Missing Mechanisms of Manipulation in the EU AI Act

[1]Matija Franklin, [1]Hal Ashton, [2]Rebecca Gorman, [2]Stuart Armstrong

[1]University College London, London, UK [2]Aligned AI, Oxford, UK
matija.franklin@ucl.ac.uk

Abstract

The European Union Artificial Intelligence (AI) Act proposes to ban AI systems that "manipulate persons through subliminal techniques or exploit the fragility of vulnerable individuals, and could potentially harm the manipulated individual or third person". This article takes the perspective of cognitive psychology to analyze and understand what algorithmic manipulation consists of, who vulnerable individuals may be, and what is considered as harm. Subliminal techniques are expanded with concepts from behavioral science and the study of preference change. Individual psychometric differences which can be exploited are used to expand the concept of vulnerable individuals. The concept of harm is explored beyond physical and psychological harm to consider harm to one's time and right to an un-manipulated opinion. The paper offers policy recommendations that extend from the paper's analyses.

Introduction

The European Union (EU) Artificial Intelligence (AI) Act proposes a banning of AI systems that "manipulate persons through subliminal techniques or exploit the fragility of vulnerable individuals, and could potentially harm the manipulated individual or third person" (EU Commission 2021). The EU AI ACT proposes two practices for regulating manipulation (Veale and Borgesius 2021); namely, prohibiting:

1. AI systems that use subliminal techniques that a person is not consciously aware of to "materially distort" a person's behavior in a way that either causes or is likely to cause that person (or a third party) psychological or physical harm;

2. AI systems that exploit a specific group's vulnerabilities due to their age, physical or mental disability, to "materially distort" the behavior of an individual from that group in a way that either causes or is likely to cause that person (or a third party) psychological or physical harm.

The EU AI Act in its current form poses many questions for AI developers. This paper approaches the problem from the perspective of Cognitive Psychology to understand, analyze and expand on what subliminal techniques might constitute, who vulnerable individuals may be, and what is considered a significant harm.

Existing laws concerning human experimentation

AI systems are often developed through experimental interaction with human beings with the objective of modifying human behavior, whether to increase engagement, click-through, conversion, time-on-site, return visits, or some other measure (Tamburrelli and Margara 2014).

Developers of AI systems designed to study and modify human behavior tend not to use an IRB boards or equivalent when testing out the psychological and behavioral outcomes of their AI systems.[1] Even when they do, the academic community denies the requirement of an IRB board when AI systems are deployed by a commercial entity. A famous example comes from an experiment on Facebook showing that emotional states transfer from person to person in a social network - a finding known as *emotional contagion* (Kramer, Guillory, and Hancock 2014). In an editorial expression of concern by *PNAS*, which published the original article, the journal stated that as Facebook was a private company and it "was under no obligation to conform to the provisions of the Common Rule when it collected the data used by the authors..." (Verma 2014). We argue that EU regulations should define how these standards apply to commercial entities.

The EU AI Act arguably codifies for many forms of non-consensual human experimentation, by implicitly classifying AI systems which materially modify user behavior without voluntary informed consent as low-risk systems. Article 7, in the United Nations' "International Covenant on Civil and Political Rights" (1966) reads "...no one shall be subjected without his free consent to medical or scientific experimentation" (United Nations General Assembly 1966). The covenant has 173 state parties, including EU member states. Currently in the "Charter of Fundamental Rights of the European Union", Article 3 proposes that "...the following must be respected in particular: the free and informed consent of the person concerned, according to the procedures laid down by law..." (European Union 2010). Given the impact experimentation by AI systems can have on human subjects, regulators should incorporate these principles.

[1]Institutional Review Boards (IRBs) enforce international laws for consensual human experimentation at universities.

Subliminal techniques and beyond

It has been claimed that *Subliminal stimuli* can influence behavior. The Psychological research community has not drawn a firm consensus about the efficacy of subliminal techniques except perhaps that they are weaker than feared. Brooks et al. (2012) find that subliminal stimuli - sensory stimuli which are bellow the threshold for conscious perception - can only trigger actions that an individual intends to do either way. A meta analysis of the effectiveness of subliminal stimuli found that it had a low effect size which was not statistically significant (Trappey and Woodside 2004).

Criticisms of the EU AI Act suggest that subliminal techniques should be replaced with a broader range of manipulation techniques (Uuk 2022). In agreement with this we aim to provide an overview of these manipulation techniques.

People are aware of many aspects of their environment, but not necessarily aware of how these environmental aspects exert influence over their thoughts and behavior. Such influences are not perceptually subliminal, yet still subliminal in their influence. Aspects of an environment that can be changed to impact a person's behavior are known as *Choice Architecture* (Thaler and Sunstein 2021). The act of changing choice architecture to influence people's behavior without limiting or forcing options, or significantly changing their economic incentives is called *nudging* (Thaler and Sunstein 2021). All environments will to some extent influence behavior (Sunstein 2016). However, not all aspects of choice architecture are equally influential (Mertens et al. 2022). This raises questions about what aspects of choice architecture are manipulative; and thus ban-worthy.

Franklin et al. (2022) review guidance from behavioral scientists concerning the ethics of behavior change. First, Sunstein (2021) argue that people have a right not to be manipulated". Behavior change practices are manipulative when they do not engage with people's capacity for reflective choice. Second, behavioral scientists argue that it is acceptable to make welfare-promoting behaviors easier to do, but not acceptable to use *sludge* – behavior change that results in outcomes that would be in the practitioner's best interest rather than that of the target individual (Thaler 2018). Two common forms of sludge either promote self-defeating behaviors or discourage a person's best interest. Finally, proponents of *Libertarian Paternalism* argue that behavior change which avoids material incentives and coercion is more ethical. Behavior change should be used to give people guidance given their own goals, rather than changing their goals (Thaler and Sunstein 2021).

Changes in behavior over time they can compound into more fundamental changes to a person's life, which can be considered as manipulative. These manipulative behavioral changes are fortified by negative externalities. As no behavior sits in a vacuum, the occurrence of one behavior can create *behavioral Spillovers* (Dolan and Galizzi 2015). Behaviors that occur sequentially are often causally linked. The first behavior can cause a subsequent behavior, thus promoting a spillover. Spillovers can be both positive or negative, occurring in the same or opposite direction, respectively.

Another externality of manipulative AI influence is preference change. Evidence suggests that a person's behavioral history influences their preference (Ariely and Norton 2008). Preference influences behavior, but behavior often predates and leads to the emergence of new preference. Recommender systems often use Machine Learning (ML) to learn the preferences of users to optimise the delivery of some service. When an iterative ML approach is applied to recommender systems, it becomes increasingly difficult to identify whether the system is learning about its users' preferences or whether the recommender system has nudged its users to behave in a certain way in order to maximise its objective function (often metrics like user-attention or click-through) (Ashton and Franklin 2022). Typically more popular items are recommended more, making them even more popular (Mansoury et al. 2020). The recommender system's algorithms uses this behavioral data to train. Given that the behavioral data comes from behavior influenced by the recommender system, training on that data creates a feedback loop (Chaney, Stewart, and Engelhardt 2018). Such systems may over time change preferences to a increasingly narrow band of content. To counter this, Franklin et al. (2022) propose a interdisciplinary effort to understand mechanisms of preference change - *Preference Science*.

Psychometric differences as vulnerabilities

The EU AI act proposes the prohibition of AI systems that exploit a specific group's vulnerabilities due to their age, or physical or mental disability. We argue that the increased predictive power of online AI systems, when contained with extensive data about the online subject (user), their behavior, and their preferences, create an environment in which our natural psychometric differences can become material vulnerabilities. [2] Thus, from a cognitive perspective, to be vulnerable is to deviate from other people on a psychometric trait, so that this deviation can be exploited.

Psychometric differences can be used to create a simple predictive models of how certain groups of people will respond to a particular stimulus, and thus these predictable individual differences can be exploited. More recently, people's individual susceptibility towards the influence of different choice architectures has been labelled as *nudgeability* (de Ridder, Kroese, and van Gestel 2021). On a group level, exploiting small differences in a particular psychological construct can be materially effective. This is relevant in light of proxy measures for people's psychometric profiles, which use secondary data available online (Stark 2018).

People's individual differences in Big Five personality traits - openness to experience, conscientiousness, extraversion, agreeableness, neuroticism - have been successfully measured using "digital records of human behavior" (Kosinski, Stillwell, and Graepel 2013), which can be used to predict behavior (Kosinski et al. 2016), and thus are effective for "the adaptation of persuasive appeals to the psychological characteristics of large groups of individuals with the goal of influencing their behavior" (Matz et al. 2017). Digital footprints used for measuring personality include likes, social media posts, mobile device logs, browsing

[2] The sub-field concerned with designing measures of psychological constructs is called *Psychometrics*

logs and music collections (Lambiotte and Kosinski 2014). These algorithmic judgments of people's personalities have a higher accuracy than those made by their close acquaintances (Youyou, Kosinski, and Stillwell 2015).

Personality traits have been exploited as a vulnerability in influencing voting behavior (Gerber et al. 2011). Infamously in the "Facebook–Cambridge Analytica scandal" around 87 million Facebook users got their personal data collected (ur Rehman 2019). This data was used to analyse people's personality traits in order to make psychometrically-targeted political advertising.

Expanding the scope of harm

The EU AI Act has focused on regulating manipulation that causes or is likely to cause a person (or a third party) psychological or physical harm. Recent criticisms of the act have proposed that the acts should add societal harm to the list of harms, which would include things such as AI systems "harming the democratic process, eroding the rule of law, or exacerbating inequality" (Uuk 2022). We further propose that the AI Act should further include an additional two forms of harm - harm towards time and autonomy.

First, we argue that manipulation can result in a harm towards people's time, in that it steers people away from how they would use their time if they were not manipulated. This is in line with Cass Sunstein's argument for "manipulation as theft", in that self-interested manipulators can be "thieves – of money, emotions, time and attention, or something else" (Sunstein 2021). Examples of "theft" include recommendation systems predicting the kinds of content that will keep users on a platform (Zakon 2020), sometimes resulting in social media addiction (Hou et al. 2019). The consequences of social media addiction can also be seen as a form of theft, with a study finding that social media addiction was negatively related to job performance and positively related to job burnout (Zivnuska et al. 2019).

Second, AI manipulation can harm a person's autonomy. People's behavior is getting changed everyday, which compounds into fundamental changes to their life. People should have "a right not to be manipulated" (Sunstein 2021).

Policy recommendations

Given the present analyses, we propose the following policy recommendations for the EU AI Act.

Recommendation 1: In acknowledgement that many technological techniques are not classed as 'subliminal' that materially distort a person's behavior without their conscious awareness, Article 5 Section 1 (a) be modified from its current text to read:

"the placing on the market, putting into service or use of an AI system that materially distorts a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm"

Recommendation 2: In acknowledgement that all human beings, whether targeted in groups or as individuals, are vulnerable to substantial harm when their unique weaknesses are known and exploited, article 5 Section 1 (b) be modified from its current text to read:

"the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a person or group of persons in order to materially distort the behavior of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;"

Recommendation 3: In acknowledgement that any human experimentation by any entity should not be conducted under any less stringent standard than that established by Standard 1 of the Nuremburg Code, an item be added to Chapter 1 Article 5 Section 1 reading:

"the placing on the market, putting into service or use of an AI system that materially alters a person's behavior without their voluntary informed consent"

Recommendation 4: Add to the list of high risk AI systems in Annex III:

"9. AI systems that materially alter a person's behavior with their voluntary informed consent"

Recommendation 5: We propose a requirement added to Chapter 2, Article 13 for *sludge audits* before deployment and after monitoring, identifying the presence of mechanisms that change behavior and preference, and identifying and removing those that fit the criterion of sludge (Sunstein 2020). [3]

Recommendation 6: We propose the additions to the list of harms in Article 5 of: *"harm to one's time"* and *"harm to one's autonomy"*.

Conclusion

The European Commission's EU AI Act has set up some initial regulation on AI manipulation. In this paper, we have outlined the ways in which the EU AI Act's proposed subliminal techniques, vulnerable individuals, and harms can be expanded in light of research in cognitive psychology, behavioral science and the study of preference change. The aim of the article was to contribute to the discussion around ways in which the EU AI Act can be expanded to ensure that individuals and society are not harmed by AI manipulation.

Conflict of Interest Statement Rebecca Gorman and Stuart Armstrong work for Aligned AI, an AI safety company. Aligned AI is funding Matija Franklin's participation in this conference. Henry Ashton declares no conflict of interest.

References

Ariely, D.; and Norton, M. I. 2008. How actions create – not just reveal – preferences. *Trends in Cognitive Sciences*, 12(1): 13–16.

Ashton, H.; and Franklin, M. 2022. The problem of behaviour and preference manipulation in AI systems. In *The AAAI-22 Workshop on Artificial Intelligence Safety (SafeAI 2022)*.

Brooks, S.; Savov, V.; Allzén, E.; Benedict, C.; Fredriksson, R.; and Schiöth, H. 2012. Exposure to subliminal arousing stimuli induces robust activation in the amygdala, hip-

[3]The "Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government" aims to reduce "time taxes" (Martorana et al. 2021).

pocampus, anterior cingulate, insular cortex and primary visual cortex: A systematic meta-analysis of fMRI studies. *NeuroImage*, 59(3): 2962–2973.

Chaney, A. J. B.; Stewart, B. M.; and Engelhardt, B. E. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. *Proceedings of the 12th ACM Conference on Recommender Systems*, 224–232.

de Ridder, D.; Kroese, F.; and van Gestel, L. 2021. Nudgeability: Mapping conditions of susceptibility to nudge influence. *Perspectives on Psychological Science*, 1745691621995183.

Dolan, P.; and Galizzi, M. M. 2015. Like ripples on a pond: behavioral spillovers and their implications for research and policy. *Journal of Economic Psychology*, 47: 1–16.

EU Commission, T. 2021. 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.

European Union. 2010. *Charter of Fundamental Rights of the European Union*, volume 53. Brussels: European Union.

Franklin, M.; Ashton, H.; Gorman, R.; and Armstrong, S. 2022. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. *AAAI-22 Workshop on AI For Behavior Change*.

Gerber, A. S.; Huber, G. A.; Doherty, D.; and Dowling, C. M. 2011. The big five personality traits in the political arena. *Annual Review of Political Science*, 14: 265–287.

Hou, Y.; Xiong, D.; Jiang, T.; Song, L.; and Wang, Q. 2019. Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 13(1).

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15): 5802–5805.

Kosinski, M.; Wang, Y.; Lakkaraju, H.; and Leskovec, J. 2016. Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, 21(4): 493.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24): 8788–8790.

Lambiotte, R.; and Kosinski, M. 2014. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12): 1934–1939.

Mansoury, M.; Abdollahpouri, H.; Pechenizkiy, M.; Mobasher, B.; and Burke, R. 2020. Feedback Loop and Bias Amplification in Recommender Systems. *arXiv:2007.13019 [cs]*.

Martorana, C.; et al. 2021. Using Technology to Improve Customer Experience and Service Delivery for the American People. *https://www. whitehouse. gov/omb/briefing-room/2021/12/13/using-technology-to-improve-customer-experience-and-service-delivery-for-the-american-people/*.

Matz, S. C.; Kosinski, M.; Nave, G.; and Stillwell, D. J. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48): 12714–12719.

Mertens, S.; Herberz, M.; Hahnel, U. J.; and Brosch, T. 2022. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1).

Stark, L. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2): 204–231.

Sunstein, C. R. 2016. *The ethics of influence: Government in the age of behavioral science*. Cambridge University Press.

Sunstein, C. R. 2020. Sludge audits. *Behavioural Public Policy*, 1–20.

Sunstein, C. R. 2021. Manipulation As Theft. *Available at SSRN 3880048*.

Tamburrelli, G.; and Margara, A. 2014. Towards automated A/B testing. In *International Symposium on Search Based Software Engineering*, 184–198. Springer.

Thaler, R. H. 2018. Nudge, not sludge.

Thaler, R. H.; and Sunstein, C. R. 2021. *Nudge: The final edition*. Penguin.

Trappey, R. J.; and Woodside, A. 2004. *Brand choice: revealing customers' unconscious-automatic and strategic thinking processes*. Springer.

United Nations General Assembly. 1966. International Covenant on Civil and Political Rights. *Treaty Series*, 999: 171.

ur Rehman, I. 2019. Facebook-Cambridge Analytica data harvesting: What you need to know. *Library Philosophy and Practice*, 1–11.

Uuk, R. 2022. UManipulation and the AI Act. *The Future of Life Institute*.

Veale, M.; and Borgesius, F. Z. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97–112.

Verma, I. M. 2014. Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 201412469.

Youyou, W.; Kosinski, M.; and Stillwell, D. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4): 1036–1040.

Zakon, A. 2020. Optimized for addiction: Extending product liability concepts to defectively designed social media algorithms and overcoming the communications decency act. *Wis. L. Rev.*, 1107.

Zivnuska, S.; Carlson, J. R.; Carlson, D. S.; Harris, R. B.; and Harris, K. J. 2019. Social media addiction and social media reactions: The implications for job performance. *The Journal of social psychology*, 159(6): 746–760.