scripted

Volume 20, Issue 1, February 2023

Regulating Manipulative Artificial Intelligence

*Tegan Cohen**

DOI: 10.2966/scrip.200123.203

Abstract

AI scientists are rapidly developing new approaches to understanding and exploiting vulnerabilities in human decision-making. As governments around the world grapple with the threat posed by manipulative AI systems, the European Commission (EC) has taken a significant step by proposing a new *sui generis* legal regime (the AI Act) which prohibits certain systems with the 'significant' potential to manipulate. Specifically, the EC has proposed prohibitions on AI systems which deploy subliminal techniques and exploit vulnerabilities in specific groups. This article analyses the EC's proposal, finding that the approach is not tailored to address the capabilities of manipulative AI. The concepts of subliminal techniques, group-level vulnerability, and transparency, which are core to the EC's proposed response, are inadequate to meet the threat arising from growing capabilities to render individuals susceptible to hidden influence by surfacing and exploiting vulnerabilities in individual decision-making processes. In seeking to secure the benefits of AI while meeting the heightened threat of manipulation, lawmakers must adopt new frameworks better suited to addressing new capabilities for manipulation assisted by advancements in machine learning.

Keywords
Artificial intelligence; manipulation; AI Act; regulation; subliminal techniques; vulnerability; transparency

* Lecturer, Law School, Queensland University of Technology, Brisbane, Australia, Tegan.cohen@qut.edu.au. The author would like to thank Dr Henry Fraser, Associate Professor Mark Burdon, Dr Ariadna Matamoros-Fernández and Dr Kylie Pappalardo, as well as the two anonymous reviewers, for their time and insightful comments on an earlier draft of this article.

1 Introduction

AI scientists are rapidly developing new approaches to understanding and exploiting vulnerabilities in human decision-making.[1] Outside the lab, millions of people interact daily with complex machine learning systems designed to 'learn' their behavioural patterns and adapt stimuli (such as newsfeeds and ads) to induce choices which align with the systems' objectives. As governments around the world grapple with the quandary posed by manipulative AI, the European Commission (EC) has taken a significant step toward imposing legal restrictions, proposing a new *sui generis* legal regime for AI systems.[2] The draft AI Act prohibits two kinds of AI systems which the EC considers 'have a significant potential to manipulate persons': broadly, systems that deploy subliminal techniques and systems that exploit vulnerabilities of specific groups due to age, physical or mental disability.[3]

This article examines the approach to manipulative AI systems proposed in the AI Act. Drawing on the definition of manipulation developed by Susser, Roessler and Nissenbaum,[4] it argues that the EU approach is ill-designed to meet the specific threat imposed by manipulative AI. Although the AI Act correctly targets two defining features of manipulation, hidden influence and the exploitation of vulnerabilities, the model is deficient in certain fundamental

---

1 See, eg: Amir Dezfouli, Richard Nock and Peter Dayan, 'Adversarial Vulnerabilities of Human Decision-Making' (2020) 117 *Proceedings of the National Academy of Sciences* 29221; Jon Whittle, 'AI Can Now Learn to Manipulate Human Behaviour' (*The Conversation*, 11 February 2021), <https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031> accessed 26 July 2022.

2 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 (hereinafter 'AI Act').

3 Explanatory Memorandum to the 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, 12.

4 Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World' (2019) 4(1) *Georgetown Law Technology Review* 1.

respects. First, the AI Act targets hidden content, namely subliminal techniques[5] – a highly ambiguous category of practices involving non-perceptible stimuli. While clearly a form of hidden influence, the power of manipulative AI systems lies not in sharpened capabilities to tailor and deliver subliminal stimuli, but in growing capabilities to surface and exploit individual vulnerabilities in perception, attention and other factors which affect decision-making. Second, while the AI Act also contains a prohibition on AI systems which exploit vulnerabilities, it reflects a solely group-level model of vulnerability,[6] deeming some groups as susceptible to manipulation based on characteristics such as age, and the rest of the population as not.[7] As such, both proposed prohibitions only partially engage with the specific problem posed by manipulative AI – growing capabilities to render individuals susceptible to hidden influence by surfacing and exploiting weaknesses in individual decision-making processes.

If adopted, the draft regulation will be a pioneering attempt to shape the future direction of AI through law. The approach taken in the AI Act, including with respect to manipulative AI, will likely become a lodestar for legal developments far beyond the borders of the European Union. An effective legal response must be appropriately adapted to the specific threat of covert and tailored attempts to exploit vulnerabilities in decision-making processes using machine learning techniques.

This article proceeds in six sections. Section II situates the proposed bans in the broader context of the draft AI Act, outlining key aspects of the proposed regulation relevant to the problem of manipulative AI systems. Section III

---

5 AI Act, art. 5(1)(a).

6 On the group-level model of vulnerability, see Lisa Waddington, 'Exploring Vulnerability in EU Law: An Analysis of "Vulnerability" in EU Criminal Law and Consumer Protection Law' (2020) 45(6) *European Law Review* 779; Florencia Luna, 'Elucidating the Concept of Vulnerability: Layers Not Labels' (2009) 2(1) *International Journal of Feminist Approaches to Bioethics* 121.

7 AI Act, art. 5(1)(b).

explores the concept of manipulation, drawing on recent legal scholarship to delineate the specific risks posed by AI-facilitated manipulation. Section IV examines the EC's proposed bans, arguing that neither is suitably designed to address the imminent threats posed by AI-facilitated manipulation. The section discusses the need for the EC to move away from the vision of a potential manipulee as a rational, independent, and self-sufficient person devoid of idiosyncrasies toward one of the manipulee as a potentially vulnerable subject. Section V briefly considers the limitations of current transparency and consent requirements in dealing with the threat of manipulative AI before section VI outlines, in broad strokes, an alternative response to the problem of manipulative AI.

2 The AI Act

If enacted, the AI Act would apply to 'AI systems', which is currently defined to capture a broad range of technologies,[8] including some applications not typically considered artificial intelligence.[9] Among the branches of AI captured by the definition is 'machine learning'. Machine learning is an umbrella term for models

[8] At the time of writing, the European Council presidency had proposed amendments to narrow the definition: Council of the European Union Presidency, '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 - Presidency Compromise Text' (2021), 3, 33 <https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf> accessed 4 February 2022.

[9] Some argued that the originally proposed definition 'covers almost every computer program': Martin Ebers et al, 'The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)' (2021) 4 *J* 589; Cailean Osborne, 'The European Commission's Artificial Intelligence Act Highlights the Need for an Effective AI Assurance Ecosystem - Centre for Data Ethics and Innovation Blog' (*Centre for Data Ethics and Innovation*, 11 May 2021), <https://cdei.blog.gov.uk/2021/05/11/the-european-commissions-artificial-intelligence-act-highlights-the-need-for-an-effective-ai-assurance-ecosystem/> accessed 18 August 2021.

which 'learn' from training data over time,[10] using methods which are 'supervised' (i.e., with labelled input data and human-defined desired outcomes), 'unsupervised' (i.e., without labelled data or desired outcomes) or 'reinforced' (i.e., with feedback on the system's failure/success).[11]

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. As the perceived risk of an AI system increases, intervention under the Act escalates. For 'minimal' to 'low-risk' AI systems, providers will be encouraged to develop voluntary codes of conduct but will otherwise remain un- or self-regulated.[12] 'Limited' risk systems, including systems that interact with humans such as chatbots, will be subject to transparency obligations.[13] 'High-risk' AI systems will be subject to a raft of new rules regarding transparency, data quality and governance, accuracy, human oversight, risk management and record-keeping.[14] Essentially, a system deployed to detect cat faces[15] will not be treated the same as a system deployed by law enforcement to identify humans for criminal investigation or arrest. AI trained to play *Grand Theft Auto* will not be treated the same as an autonomous vehicle circling the streets of Phoenix, Arizona.[16] Spam filters will not be treated the same as CV-sorting systems.

A short list of systems deemed to pose an 'unacceptable risk' would be banned altogether under the Act. Two of the four proposed bans target

---

10 Constance de Saint Laurent, 'In Defence of Machine Learning: Debunking the Myths of Artificial Intelligence' (2018) 14(4) *Europe's Journal of Psychology* 734.

11 Margaret A Boden, *AI: Its Nature and Future* (OUP 2016), 47-48.

12 AI Act, art. 69.

13 AI Act, art. 52.

14 AI Act, arts. 8-17.

15 Liat Clark 'Google's Artificial Brain Learns to Find Cat Videos' (*WIRED*, 26 June 2012), <https://www.wired.com/2012/06/google-x-neural-network/> accessed 18 August 2021.

16 Though an AI system trained to play 'Call of Duty' will be subject to requirements under the AI Act while an autonomous weapons systems deployed exclusively for military purposes will not: AI Act, Recital 12, art. 2(3).

manipulation: namely, the placing on the market, putting into service or use of any AI system that: (i) deploys *subliminal techniques*; or (ii) *exploits any of the vulnerabilities* of a specific group of persons due to their age, physical or mental disability,[17] in each case to materially distort a person's behaviour in a manner that causes or is likely to cause a person physical or psychological harm.[18]

Concerns about the threat of manipulation appear to have motivated the inclusion of these prohibitions. The EC's stated 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.[19] Concerns about AI-enabled manipulation surfaced repeatedly during public consultations in the lead up to the release of the draft AI Act.[20] Numerous civil society organisations, citizens, public authorities and academics raised alarms about the use of AI systems to manipulate human behaviour, opinions and decisions in their submission, citing diverse harms such as identity theft, pricing discrimination, and threats to democratic processes and freedoms.[21]

Despite concerns about the risk of manipulation animating debate and precipitating a proposal by the EC to prohibit certain AI systems, a univocal definition of manipulation and its harms is lacking. Although mentioned

---

17 The European Council presidency has also proposed changes to this aspect of the AI Act by including reference to persons vulnerable due to economic or social situation: Council of the European Union Presidency (n 8) 4 and 38. This proposed amendment does not affect the analysis in this article.

18 AI Act, arts. 5(1)(a)-(b). Article 5(1) sets out two other prohibitions which are not discussed in this article.

19 Explanatory Memorandum (n 3) 12.

20 European Commission, 'Public Consultation on the AI White Paper' (Final Report, November 2020), 16 <https://www.standict.eu/sites/default/files/2021-02/PublicConsultationAIWhitePaper_Finalreportpdf.pdf> accessed 21 December 2022.

21 European Commission, 'Contributions to White Paper on Artificial Intelligence: Public Consultation Towards a European Approach for Excellence and Trust' (20 February-14 June 2020) <https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12270-White-Paper-on-Artificial-Intelligence-a-European-Approach/public-consultation_en> accessed 2 February 2021.

repeatedly throughout the AI Act, neither the text of the Act nor the travaux elaborate upon the nature of the threat of AI-enabled manipulation. A common understanding of manipulation and the harms which flow from it cannot be taken for granted. Further, a coherent account is required if we are to develop sound legal responses to the problem.

3 AI-Facilitated Manipulation and its Harms

Acknowledging the absence of a universal definition of manipulation, various legal scholars have attempted to identify the central features of manipulation.[22] Manipulation has been characterised as a failure to 'sufficiently engage or appeal to [the] capacity for reflection and deliberation'[23] in socially unacceptable ways.[24] Such definitions do seem to capture something essential about manipulation. Yet, open-textured notions of 'sufficiency' and 'social acceptability' leave an unsatisfying degree of room for subjectivity. To regulate manipulative conduct, it is necessary to pin down the criteria which make it socially unacceptable or undesirable, which differentiate manipulation from other types of influence. In a comprehensive and compelling treatment of the subject, Susser, Roessler and Nissenbaum proposed a definition which isolates two features of online manipulation that distinguish it from other attempts to influence decision-making.[25] First, manipulation is an attempt to subvert *conscious* decision-making through *hidden* influence.[26] It involves bypassing reflective and deliberative

[22] Susser, Roessler and Nissenbaum (n 4); Tal Z Zarsky, 'Privacy and Manipulation in the Digital Age' (2019) 20(1) Theoretical Inquiries in Law 157; Ryan Calo, 'Digital Market Manipulation' (2014) 82(4) George Washington Law Review 995; Shaun B Spencer, 'The Problem of Online Manipulation' (2020) 2020(3) University of Illinois Law Review 959.

[23] Cass R. Sunstein, *The Ethics of Influence: Government in the Age of Behavioral Science* (CUP 2015), 88.

[24] Tal Z. Zarsky (n 22) 158.

[25] Susser, Roessler and Nissenbaum (n 4).

[26] Ibid 17.

capacities through covertness. Second, it often involves the exploitation of 'cognitive, emotional, or other decision-making vulnerabilities'.[27] Adopting this definition of manipulation as a starting point, the remainder of this section expands on these elements, the harm of manipulation and the enhanced threat imposed by AI-facilitated manipulation.

A manipulation attempt involves the exploitation of a cognitive, emotional, or other decision-making vulnerability in the manipulee. Vulnerability can describe both a state and a quality. To be in a state of vulnerability means to face 'a risk and lack the resources to avoid the risk or respond adequately to the risk if it materialised'.[28] Vulnerability, as a quality, is best understood as a particular susceptibility to physical and non-physical harm. Manipulators may identify vulnerabilities in targets in different ways, such as by drawing on general learnings about widely held cognitive biases from behavioural science, or personal knowledge of the manipulee.

The surfacing and exploitation of vulnerabilities in decision-making alone does not distinguish manipulation attempts from other modes of influence. The exploitation of cognitive biases may be deployed in persuasion attempts too, where the subject is aware of the intent to influence. For example, a mobile app designed to help a person quit smoking may deploy techniques such as peer pressure or emotive appeals. Even though the app may be exploiting cognitive or emotional vulnerabilities, the subject has not been deprived of their ability to reflect on their decisions. As Naomi Jacobs posits, what separates these sorts of 'persuasive technologies' from manipulation, is not that persuasion is restricted

---

[27] Ibid 27. Other authors have identified the targeting of vulnerabilities as a particularly concerning mode of technologically enabled manipulation: see Calo (n 22) 999; Spencer (n. 22) 978-993.

[28] Jonathan Herring, 'Foreword: Vulnerability and the Law' 41(3) University of New South Wales Law Journal 624, 624.

to purely rational appeals.[29] Instead, influence wielded through persuasive technologies 'is in alignment with the personal goals of the person being influenced'[30] and the subject is aware of the intent to influence.

As Susser et al argue, it is the *hidden* nature of manipulative appeals which separates them from open attempts at persuasion, as well as coercion, which 'leaves [the individual's] capacity for conscious decision-making intact' but 'deprives them of choice'.[31] Where an influence attempt is hidden, the subject is deprived of the chance to identify and critically reflect on the external factors shaping their decisions.[32]

Defining manipulation as covert influence raises an obvious question: what aspect of the manipulative appeal needs to be covert? Philosopher Radim Bělohrad provides some insight by delineating between explicit information communicated by the manipulator to the manipulee ('content') and implicit information about the manipulator's intention behind supplying that information ('intent').[33] If the content of an influence attempt is hidden (say, because it is subliminal), then the intent will necessarily be hidden as well. Take, for instance, an attempt to influence an audience of viewers to quit smoking by flashing images of cancerous lungs for a millisecond during a film. If the images are not consciously perceived by the audience, then the intent to influence their behaviour through exposure to those images is also undetectable. As Bělohrad argues, however, it is not the concealment of content, but the concealment of *intent* to influence the manipulee's decision-making which renders an influence

29 Naomi Jacobs, 'Two Ethical Concerns about the Use of Persuasive Technology for Vulnerable People' (2020) 34(5) Bioethics 519, 520.

30 Ibid.

31 Susser, Roessler and Nissenbaum (n 4) 15.

32 Ibid.

33 Radim Bělohrad, 'The Nature and Moral Status of Manipulation' (2019) 34(4) Acta Analytica 447.

attempt manipulative.[34] Consider the example of government policies designed to promote smoking cessation. To achieve that objective, the government might adopt a coercive approach, prohibiting the sale and purchase of tobacco products. Alternatively, the government might seek to persuade consumers to quit through 'fear appeals' by placing visceral warning labels on tobacco products. In both cases, even if the government's policy objective has not been explicitly articulated to the population, the government's intention to change citizens' behaviour is apparent from the context and nature of the appeal.[35] Contrast these open influence attempts with a tobacco company seeking to increase take up of smoking amongst teenagers by engaging social influencers to place tobacco products in their social media content. In this case, the appeal is not explicit and occurs in an unexpected context, thus obscuring the company's intent to influence.

The covert exploitation of a subject's perceived vulnerabilities in order to induce them to take a particular action gives rise to different kinds of harm depending on the context. Compare, for instance, the implications of manipulation in a commercial context with the potential harms that flow from manipulation in an electoral context. Inducing individuals to shop or gamble in a manner contrary to their short- or long-term interests can lead to financial and, in some cases, psychological harms, to the individual concerned. At a macro-level, such practices may create market inefficiencies by inducing individuals to excessively consume harmful goods or expend time and money avoiding profiling.[36] In contrast, manipulating an individual to vote (or abstain from voting) for a particular candidate deprives the voter of the authorship of their

---

[34] Ibid 455. This view appears to be consistent with Susser, Roessler and Nissenbaum's discussion: Susser, Roessler and Nissenbaum (n 4) 24.

[35] Susser, Roessler and Nissenbaum (n 4), 24.

[36] Calo (n 22) 1026-1027.

vote and infringes upon their democratic rights. At a macro-level, widespread voter manipulation undermines the legitimacy of the democratic process, which is premised on electors making a free and informed decision at the ballot box; even inchoate voter manipulation can erode citizen trust in the legitimacy of democratic processes and institutions. In short, the consequences that flow to individuals, communities and broader society from manipulation will vary on a context-to-context basis.

The consequences of AI-facilitated manipulation also depend on the human-defined objective of the relevant system. An AI system programmed to detect and curb impulse shopping or gambling addiction[37] will have a vastly different impact on the manipulee than a system designed to encourage those habits. An AI system designed to mobilise an electorate to participate in an election (and deployed in a non-partisan manner) will have different consequences than a system designed to psychographically profile and target voters in order to dissuade them from voting.[38] In sum, the consequences of manipulation vary depending not only on the context of the manipulation but the manipulator's objectives. This raises the question of what harms, if any, arise from AI-facilitated manipulation for altruistic purposes?

Manipulation, like persuasion and coercion, can be deployed to improve the welfare of subjects. However, in contrast to persuasion, manipulation should be considered an undesirable and unethical practice regardless of the nature of the manipulator's objectives because it deprives the subject of the opportunity to

---

[37] Martin Eden, 'The Influence of Artificial Intelligence on Online Gambling' (*Melanor Games Blog*) <https://meliorgames.com/gambling/the-influence-of-artificial-intelligence-on-online-gambling/> accessed 4 October 2021.

[38] Matthew Rosenberg, Nicholas Confessore and Carole Cadwalladr, 'How Trump Consultants Exploited the Facebook Data of Millions' (*The New York Times*, 17 March 2018), <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html> accessed 4 October 2021.

critically reflect on those objectives.[39] Manipulation obstructs the subject's capacity to pursue alternative and self-defined objectives. By seeking to covertly induce the subject to act in accordance with the *manipulator's* and not the *subject's* objectives, a manipulator interferes with the subject's ability to pursue their own notion of good. Even where the defined objective of a manipulative AI system is benevolent or beneficent or coincidentally aligned with the subject's objectives, the system is nonetheless indifferent to the subject's objectives, goals, and desires. There may well be situations in which manipulation is justified by welfare concerns which outweigh concerns about encroachments on autonomy. However, such situations – where the best or only option is to *covertly* exploit the vulnerabilities of a person to protect their welfare or the welfare of others in the community, rather than persuade or even coerce – are likely to be exceptional.[40] In other words, what makes manipulation, as defined above, undesirable in any context, regardless of the consequences, is that it deprives the manipulee of the opportunity to consciously pursue their own objective, goal, or desires.

Does the threat to autonomy imposed by AI-facilitated manipulation justify regulation? There are reasons to be doubtful. Manipulation is an ancient pattern of behaviour, permeating social, economic, and institutional relationships long before, and without the assistance of, recent advances in machine learning models. Eradicating manipulation from the social fabric altogether is a fanciful goal; so long as the natures and conditions which motivate manipulation exist, so too will the threat of manipulation. However, the inevitability of manipulation in human-to-human interactions should not obscure the heightened and growing threat to autonomy arising from manipulative AI systems. That heightened threat arises from certain capabilities

---

[39] Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Technology, Autonomy, and Manipulation' (2019) 8(2) Internet Policy Review 1, 10-11.

[40] Bělohrad (n 33) 460-461.

of machine learning and the interaction of machine learning with other elements of socio-technical systems.

The first is the capacity to *systematically* uncover vulnerabilities in decision-making processes specific to an individual or small group and adapt an appeal accordingly. Using massive datasets and machine-learning techniques, manipulators can personalise appeals at an individual-level, and further refine and adapt appeals based on feedback from the individual, all while the individual remains oblivious to those processes.[41] Compare this to manipulation without the aid of machine learning; a human manipulator may act on knowledge about a manipulee's compulsions or addictions acquired over the course of their relationship, or knowledge on cognitive biases acquired through general research, to steer a manipulee toward certain choices. While a human manipulator may have the benefit of personal knowledge of the prospective manipulee or general insights from behavioural science, a machine learning system may have access to, or may construct, granular and dynamic behavioural profiles to identify and exploit a vulnerability in the manipulee's decision-making processes. These capabilities are the product not only of machine learning techniques but other features and functionality of digital environments. Device mobility, the proliferation of sensors and tracking technologies, the interoperability of web-based services[42] and advances in storage capacity, among other things, have contributed to the capture of vast pools of continuously replenished data. Machine learning techniques can be applied to rapidly convert these streams of data into predictions about vulnerabilities, to uncover hidden

---

41 Calo (n 22) 1003-1018; Zarsky (n 22) 169; Christian Ernst, 'Artificial Intelligence and Autonomy: Self-Determination in the Age of Automated Systems' in Thomas Wischmeyer and Timo Rademacher (eds.), *Regulating Artificial Intelligence* (Springer International Publishing 2020).

42 Seda Gürses and Joris van Hoboken, 'Privacy After the Agile Turn' in Evan Selinger, Jules Polonetsky and Omer Tene (eds), *The Cambridge Handbook of Consumer Privacy* (CUP 2018).

patterns, associations, clusters, or correlations, 'learning' from new data over time without explicit programming. In agile and dynamic digital environments, machine learning algorithms can help accelerate the tailoring and targeting of stimuli and environments based on those predictive outputs. As Calo pointed out, human manipulators cannot tailor their environment or appearance to better appeal to their targets.[43] The persistent tracking and analysis of an individual's behaviour, and the tailoring of stimuli and environments to exploit individual-level vulnerabilities, poses a novel and heightened threat to autonomy.[44]

The second feature is the volume and velocity of appeals tailored using machine learning in digital environments. For instance, automated systems for ad buying and placement on digital media, which algorithmically match ads with millions of consumers in less than a second,[45] allow for dynamic, personalised stimuli to be delivered at opportune moments, on a massive scale. Dominant digital platforms deploy machine learning in their ad delivery processes to predict the likelihood of a user taking an advertiser's desired action in response to the ad, and to swiftly 'learn' and refine those predictions based on response signals: clicks, views, conversions.[46] The appearance of tailored and targeted

---

43 Calo(n 22), 1021.

44 Ibid.

45 Ibid; Jeff Chester and Kathryn C Montgomery, 'The Role of Digital Marketing in Political Campaigns' (2017) 6(4) Internet Policy Review <https://policyreview.info/articles/analysis/role-digital-marketing-political-campaigns> accessed 21 December 2022.

46 For example, Facebook's documentation states that '[t]o find the estimated action rate, machine learning models predict a particular person's likelihood of taking the advertiser's desired action, based on the business objective the advertiser selects for their ad, like increasing visits to their website or driving purchases. To do this, our models consider that person's behavior on and off Facebook, as well as other factors, such as the content of the ad, the time of day, and interactions between people and ads.': Facebook Business, 'Good Questions, Real Answers: How Does Facebook Use Machine Learning to Delivery Ads?' (2020) <https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads> accessed 26 July 2022. See also Google Ads Help, 'Putting Machine Learning into the Hands of Every User' (2018) <https://support.google.com/google-ads/answer/9065075?hl=en-AU> accessed 26 July 2022. As various academic studies and journalistic investigations explicate, numerous factors

stimuli as 'dark posts' on digital platforms can help to ensure that the intent to modify behaviour remains hidden to the subject. Collective evaluation and public discussion may help to uncover broad-based manipulation attempts, making more people aware and therefore better equipped to avoid the attempt. For example, many are aware of the intent behind common retail practices such as 'charm' and 'open the wallet' pricing by media reporting, and public scrutiny and discussion.[47] By contrast, manipulative stimuli produced by black box AI systems, tailored and targeted at the individual or small group rather than at population-level, are less susceptible to the same kind of public scrutiny.

The third feature, a relative advantage of most machines over humans, is the relentlessness and stamina of manipulative AI systems. As Jacob's put it, 'technology is inherently persistent: a computer does not get tired, discouraged or frustrated like humans do'.[48] Unlike a human manipulator or manipulee, the processing power of an AI system is not limited by biological inconveniences like fatigue.

shape the predictive outputs of the machine learning models (sometimes leading to discriminatory outcomes), including the targeting parameters selected by advertisers and image classifications: Julia Angwin, Ariana Tobin and Madeleine Varner, 'Facebook (Still) Letting Housing Advertisers Exclude Users by Race' (*ProPublica*, 21 November 2017)<https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin> accessed 26 July 2022; Muhammad Ali et al, 'Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes' (2019) 3 Proceedings of the ACM on Human-Computer Interaction 199.

47 'Charm' pricing is the widespread practice of setting prices just-below round numbers for goods and services (for e.g. $9.99): Natasha Noman, 'Yes, Grocery Stores are Tricking You into Spending More Money' (*Business Insider*, 23 February 2017) <https://www.businessinsider.com/yes-grocery-stores-are-tricking-you-into-spending-more-money-2017-2> accessed 14 October 2021. 'Open the wallet' pricing involves the placement of cheaper goods in prominent areas of a physical store: Kristin Wong, '5 Sneaky Tricks Grocery Stores Use to Make You Spend More Money' (*Mental Floss*, 4 March 2016) <https://www.mentalfloss.com/article/76469/5-sneaky-tricks-grocery-stores-use-make-you-spend-more-money> accessed 14 October 2021.

48 Jacobs (n 29) 521.

4 The EC's Approach to AI-Facilitated Manipulation

The bans proposed by the EC in the AI Act respond to the two core features of manipulation – hiddenness and the exploitation of vulnerabilities – but fail to properly grapple with the specific threat posed by the use of machine learning for manipulation. The first ban targets covert attempts at influence but focuses solely on hidden content. In doing so, the ban overlooks individual differences in processing sensory information and the ability of AI systems to induce or detect lapses in perception and attention. The second ban adopts a group-level model of vulnerability, overlooking the growing capabilities of machine learning models to covertly surface and exploit individual vulnerabilities.

Both prohibitions also include an element of harm which significantly narrows their ambit. Specifically, for the prohibitions to apply, the relevant AI systems must be placed on the market, put into service or used 'in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm'.[49] Establishing a causal link between the operation of the relevant AI system and actual or likely physical or psychological harm to a person will be a significant hurdle due to the lack of explainability and potential foreseeability of a deep learning system's 'thought process'.[50] Further, the link is presumed broken where harm arises from external factors outside the control of the user or provider of

[49] AI Act, art. 5(1)(a). As drafted, there is some ambiguity in the provisions as to whether the intention of the user or provider of the AI system relates solely to the material distortion of a person's behaviour, or whether the user or provider must also intend to cause the resulting or likely psychological or physical harm. Recital 16 of the AI Act appears to suggest that the harmful result may not need to be intended but merely risked, stating that '[t]he placing on the market, putting into service or use of certain AI systems intended to distort human behaviour, *whereby physical or psychological harms are likely to occur*, should be forbidden'.

[50] See Zhao Yan Lee, Mohammad Ershadul Karim and Kevin Ngui, 'Deep Learning Artificial Intelligence and the Law of Causation: Application, Challenges and Solutions' (2021) 30(3) Information & Communications Technology Law 225.

the AI system,[51] potentially excluding harm resulting from interactions between AI and other actors, functions or technologies.[52] As Veale and Borgesuis observe, the harm requirement also overlooks cumulative and collective harms,[53] including the macro-level consequences touched upon in section III. For example, widespread voter manipulation may not result in physical or psychological harm to affected individuals but may undermine or damage democratic institutions and processes. In sum, the proposed provisions cover a narrow subset of the harmful consequences that could flow from the deployment of manipulative AI.

The focus of the remainder of this section will be on the conception of manipulation, specifically the constructs of hidden influence and vulnerability, embedded in the proposed prohibitions.

4.1 Subliminal Techniques

The covert nature of AI-enabled subliminal techniques appears to be the EC's key reason for affording special treatment to subliminal techniques.[54] The prohibition targets concealed stimuli intended to induce behavioural changes (what would be considered 'content' in Bělohrad's dichotomy). Indeed, subliminal techniques are not merely covert; such techniques are theoretically imperceptible by the conscious mind.

To understand the difficulties inherent in implementing a legal ban on subliminal techniques, AI-facilitated or otherwise, it is necessary to understand some key concepts. Subliminal techniques, by definition, target below (*sub*) the

51 AI Act, Recital 16.

52 Michael Veale and Frederik Zuiderveen Borgesuis, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 4 Computer Law Review International 97.

53 Ibid, 99.

54 AI Act, Recital 16; Explanatory Memorandum (n 3) 12.

threshold of consciousness (*limen*).[55] Such techniques are premised on the notion of subliminal *perception*, which occurs when a subject unconsciously perceives a stimulus. Subliminal *persuasion* involves targeting stimuli below the threshold of a person's consciousness in order to influence their behaviour, thoughts, or emotions.[56] Subliminal persuasion thus depends upon a subject unconsciously perceiving stimuli intended to influence their behaviour, and on that stimulus actually shaping the subject's behaviour. The proposed ban applies to the use of 'an AI system that deploys subliminal techniques […] in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm',[57] and therefore applies to techniques intended for subliminal *persuasion*, not techniques which merely target subliminal perception.[58]

The first obstacle to applying the ban will be establishing a causal link between subliminal stimuli and a change in behaviour which leads to psychological or physical harm. Empirical evidence on the efficacy of subliminal persuasion is mixed.[59] Further, proving that the relevant stimuli is subliminal presents a challenge. The concept of 'subliminal' is indeterminate due to the

---

55 'subliminal, adj. and n.', *OED Online* (3rd edn, OUP March 2022) <https://www.oed.com/view/Entry/192773?redirectedFrom=subliminal#eid> accessed 21 December 2022.

56 Laura Smarandescu and Terence A Shimp, 'Drink Coca-Cola, Eat Popcorn, and Choose Powerade: Testing the Limits of Subliminal Persuasion' (2015) 26(4) Marketing Letters 715, 716; Nicholas Epley, Kenneth Savitsky and Robert A Kachelski 'What Every Skeptic Should Know About Subliminal Persuasion' (1999) 23(5) Skeptical Inquirer 40, 41.

57 AI Act, art. 5(1)(a). The section also applies to 'the placing on the market' and 'putting into service' of such an AI system.

58 In contrast to Article 9 of Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States Concerning the Provision of Audiovisual Media Services [2010] OJ L95/1 (the 'Audiovisual Media Services Directive'), which takes a strict liability approach, specifying that Member States should ban the use of subliminal techniques in 'audiovisual commercial communications'.

59 See Timothy E Moore, 'Subliminal Advertising: What You See Is What You Get' (1982) 46(2) Journal of Marketing 38; Ben R Newell and David R Shanks, 'Unconscious Influences on Decision Making: A Critical Review' (2014) 37 Behavioral and Brain Sciences 1, 13-15.

instability of the conscious/unconscious dichotomy which underpins it. A brief tour through some well-known examples reveals the key difficulties in categorising techniques as subliminal.

Subvisual and subaudible cueing are undoubtedly the most well-known and perhaps clear-cut examples of subliminal techniques. Subvisual cueing involves the flashing of visual stimuli for a fraction of a second – at a rate too rapid to be detected by a human eye; a technique made notorious by the market researcher James Vicary who claimed that he was able to increase the sales of Coca-Cola and popcorn by repeatedly flashing the slogans 'Drink Coca-Cola' and 'Eat Popcorn' for 1/3000 of a second during a film screening.[60] Similarly, subaudible messaging involves playing sounds at a volume which is inaudible to a human ear.[61] Take, for example, the 'Dr Becker's Blackbox', an 'honesty enhancement program' deployed in a shopping mall in the US, which was designed to play an audio recording of phrases such as 'I am honest. I will not steal' repeatedly at an inaudible volume, in an effort to subliminally deter shoppers from engaging in theft.[62] The theory underlying subvisual and subaudible cueing is that a stimulus which is either too rapid or too quiet to be detected by sensory receptors can still be perceived, only at an *unconscious* level.

Subvisual and subaudible cues are not the only techniques commonly labelled subliminal. Another visual technique often labelled as subliminal is the use of embedded images (or 'embeds').[63] An oft-cited example is the KFC TV commercial which contained an embedded image of a US dollar bill nestled in

---

60 Sheri J. Broyles, 'Misplaced Paranoia Over Subliminal Advertising: What's the Big Uproar this Time?' (2006) 23(6) The Journal of Consumer Marketing 312, 312; Ronald A Fullerton, '"A Virtual Social H-Bomb": The Late 1950s Controversy over Subliminal Advertising' (2010) 2(2) Journal of Historical Research in Marketing 166, 166-167.

61 Moore (n 59) 43-44.

62 Dave Kindred, 'Message Creeps in on Little Feet' (*Washington Post*, 11 February 1981) <https://www.washingtonpost.com/archive/sports/1981/02/11/message-creeps-in-on-little-feet/f784be2a-0d93-45ce-8ae4-4557db0c9d44/> accessed 14 October 2021.

63 Moore (n 59) 45-46.

the lettuce layer of a chicken burger. Many speculated that the bill was an attempt to associate the product with wealth.[64]

These examples highlight some of the challenges inherent in categorising techniques as subliminal. The first is a practical concern. Subliminal stimuli can either be uncovered when revealed by the manipulator or detected by a viewer or listener. Manipulators acting in contravention of a prohibition on subliminal techniques are unlikely to self-incriminate, yet once a stimulus is consciously detected by a viewer/listener, it arguably crosses the threshold into supraliminal.

A second problem with the concept of 'subliminal techniques' arises from the interdependence between perception and attention. Embeds like the KFC dollar bill might be better viewed as *obscure* or *unattended* rather than subliminal. The line between unattended and imperceptible is further blurred in cases of misdirection – where a manipulator purposely pulls a viewer's attention toward certain visual stimuli in an effort to prevent them from attending to others – a common design feature of websites. Harry Brignull, who researches and documents 'dark patterns', illustrates the technique using the example of a budget airline website which pre-selects the customer's seats (at an additional cost) while including the alternative (free) option of skipping seat selection in small print at the bottom of the webpage.[65] The technique relies on insights from behavioural science about common biases and vulnerabilities in attention. It is not difficult to imagine the super-charging of such dark patterns using machine

---

64 Dylan Love, 'The Shocking and Incredible Coke History of Subliminal Advertising' (*Business Insider Australia*, 27 May 2011) <https://www.businessinsider.com.au/subliminal-ads-2011-5?r=US&IR=T#a-hidden-dollar-in-this-kfc-sandwich-links-it-to-power-and-wealth-8> accessed 14 October 2021.

65 Harry Brignull, 'Misdirection' (*Dark Patterns*) <https://www.darkpatterns.org/types-of-dark-pattern/misdirection> accessed 4 February 2022.

learning models to dynamically personalise a digital environment based on individual-level biases and vulnerabilities to inattention.[66]

In this example, the 'skip seat selection' text, like the dollar bill in KFC's chicken burger, is obscured. Neither stimulus is technically 'below the threshold of consciousness'. With sufficient attention to detail, the stimuli can be perceived. However, 'sludge'[67] on the airline website obfuscates the optimal choice, gaming the traveller's selective attention. Should misdirection be classified as subliminal technique? Or should a line be drawn between stimuli which *cannot* be consciously perceived and stimuli which is visible or audible but *unattended*.[68] The problem with such a distinction is that sensory perception is contingent upon attentional processes.[69] That link is writ large in cases of inattentional blindness, which occur when a person fails to detect an unexpected yet visible stimulus.[70] This interdependence between attention and perception means that drawing a line between non-perceptible and unattended stimuli will, in many cases, be impossible.

This leads to another fundamental problem with the proposed prohibition, which is that there is no bright line between the sub- and supra-liminal. Not only does the boundary between conscious and unconscious – the

---

66 Zakary Kinnaird, 'Dark Patterns Powered by Machine Learning: An Intelligent Combination' (*Medium*, 16 October 2020) <https://uxdesign.cc/dark-patterns-powered-by-machine-learning-an-intelligent-combination-f2804ed028ce> accessed 1 October 2021.

67 Richard R Thaler and Cass R Sunstein, *Nudge* (Final Edition, Allen Lane 2021), ch. 8.

68 Sheri J Broyles, 'Misplaced Paranoia over Subliminal Advertising: What's the Big Uproar This Time?' (2006) 23(6) Journal of Consumer Marketing 312, 312.

69 Ronald A Resnick, 'Perception and Attention' in D. Reisberg (ed.), *Oxford Handbook of Cognitive Psychology* (OUP 2013).

70 Arien Mack and Irvin Rock, *Inattentional Blindness* (MIT Press 1998) cited in *ibid*. There is evidence that even where individuals fail to detect a stimulus, due to inattentional blindness, it can still influence or 'prime' their later responses to tasks, indicating the occurrence of nonconscious perception: Arien Mack, 'Inattentional Blindness: Looking without Seeing' (2003) 12(5) Current Directions in Psychological Science 180, 181–182.

'absolute threshold'[71] – vary between individuals,[72] a given individual's ability to consciously detect stimuli also fluctuates.[73] Psychology researchers have found that a range of factors may affect thresholds of conscious perception, including physiological conditions, such as fatigue or hunger,[74] age,[75] motivation, personality traits,[76] and environmental conditions.[77] This variability has led some to suggest that perception is better viewed 'as a continuum of sensory states than a binary state'.[78] The question that arises, then, is: for whom or how many and at what times must the stimuli be subliminal?

Attempting to apply a universal standard for subliminal perception is inherently fraught. Constructing an 'ordinary' or 'average' standard for processing sensory information would be discriminatory in its effect, affording less protection to those with temporary or permanent sensory perception abilities below the so-called 'average'. Further, an objective legal standard for sensory perception is incongruent with the dynamic and personalised nature of AI-enabled manipulation. On the topic of one ubiquitous objective standard – the 'reasonable person', Wendy Parker observed: 'one of the functions of the standard is to eliminate idiosyncrasies; to remove the personalised nature of the

---

71 The absolute threshold is 'the lowest or weakest level of stimulation (e.g., the slightest, most indistinct sound) that can be detected on 50% of trials.': Gary R. VanderBos (ed.), *APA Dictionary of Psychology* (2nd edn, APA 2015) 4.

72 Fullerton (n 60) 167.

73 As noted in the APA Dictionary of Psychology definition, 'Although the name suggests a fixed level at which stimuli effectively elicit sensations, the absolute threshold fluctuates according to alterations in receptors and environmental conditions': VanderBos (n 71) 4.

74 James V McConnell, Richard L Cutler and Elton B McNeil, 'Subliminal Stimulation: An Overview' (1958) 13(5) American Psychologist 229, 232.

75 Larry E Humes et al, 'The Effects of Age on Sensory Thresholds and Temporal Gap Detection in Hearing, Vision, and Touch' 71(4) Attention, Perception & Psychophysics 860.

76 Stuart L. Smith, 'Extraversion and Sensory Threshold' (1968) 5(3) Psychophysiology 293.

77 VanderBos (n71) 4.

78 Stefan Wiens, 'Subliminal Emotion Perception in Brain Imaging: Findings, Issues, and Recommendations' (2006) 156 Progress in Brain Research 105, 105.

conduct by measuring it against a commonly accepted standard'.[79] This is true of all objective legal standards – they remove the need for individual assessments of idiosyncrasies and vulnerability. The fundamental problem with ignoring the existence of individual idiosyncrasies and vulnerabilities when applying a law designed to restrict manipulative AI systems is that the manipulative power of such a system largely arises from its capacity to learn and identify those individual-level idiosyncrasies and vulnerabilities.

Machine learning provides manipulators with greater opportunities to learn and detect individual vulnerabilities in a manipulee's decision-making processes in real-time, and to tailor stimuli to those vulnerabilities and target them at optimal moments. Such techniques could be applied to detect the presence of factors which render a subject more vulnerable to subliminal stimuli. For instance, researchers have made strides in the use of artificial neural networks and other AI techniques to detect the presence of fatigue and anxiety.[80] Such research holds promise for improving human health and safety but could be misappropriated for the purpose of discerning when a subject is potentially more vulnerable to subliminal targeting due to fatigue. A Facebook pitch document, leaked in 2017, offered a glimpse into the possibilities. In the pitch, Facebook Australia touted its ability to infer the emotional states of young users, such as stress, anxiousness, silliness, uselessness, and nervousness, and to target them at moments of insecurity, such as when they are interested in 'working out and losing weight'.[81] While the company denied engaging in such targeting, the

---

79 Wendy Parker, 'The Reasonable Person: A Gendered Concept Claiming the Law - Essays by New Zealand Women in Celebration of the 1993 Suffrage Centennial' (1993) 23 Victoria University of Wellington Law Review 105, 107.

80 Vidhi Parekh, Darshan Shah and Manan Shah, 'Fatigue Detection Using Artificial Intelligence Framework' (2019) 5 Augmented Human Research 5.

81 Darren Davidson, 'Facebook Exploits "Insecure" to Sell Ads' (*The Australian*, 1 May 2017).

incident highlights the risk of AI being utilised to identify individuals at vulnerable moments for targeted stimuli.

The hypothetical figures which underpin objective legal standards such as the 'reasonable person', the 'bonus pater familias', the 'average consumer', are devoid of individual vulnerabilities. Instead, such legal fictions are intended to reflect the 'average' or 'common' abilities of persons in order to consistently apportion responsibility, construct appropriate levels of care and discern 'normal' responses in a given situation. Such an approach is fraught and incompatible with the operating mode of manipulative AI systems designed to surface and exploit individual vulnerabilities in manipulees.

The alternative is to apply a subjective standard to assessing subliminality, taking into account the particular characteristics and circumstances of the subject. However, a subjective standard could lead to unreasonable allocations of liability in cases where a subject suffers from a lapse of sensory perception or attention for reasons unrelated to the methods deployed by the AI system. Of course, as discussed above, a manipulator might use machine learning techniques to intentionally *induce* a lapse in attention, and in such cases, it seems reasonable that the manipulator should bear responsibility for harm arising from that inducement. In such cases, however, it is the use of an AI system to detect, exploit and induce a *vulnerability* in the subject, rather than the use of the subliminal technique, which poses a threat of harm.

4.2 Exploiting Vulnerability

The second proposed prohibition acknowledges the growing AI-driven capabilities to detect and exploit vulnerabilities but is limited to 'practices that have a significant potential to […] exploit vulnerabilities of specific vulnerable

groups'.[82] The prohibition assigns vulnerable status to certain groups on the basis of specific, shared characteristics: age and physical or mental disability.[83] Under this approach, defined groups are deemed to be inherently vulnerable while the rest of the population is not.

A key assumption underpinning the approach is that individuals outside of specific groups possess sufficient resources to resist covert and systematic attempts to influence their behaviour. The approach is built upon an idealised vision of the citizen and consumer as an autonomous being who, with reasonable information, is independent, rational, and self-sufficient. The assumption of the citizen and consumer as an autonomous being supports a regulatory model that focuses on providing individuals with enough information and opportunities to withhold consent. The approach reflects a group-level model of vulnerability which is commonly adopted in legal and policy contexts to locate or justify the need for additional care or protection by the state.[84] Under these legal frameworks, vulnerable groups are presumed to lack sufficient resources and capabilities to protect themselves against harm and that presumption is used to justify not only additional protection, but at times paternalistic interventions into decision-making and areas of life usually beyond the reach of the state.[85] Thus,

---

82 Explanatory Memorandum (n 3) 12.

83 AI Act, art. 5(1)(b).

84 For instance, EU consumer protection law recognises the 'vulnerable consumer' who is deemed under the law to more susceptible to unfair commercial practices due to 'mental or physical infirmity, age or credulity': 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 [2005] OJ L149/22 (hereinafter 'UCPD'), art. 5(3); Eleni Kaprou, 'The Legal Definition of "Vulnerable" Consumers in the UCPD' in Christine Riefa and Severine Saintier (eds.), *Vulnerable Consumers and the Law: Consumer Protection and Access to Justice* (Routledge 2020).

85 See Michael C Dunn, Isabel CH Clare and Anthony J Holland, 'To Empower or to Protect? Constructing the "Vulnerable Adult" in English Law and Public Policy' (2008) 28(2) Legal Studies 234.

the dominant conception of vulnerability under the law is incongruent with classical liberal notions of autonomy.[86]

Broadly, the approach adopted in the AI Act does not sufficiently engage with the central problem posed by manipulative AI systems. The approach not only assumes that groups of individuals assigned vulnerability status under the Act are monolithic but ignores the existence of vulnerability outside these groups.[87] A law designed to address manipulative AI must acknowledge the enhanced capabilities to surface and exploit individual-level vulnerabilities afforded by machine learning systems. Finally, the approach rests on a false dichotomy between vulnerability and autonomy.

A legal response adapted to the threat of manipulative AI requires a paradigm shift, away from the classical liberal assumption of the autonomous being toward the notion of all manipulation targets as potentially vulnerable subjects.[88] This would require that legislation be drafted in a manner which recognises that vulnerability may result from the interaction of an individual's particular characteristics and an AI system (or an environment shaped by an AI system).[89] To be clear, a shift away from the classical liberal notion of the rational, self-sufficient individual does not require a rejection of the value and right to autonomy. Rather, it requires an acknowledgement of the actual conditions of human decision-making, of the socially and technologically embedded nature of autonomous agents, and of their cognitive and attentional limitations. Most importantly, it requires a rejection of manipulative stimuli generated by AI systems as an acceptable burden on autonomous decision-making which

---

86 Martha Albertson Fineman, 'The Vulnerable Subject: Anchoring Equality in the Human Condition' 20(1) Yale Journal of Law and Feminism 1, 11.

87 Kaprou (n 84), 67.

88 A perspective broadly in line with the 'universal' conception of vulnerability articulated by Martha Fineman: Fineman (n 86).

89 In line with Florencia Luna's observation that vulnerability can be relational in that a particular situation '*makes* or *renders* someone vulnerable': Luna (n 6) 129.

individuals must overcome with information and hyper-observance. In practical terms, such a shift would entail a shift in emphasis from transparency and consent mechanisms toward enhanced protections to ameliorate the threat of AI-facilitated manipulation. Before exploring what form such protection could take, it is necessary to consider the limitations of existing transparency and consent mechanisms in addressing the threat posed by manipulative AI systems.

5 Transparency and Consent

The Explanatory Memorandum to the AI Act reflects the view that the transparency and consent mechanisms afforded by other legislation provide sufficient protection against manipulative practices for adults outside the identified 'vulnerable' groups. Section 5.2.2 of the Explanatory Memorandum states that:

> [o]ther 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 (emphasis added).

EU data protection legislation and consumer protection legislation contains provisions aimed at addressing information asymmetries between data subjects and data controllers/processors, and consumers and business. The shortcomings of the transparency and consent model underpinning these legal frameworks has been the subject of extensive critique elsewhere.[90] This section considers the adequacy of recent and proposed enhancements to transparency requirements

[90] See, eg: Daniel J Solove, 'Introduction: Privacy Self-Management and the Consent Dilemma' (2013) 126 Harvard Law Review 1880.

regarding automated decision-making and online advertisement in revealing manipulation attempts.

Data is fuel for AI systems. As Kate Crawford explains,

> [m]achine learning models require ongoing flows of data to be more accurate. But machines are asymptomatic, never reaching full precision, which propels the justification for more extraction from as many people as possible to fuel the refineries of AI.[91]

Generally speaking, in order to accurately detect and exploit vulnerabilities in an individual's decision-making processes, an AI system requires fresh and continuous streams of data about the individual's behaviour. Thus, an obvious way to curtail AI-facilitated manipulation is to stem the flow of behavioural data for such purposes. However, evading behavioural profiling while participating in the digital economy is an onerous endeavour. Transparency notices, mandated under the General Data Protection Regulation[92] and ePrivacy Directive,[93] have limited utility for many 'data subjects' who encounter thousands of pages of data protection policies and cookie banners as they glide seamlessly between digital services which are increasingly integral to their social, economic and political lives, and which collect data about their online behaviour as a condition of service. Machine learning models often use a representative training sample to draw inferences about a larger group of people about whom a smaller amount of

---

91 Kate Crawford, *Atlas of AI* (Yale University Press 2021).

92 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) [2016] OJ L119/1 (hereinafter 'GDPR').

93 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) [2002] OJ L 201/37 (hereinafter 'ePrivacy Directive').

data is available but who share the same observable traits.[94] Consequently, the most diligent and circumspect data subjects, who carefully read privacy policies and adjust their cookie preferences, and who deliberately withhold certain data, may still be affected by the willingness of others to share it.[95] All this makes for an environment in which it is increasingly arduous, sometimes futile, to attempt to stem the flow of behavioural data which would feed manipulative AI.

Privacy notices given at the point of data collection provide limited protection against future manipulation for individuals who are ill-equipped to predict how data collected in one context will be transformed and repurposed for manipulative ends in a myriad of other contexts. In short, collection notices are ineffective because manipulation is largely unforeseeable. Transparency around the *outputs* (rather than the data *inputs*) of AI systems, set out under proposed and impending EU legislation, may partly overcome this issue, helping to expose manipulation attempts at the point at which they occur. Under the proposed Digital Services Act,[96] online platforms that display advertising will be required to furnish recipients with 'meaningful information about the main parameters used to determine the recipient to whom the advertisement is displayed'.[97] The major digital platforms currently provide users with information about why an ad is shown to them, but such information is high-level, providing little insight into the underlying logic of the platforms' advertising algorithms. Hence, the efficacy of the new transparency requirements in the Digital Services Act will turn on what constitutes *meaningful information*. The AI Act similarly requires transparency around certain outputs of AI systems

94 Solon Barocas and Helen Nissenbaum, 'Big Data's End Run Around Procedural Privacy Protections' (2014) 57(11) Communications of the ACM 31.

95 Ibid.

96 Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market For *Digital* Services (Digital Services Act) and Amending Directive 2000/31/EC' COM/2020/825 final, 15 December 2020 (hereinafter 'Digital Services Act').

97 Ibid, art. 24(c).

which may be manipulative. Specifically, users of AI systems that generate manipulated images, audio or video ('deep fakes') would be required to inform viewers that the content is manipulated.[98] These new transparency requirements would go some way toward exposing hidden influence attempts as a manipulator's intent will often (though not always) be apparent from information about the parameters used to determine the audience for an advertisement, or the fact that a video has been manipulated. However, advertising and deep fakes do not cover the field when it comes to manipulative stimuli. As discussed earlier, manipulation may take the form of dark patterns, embeds or targeted ads served at moments of vulnerability.

In addition to transparency requirements, the EC cites existing rights not to be subject to profiling as another safeguard against manipulation. Article 22 of the GDPR establishes a right not to be subject to automated decisions, including 'profiling'. In order for the right to apply, an automated decision must produce 'legal' or 'similarly significant' effects.[99] Further, automated decision making, including profiling, is permitted where necessary for performance of a contract or the subject provides consent.[100] In collection notices, individuals must be informed about 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject'.[101] Thus, the right is displaced where organisations obtain consent after furnishing individuals with prescribed information.

---

98 AI Act, art. 52(3).
99 GDPR, art. 22(1).
100 Ibid, art. 22(2)(a) and (c).
101 GDPR, arts. 13(2)(f), 14(2)(g), and 15(1)(h).

Given that manipulation involves influencing an individual's decisions, rather than subjecting them to a decision, the relevance of article 22 to the problem of manipulative AI may seem limited. However, guidance from the Article 29 Working Party indicates that the right extends beyond decisions like automatic refusals of credit or university applications.[102] Importantly, a decision to target content to an individual may constitute a 'decision' for the purpose of article 22. The Working Party advised that targeted advertising based on profiling could be construed as having 'significant effects' on an individual for the purpose of article 22 if based on 'knowledge of the vulnerabilities of the data subjects targeted'.[103] However, the Working Party appears to adopt a group-level understanding of vulnerability, stating that '[p]rocessing that might have little impact on individuals generally may in fact have a significant effect for certain groups of society, such as minority groups or vulnerable adults'.[104] It does, however, go on to give the example of an entity targeting a person in a vulnerable financial position with ads for high interest loans,[105] suggesting it has in mind a more expansive list of vulnerable groups than that reflect in the AI Act.

The EC also cites EU consumer protection laws, which seek to address information asymmetries between consumers and businesses by requiring that businesses provide all 'material information that the average consumer needs, according to the context, to take an informed transactional decision'.[106] Although, as noted above, these laws seek to accommodate vulnerability, the law also reflects a group-level model.[107] Under the UCPD, even consumers who are

---

[102] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' wp251rev.01, adopted on 3 October 2017, 22.

[103] Ibid.

[104] Ibid.

[105] Ibid.

[106] UCPD, art. 7.

[107] Waddington (n 6) 796.

classified as 'particularly vulnerable' are assessed against an objective standard – an average consumer of the group to which they belong.[108]

As discussed above, the group-level model of vulnerability, which is reflected in the UCPD and the Working Party's interpretation of article 22, does not sufficiently meet the threat imposed by manipulative AI.[109] The model does not contemplate heterogeneity within 'vulnerable groups' nor vulnerability outside those groups.[110] It overlooks the capabilities of manipulative AI to render persons outside the identified groups vulnerable by surfacing and exploiting biases and other weaknesses in their decision-making.

Approaches which rely on transparency and consent mechanisms to empower individuals to resist manipulative AI optimistically overestimate the capacities of the vast majority of people to compete with the processing power, stamina and persistence of an AI system designed to surface weaknesses and biases in their decision-making, and to individually-tailor conditions and stimuli to exploit those biases and vulnerabilities.

Undoubtedly, a small minority of people will be able to resist AI-facilitated manipulation in certain situations with the assistance of transparency and consent mechanisms. However, policy and lawmakers must consider the long-term, corrosive effects of insidious and widespread AI-based manipulation. At a time when privacy seeking individuals need to refrain from participation in digital spaces or engage in obfuscation to avoid incessant data profiling, and

---

[108] UCPD, art. 5(3).

[109] Recognising the deficiency in the current approach, one European Parliament Committee adopted the view that 'the concept of vulnerable consumers should also include consumers in a situation of vulnerability, meaning consumers who are placed in a state of temporary powerlessness resulting from a gap between their individual state and characteristics on the one hand, and their external environment on the other hand': Committee on the Internal Market and Consumer Protection Rapporteur, 'Motion for a European Parliamentary Resolution on a Strategy for Strengthening the Rights of Vulnerable Consumers' 2011/2272(INI), <https://www.europarl.europa.eu/doceo/document/A-7-2012-0155_EN.html> accessed 4 February 2022.

[110] Kaprou (n 84) 67.

social media users subjected to relentless abuse and harassment shrink their presence or withdraw from digital life entirely, AI-facilitated manipulation poses a further impediment to human flourishing in digital spaces. From an economic perspective, allowing AI-facilitated manipulation to run rife in markets creates market inefficiencies, requiring consumers to deploy resources they otherwise would not.[111] The time and labour expended detecting and resisting AI-facilitated manipulation could be put to better use. From a democratic perspective, permitting AI-facilitated manipulation to shape electoral contests strikes at the heart of the legitimacy of democratic systems of government. Citizens who wish to avoid a steady diet of stimuli tailored to exploit individual vulnerabilities in their decision-making may simply opt to withdraw from digital spaces. If the benefits of social platforms touted by their owners are to be preserved – the increased opportunities for democratic discussion, social connection and civic collaboration, then digital spaces of congregation must not be left to degenerate into hotbeds of AI-facilitated manipulation. Currently, laws which supposedly guard against manipulative AI envisage many targets as rational and self-sufficient, with abundant time to read collection notices and anticipate manipulative outputs – an ideal few of us measure up to.

6 Responding to AI-Facilitated Manipulation

So far, this article has focused on the need for regulatory attempts to constrain manipulative AI to engage with the specific capabilities of machine learning systems: detecting and exploiting individual-level vulnerabilities in human decision-making processes in order to covertly influence human behaviour. In this final section, I offer some thoughts on an alternative approach to grappling with the problem of manipulative AI.

[111] Calo (n 22) 1027.

It will be important to tread carefully when designing restrictions on manipulative AI, including any prohibition. Before escalating regulatory intervention into the development and use of AI, it is incumbent upon law and policy makers to not only consider the harms of failing to intervene, but to anticipate and seek to avert the harms of ill-fitting or draconian legal responses.[112] A prohibition, the most severe option for regulatory intervention available, could carry significant social and economic costs if inappropriately scoped. An excessively restrictive law which blocks certain avenues of intellectual and experimental inquiry may stifle useful and beneficial innovation. A poorly crafted law may have similar effects by encouraging over-compliance (particularly amongst smaller enterprises without the resources to absorb significant penalties), chilling beneficial AI developments not intended to be caught by the prohibition. Finally, a prohibition should not, counterproductively, disempower those it is designed to protect.

With that in mind, as a starting point, constraints on manipulative AI should be limited in at least two respects. First, such constraints should target manipulation attempts which are *tailored* to the individual based on profiling rather than broad-based efforts. As Sunstein and Thaler note, 'there is no such thing as a "neutral" design' when it comes to the 'context in which people make decisions' – the *choice architecture*.[113] Choice architects ranging from shopkeepers to website owners design decisional contexts in ways that impact human behaviour.[114] Any restriction on AI-facilitated manipulation should not operate to mandate random or neutral choice architecture in on- or offline spaces, but rather target the specific threat imposed by manipulative AI. The kinds of welfare-guided 'nudges' advocated by Sunstein and Thaler, which are

[112] Calo (n 22) 1042.
[113] Thaler and Sunstein (n 67).
[114] Ibid, ch. 5.

'transparent and subject to public scrutiny',[115] should not be prohibited. As discussed in section III, the use of complex machine learning models adds a potentially impenetrable layer of opacity to manipulative practices. The individualised and dynamic nature of AI-facilitated manipulation removes the safeguard of collective scrutiny of manipulative practices. The increased availability of behavioural data has supercharged the ability of would-be manipulators to tailor their appeals.[116] Hence, in order to target the specific threat imposed by manipulative AI, any restriction should apply to the use of AI intended to modify an individual's behaviour based on individual-level vulnerabilities inferred through data profiling, rather than generalised learnings from behavioural science.

Further, individuals wishing to use AI systems to adjust their behavioural patterns in pursuit of a self-defined goal, be it saving money, exercising more frequently, sleeping better, smoking cessation, reducing their alcohol intake or any other objective, should be able to do so. Legal constraints should target systems deployed for the *hidden* objective of shaping human behaviour. Legal constraints should not apply to AI systems that facilitate the achievement of an individual's self-selected objective, even where that objective is neutral or deleterious to their best interests and wellbeing. While the pursuit of certain harmful objectives may be restricted by other laws, a law which is intended to safeguard autonomous decision-making should not operate to prevent individuals from utilising AI to pursue their own goals. In any case, such technologies are persuasive as opposed to manipulative as the intent to influence is clear to the user from the context and circumstances. Nonetheless, there may be grey areas where the line between persuasion and manipulation is somewhat

[115] Cass R Sunstein, 'The Ethics of Nudging' 32(2) Yale Journal of Regulation 413, 428.

[116] Zarsky (n 22) 219; Susser, Roessler and Nissenbaum (n 4) 29-31; Calo (n 22), 1003-1004; Spencer(n 22) 972-973.

blurred; a restriction on manipulative AI should not inadvertently capture persuasive technologies.

However, to ensure that the intent to modify behaviour is in fact overt, individuals wishing to adopt persuasive technologies should be made aware of, and provide express consent to, the behavioural modification techniques deployed by the AI system. If enacted in its current form, the AI Act would impose an obligation upon the providers of any AI systems that interact with individuals to design their systems 'in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use'.[117] In relation to AI systems designed to detect and adjust an individual's behavioural patterns, this obligation should be extended to require that the system be designed in such a way as to inform the individual of the techniques deployed by the system to achieve their selected objective. Using the example of an app designed to influence a user to reduce alcohol intake, Jacobs outlines how such a design requirement might be implemented, suggesting a simple interface where the user is asked 'Do you consent to the use of the persuasive tool of self-monitoring, which consist of self-tracking the amount of alcohol you take in per day?' and 'Do you consent to the app sending you a maximum of five text messages throughout the day?'.[118] The designers of that system might inform users that the system will allocate rewards, rearrange stimuli or adopt patterns based on their behavioural patterns.

While we should be able to opt-in to persuasive technologies, Thaler and Sunstein have written about the scourge of 'sludge': 'friction which makes it harder for people to obtain an outcome that will make them better off'.[119] Adopters of machine learning technologies designed to distort their behaviour

[117] AI Act, art. 52(1).
[118] Jacobs (n 29) 525.
[119] Thaler and Sunstein (n 67) 153.

should have to surmount a small amount of sludge to opt-in, at least enough to ensure awareness and even invite critical reflection on their decision. Needless to say, that sludge should evaporate when adopters decide to cease using the technology.

Finally, the deployment of persuasive technologies should be limited to the same context and purpose for which they are adopted to prevent the use of such technologies becoming manipulative. Consider, again, the example of the app designed to reduce alcohol intake. Say a user consents to receive nudges in the form of push notifications from the app on their mobile phone. If the user starts to receive nudges in a different form and context, then the influence mechanisms deployed by the system are no longer transparent or apparent. The intent to influence becomes obscured, so the technology is manipulative rather than persuasive. The prevalence of 'agile programming practices', whereby digital service providers continuously swap-out, enhance, tweak, add or remove service features,[120] means that digital contexts are highly dynamic. The seamless integration of persuasive technologies with different apps, services and platforms, and the accumulation of data and insights across contexts, poses the risk of 'function creep' – 'the gradual widening of the use of a technology or system beyond the purpose for which it was originally intended'[121] – in this case, to achieve the objective selected by the subject. To guard against function creep which obscures intent to influence, any data about a user's behaviour, as well as inferences drawn about vulnerabilities in their decision-making, should be processed solely for the objective selected by the user – without the usual exceptions.

[120] Gürses and van Hoboken (n 42) 593.

[121] 'Function Creep, n.', Collins Dictionary (Collins 2022) <https://www.collinsdictionary.com/dictionary/english/function-creep> accessed 2 February 2022.

7 Conclusion

Understanding the complexities of human decision-making has long been,[122] and will undoubtedly remain, a core pursuit of AI scientists. Advances in our understanding of human decision-making will likely bring with it enhanced capabilities to manipulate those processes.

The flourishing of ethical, trustworthy, and beneficial AI will depend upon the implementation of well-defined, certain, and adapted constraints on harmful applications of machine learning capabilities, including manipulation. Legal responses to manipulative AI must therefore be adapted to address the specific features of AI-facilitated manipulation.

The approach proposed by the EC in the AI Act appears to be responsive to the core elements of manipulation – covert influence and the exploitation of vulnerabilities in decision-making. However, the approach is built upon concepts and devices which are incongruent with the dynamic, personalised, and persistent nature of manipulation enabled by machine learning. The proposed prohibition on AI systems which deploy subliminal techniques overlooks individual differences in processing sensory information and capabilities to induce or detect lapses in perception and attention. The group-level construction of vulnerability underpinning the second prohibition is ill-suited to the growing capabilities of machine learning models to covertly surface, induce and exploit individual vulnerabilities. Equally inadequate is the conception of the potential manipulee as an inherently rational, self-sufficient subject who only requires reasonable information to resist hidden, hyper-tailored influence attempts. In seeking to secure the benefits of AI while meeting the heightened threat of manipulation, lawmakers must eschew prevalent yet ill-fitting conceptions and

[122] See Allen Newell and Herbert A Simon, *Human Problem Solving* (Prentice-Hall 1972).

adopt new frameworks better suited to address the growing capabilities for manipulation arising from advancements in machine learning.