thilo Schroeder. "Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes." *Front. Commun.* 6:632317 (2021): 1–12 // DOI: 10.3389/fcomm.2021.632317, Marcellino, William, Nathan Beauchamp-Mustafaga, Amanda Kerrigan, Lev Navarre Chao, Jackson Smith. *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0. Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI.* (2003) // https://www.rand.org/pubs/perspectives/PEA2679-1.html
12. Morozovaitė, Viktorija. "Hypernudging in the changing European regulatory landscape for digital markets." *Policy Internet* 15 (2022): 78–99 // DOI: 10.1002/poi3.329, Nepryakhin, Nikita. "Classification of vulnerability factors in the process of psychological manipulation." *Conference: International Conference on Advanced Research in Social Sciences* (2019) // DOI: 10.33422/icarss.2019.03.93, NIST. *AI Risk Management Framework.* (2024) // https://www.nist.gov/itl/ai-risk-management-frameworkOECD Working Papers on Public Governance. *Hello, World. Artificial intelligence and its use in the public sector.* (2019): 56 // https://www.oecd-ilibrary.org/docserver/726fd39d-en.pdf?expires=1666023983&id=id&accname=guest&checksum=2B0678AFDEB937C7A5B42575C7C96F44
13. OECD. *AI principles*. (2024) // https://www.oecd.org/en/topics/sub-issues/ai-principles.html
14. OECD. "Dark Commercial Patterns". *OECD Digital Economy Papers,* n. 366, DSTI/CP (2021)12/FINAL. Paris, 2022.
15. OHCHR. *Report on the Right to Privacy in the Digital Age*. UN doc A/HRC/39/29, Geneva, 2018.
16. *Partnership on AI* // https://partnershiponai.org/, PHD Media. *New beauty study reveals days, times and occasions when U.S. women feel least attractive.* (October 2012) // https://www.prnewswire.com/news-releases/new-beauty-study-reveals-days-times-and-occasions-when-us-women-feel-least-attractive-226131921.html, Poullet, Yves. *Ethique et droits de l'Homme dans notre société du numérique (Ethics and human rights in our digital society*). Brussels: Académie Royale de Belgique, 2020.
17. Stahl, Bernd Carsten, Doris Schroeder and Rowena Rodrigues. *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*. (2023) // DOI: 10.1007/978-3-031-17040-9, Susser, Daniel, Beate Roessler and Helen Nissenbaum. "Technology, Autonomy, and Manipulation." *Internet Policy Review* 8(2)-1 (2019): 1–22 // DOI: 10.14763/2019.2.1410
18. UNDESA. *Resource Guide on Artificial Intelligence (AI) Strategies*. New York City, NY: UNDESA, 2021.
19. UNESCO. *Report of COMEST on robotics ethics – UNESCO Digital Library*. 17 // https://unesdoc.unesco.org/ark:/48223/pf0000253952

20. UNGA. *Promotion and Protection of the Right to Freedom of Opinion and Expression: Note by the Secretary-General.* UN Doc A/73/348, New York City, NY: UNGA, 2018.
21. Wilkinson, T. Martin. "Nudging and Manipulation." *Political Studies* 61, 2 (2013): 341–355.
22. WIRED. *Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them*. (2020) // https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/Yadav, Vikrant Sopan. "AI and Human Rights: A Critical Ethico-Legal Overview." *Agathos; Iasi,* Vol. 14, Issue. 1 (2023): 261–270.
23. Zuboff, Shoshana. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30(1) (2015): 75–89 // DOI: 10.1057/jit.2015.5