Manipulation and the AI Act

18th January 2022

CONTACT

Risto Uuk
The Future of Life Institute
risto@futureoflife.org

Manipulation and the AI Act

- Manipulation through AI poses risks for individuals and communities. From worsening body image issues to influencing election outcomes and targeting consumption choices by exploiting vulnerabilities.
- Manipulation is problematic because it can cause direct harm, violate the target's autonomy and treat persons as things not ends in themselves.
- We recommend two main changes in Article 5 in the EU AI Act: removing subliminal techniques and adding societal harm to the list of harms of manipulation.

Digital technologies, in general, and artificial intelligence (AI), in particular, can **intensify the risks of human manipulation** and bring about **unique issues**. First, there is a significant **level of opacity** with AI due to the lack of transparency and explainability of the vast majority of algorithms, coupled with consumers lacking technical literacy on AI's shortcomings.[1] Second, AI systems can detect or infer a person's preferences, interests, habits, and many other characteristics to **personalise content** in a precise manner.[2] Furthermore, AI systems can be used to assess people's psychological states such as mood, stress, and emotions.[3] Third, AI enables the ability to weaken the deliberative autonomy of consumers by **exploiting their decision-making vulnerabilities**.[4]

Although 'manipulation' is a difficult concept to define, the literature describes four main characteristics.[5] One is that it is a **non-rational influence** where the manipulator tries to bypass or weaken a person's deliberative decision-making capacities. Another is that manipulation requires the **use of trickery and deception**, often through hidden means, to get someone to behave in a certain way. The third is that it entails **using some degree of pressure to do as the manipulator wants**, for example, through emotional blackmailing. Lastly, it is generally **not guided by the target's own interests, goals and preferences**, but only the manipulator's.

TABLE: CHARACTERISTICS OF MANIPULATION

Characteristic	**Example**
Non-rational influence	Platform starts another video automatically
Use of trickery and deception	Website promises that the subscription is free but then automatically renews it
Some degree of pressure	Social media site tells users that many of their friends are already liking something
Often hidden influence	Company collects user data from personality quizzes for political influence
Exploiting vulnerabilities	Advertiser targets an ad for the moment when the user is sad
Not guided by the target's own interests	Gig economy drivers are urged to keep working after the end of their shift

1 Federico Galli, "AI and Consumers Manipulation: What the Role of EU Fair Marketing Law?," Católica Law Review 4, no. 2 (May 1, 2020): 35–64, https://doi.org/10.34632/catolicalawreview.2020.9320.

2 Daniel Susser, Beate Roessler, and Helen Nissenbaum, "Technology, Autonomy, and Manipulation," Internet Policy Review 8, no. 2 (June 30, 2019), https://policyreview.info/articles/analysis/technology-autonomy-and-manipulation.

3 Sandra C Matz, Ruth E Appel, and Michal Kosinski, "Privacy in the Age of Psychological Targeting," Current Opinion in Psychology, Privacy and Disclosure, Online and in Social Interactions, 31 (February 1, 2020): 116–21, https://doi.org/10.1016/j.copsyc.2019.08.010.

4 Galli, "AI and Consumers Manipulation."

5 Robert Noggle, "The Ethics of Manipulation," in The Stanford Encyclopedia of Philosophy, ed. Edward N. Zalta, Summer 2020 (Metaphysics Research Lab, Stanford University, 2020), https://plato.stanford.edu/archives/sum2020/entries/ethics-manipulation/.

Risks from AI Manipulation

Risks from AI systems are not hypothetical, they **already threaten individuals and communities** and can lead to further harms if not adequately prepared for. Recent evidence indicates that Instagram's machine learning algorithm is harmful to a significant proportion of users and especially teenage girls **worsening body image issues and increasing anxiety and depression**, but the platform does not want to reduce engagement to preserve its profit margins.[6] Furthermore, the voter micro-targeting firm Cambridge Analytica has used advanced analytics to target ads at specific voters with data gathered from the social media site Facebook to **potentially swing elections around the world**.[7] In addition, there have been leaked strategy documents indicating how advertisers plan to **target people with their ads in moments of vulnerability**.[8] In some instances, AI is already influencing consumer decisions more than standard marketing techniques.[9]

There are several ways manipulation can be unethical. Firstly, manipulators can cause **direct harm** to the people targeted, as shown above with some of the examples from social media, politics and marketing. Secondly, manipulators can **violate the target's personal autonomy** as they intend to reduce personal choice and decision-making. Finally, manipulators may **treat persons as things** to take advantage of rather than to discuss and reason with. The examples of manipulatory risks mentioned above are obviously unethical to most people, but manipulation can take a more subtle role in the case of nudging.

Nudges are behavioral practices that are developed with insights from behavioral experiments. Some nudges are clearly not manipulative because they are not hidden or implemented to exploit someone. For example, **informational nudges like nutrition labels showing calories and ingredients are not manipulative**. On the other hand, **many nudges used online like autoplay on YouTube are more controversial** as they take advantage of user vulnerabilities to fall into the trap of mindless engagement and overuse of the platform.

The EU AI Act on Manipulation

The EU AI Act directs its attention to manipulation in two main ways:

1) Identifying the practice, target population and harms: "covers practices that have a significant potential to manipulate persons through **subliminal techniques** beyond their consciousness or exploit vulnerabilities of **specific vulnerable groups such as children or persons with disabilities** in order to materially distort their behaviour in a manner that is likely to cause them or another person **psychological or physical harm**."[10]

2) Acknowledging that other regulation might cover manipulation: "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."[11]

---

6 Georgia Wells Seetharaman Jeff Horwitz and Deepa, "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," Wall Street Journal, September 14, 2021, sec. Tech, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739.

7 Marcello Ienca and Effy Vayena, "Cambridge Analytica and Online Manipulation," Scientific American, March 30, 2018, https://blogs.scientificamerican.com/observations/cambridge-analytica-and-online-manipulation/.

8 Nitasha Tiku, "Welcome to the Next Phase of the Facebook Backlash," Wired, May 21, 2017, https://www.wired.com/2017/05/welcome-next-phase-facebook-backlash/.

9 Jane Wakefield, "How Artificial Intelligence May Be Making You Buy Things," BBC News, November 9, 2020, https://www.bbc.com/news/technology-54522442.

10-11 European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206 final), pg. 12-13

We focus on paragraphs (1) and (2) of Article 5 in the AI Act because these paragraphs provide the most general principles for prohibiting AI manipulation. The rest of Article 5 focuses on specific applications such as social scoring and biometrics. A lot has already been written about these elsewhere.[12]

Many members of the civil society have expressed their thoughts on how the AI Act protects against manipulatory risks. The European Consumer Organisation (BEUC) states that AI that manipulates humans in a way that causes **economic or societal harm** is not covered by the proposal, only AI that causes physical or psychological harm through manipulation is.[13] This is also emphasised by European Digital Rights (EDRi) who state that many uses of AI will not target individuals, but rather groups.[14] Other types of harms such as **cultural and harms to democracy** have been mentioned by other civil society organisations. EDRi and Amnesty International also add that the specific vulnerabilities listed is very limited – only age and physical or mental disability are covered – whereas there are **many other protected characteristics in the EU law**.[15] In addition, the Electronic Privacy Information Center and many others say that the requirement of using '**subliminal techniques**' for manipulation to be prohibited is vague and should be removed to instead include any technique used for manipulation.[16] Finally, BEUC advocates for the removal of the wording **"in order to"** because it requires proving intent, whereas AI could cause manipulatory harm without intention.

Policy Recommendations

1. **We suggest removing "subliminal techniques" from the proposal so that the article applies to any type of manipulation technique.**

Many global industry leaders such as Facebook and Google have brought attention to the problem of lack of clarity in the AI Act regarding what counts as a 'subliminal technique'.[17] There is a broad consensus that this is a problematic aspect of Article 5 even though specific solutions may vary. We, as a member of civil society, suggest **removing the requirement for subliminal techniques and apply the article to any manipulative technique**.

The term 'subliminal' is not explicitly defined in the AI Act proposal. It usually refers to sensory stimuli that consumers **cannot consciously perceive**.[18] A stimuli would only be considered subliminal if it was presented for less than 50 milliseconds.[19] **Most uses of AI will not be subliminal** since they will be consciously perceived by users.[20] Therefore, this article **will still allow many forms of manipulation**.

For example, suppose that an AI system concludes that an individual is thirsty due to sensors that can identify the last

12 Please find some examples here:
- https://www.europarl.europa.eu/RegData/etudes/STUD/2021/696968/IPOL_STU(2021)696968_EN.pdf
- https://www.accessnow.org/cms/assets/uploads/2021/06/BanBS-Statement-English.pdf
- https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net.

13 "Feedback from: BEUC - The European Consumer Organisation," European Commission Public Consultation, August 5, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665432_en.

14 "Feedback from: European Digital Rights (EDRi)," European Commission Public Consultation, August 3, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665234_en.

15 "Feedback from: Amnesty International," European Commission Public Consultation, August 6, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665634_en.

16 "Feedback from: The Electronic Privacy Information Center (EPIC)," European Commission Public Consultation, August 6, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665484_en.

17 "Feedback from: Google," European Commission Public Consultation, July 15, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en; "Feedback from: Facebook Ireland Limited," European Commission Public Consultation, August 6, 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665607_en.

18 Nathalie A. Smuha et al., "How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, August 5, 2021), https://doi.org/10.2139/ssrn.3899991.

19 Mihai Radu Ionescu, "Subliminal Perception of Complex Visual Stimuli," Romanian Journal of Ophthalmology 60, no. 4 (2016): 226–30.

20 Smuha et al., "How the EU Can Achieve Legally Trustworthy AI."

time they filled their water bottle. If this system shows this person a commercial for Coca-Cola when they become thirsty, this would not be subliminal as long as they can consciously perceive the commercial. Even if they are not aware that they are being influenced to buy the drink because of the algorithm, or that the algorithm targeted them in a state of vulnerability, the manipulation is not subliminal because **the stimulus was not below the threshold of detection**.

2. We recommend adding societal harm to the list of harms currently included in Article 5.

We recommend **adding societal harm to the list of harms of manipulation** currently included in Article 5. Many civil society organisations, academics, and European expert groups like the AI High-Level Expert Group have discussed the importance of increasing societal wellbeing or reducing societal harm to ensure that European citizens can trust AI systems.[21] Some negative consequences of implementing certain AI systems, such as **harming the democratic process, eroding the rule of law, or exacerbating inequality**, can cause only modest harms to individuals but hurt societies at large.[22] AI systems have already been used to manipulate individuals to influence political opinions as is well-known due to the Cambridge Analytica scandal. This will only intensify in the future unless dealt with now.

We acknowledge **the importance of considering the costs of additional regulations to European businesses**. Regardless of this burden, many companies agree with the European Union's approach to ensure citizen trust. Healthcare company Healx, for example, has mentioned that if AI is not properly regulated, it can perpetuate structural racism, may not accurately represent women and minorities, and damage trust.[23] Furthermore, the European Commission has taken steps to reduce burdens by choosing a **risk-based approach to regulating AI**, which means that **a small proportion of AI systems will be considered prohibited or high risk** and require extra scrutiny. In an analysis of the cost of the AI Act, it was claimed that the European Commission expects the proportion of **high-risk AI systems to be between 5%-15% of all AI systems**.[24] For most other AI systems, there will only be transparency obligations or possible voluntary codes of conduct.

The societal harm consideration in the context of manipulation will be applicable to an even smaller proportion of AI systems and hence, will not lead to overly costly regulation. However, those limited number of systems could still do a lot of damage if not properly assessed. It is crucial to do so because as of now, existing regulations limit themselves to preventing individual harm.[25] Two ways to make the enforcement of the societal harm requirement more feasible and proportionate are to focus on (1) **protecting democracy, rule of law and equality** as the values of Article 2 in the Treaty of the European Union or (2) **using the definition of a systemic risk** as it is done in the Digital Services Act narrowly in the context of very large online platforms. These approaches are more feasible and proportionate because they address risks the EU has already experienced (e.g. political manipulation) and they are smaller and clearer categories of societal harm (e.g. protecting democracy).

21 Some supporters can be found here:
- "Feedback from: BEUC - The European Consumer Organisation";
- Nathalie A. Smuha, "Beyond the Individual: Governing AI's Societal Harm," Internet Policy Review 10, no. 3 (September 30, 2021), https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm;
- "Ethics Guidelines for Trustworthy AI," High-Level Expert Group on Artificial Intelligence (European Commission, April 8, 2019), https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

22 Smuha, "Beyond the Individual."

23 Tim Guilliams, "Europe's AI Laws Will Cost Companies a Small Fortune – but the Payoff Is Trust," VentureBeat (blog), November 21, 2021, https://venturebeat.com/2021/11/21/europes-ai-laws-will-cost-companies-a-small-fortune-but-the-payoff-is-trust/.

24 Moritz Laurer, Andrea Renda, and Timothy Yeung, "Clarifying the Costs for the EU's AI Act," CEPS (blog), September 24, 2021, https://www.ceps.eu/clarifying-the-costs-for-the-eus-ai-act/.

25 Smuha, "Beyond the Individual."

About the author

Risto Uuk is a Policy Researcher at the Future of Life Institute and is focused primarily on researching policy-making on AI to maximize the societal benefits of increasingly powerful AI systems. Previously, Risto worked for the World Economic Forum on a project about positive AI economic futures, did research for the European Commission on trustworthy AI, and provided research support at Berkeley Existential Risk Initiative on European AI policy. He completed a master's degree in Philosophy and Public Policy at the London School of Economics and Political Science. He has a bachelor's degree from Tallinn University in Estonia.

CONTACT

Risto Uuk
The Future of Life Institute
risto@futureoflife.org