Manipulative Influence via AI Systems and the EU Proposal for Regulation of Artificial Intelligence

Dr. Jan Christoph Bublitz
Faculty of Law, Universität Hamburg
christoph.bublitz@uni-hamburg.de

Prof. Thomas Douglas
Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford
thomas.douglas@philosophy.ox.ac.uk

KEY POINTS

When AI systems influence thought or behaviour in ways that bypass or weaken rational control, they are manipulative.

These manipulative influences are ethically problematic, and can threaten fundamental rights, including the rights to freedom of thought, freedom of opinion, and mental integrity.

The proposed Regulation does not address central cases of manipulative influence via AI systems; it covers only the peripheral cases of subliminal interventions and interventions that exploit specific, vulnerable groups.

Nor is the problem of manipulative influence via AI systems adequately addressed via other instruments, such as GDPR and consumer rights directives.

The creation of trustworthy AI – a key objective of the Regulation – cannot succeed without a robust stance against manipulation

RECOMMENDATION

The Regulation should address manipulative interferences more directly and comprehensively, for example, by incorporating provisions designed to ensure that AI systems are classified as high risk whenever they significantly influence thought or behaviour in ways that bypass the rational control or significantly weaken rational control.

PROPOSED AMENDMENTS

Amendment 1

Art. 5. 1 (a) should be replaced with:

“AI system that deploys subliminal techniques beyond a person’s consciousness in order **to influence her thoughts or opinions,** or materially distort a person’s behaviour **or decisions.**”

Amendment 2

Art. 5. 1 (b) should be replaced with:

“AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to **influence thoughts, opinions or weaken the behavioural control of a person of that group, or to** materially distort the behaviour **or decisions** of a person pertaining to that group.”

Amendment 3

A further class of high-risk AI should be defined:

“AI systems that typically significantly alter thought (including beliefs, opinions, desires, emotions, decisions and the mental processes that lead to them) or behaviour in ways which bypass or significantly weaken an average person’s rational control are high-risk systems. There will be reason to think that a system bypasses an average person’s rational control if it is typically highly effective in altering thought or behaviour even though most people report resisting, or wishing they had resisted, the influence. There will be reason to think that a system undermines an average person’s rational control if it typically leads to people thinking or behaving in ways which they report they resisted, or wish that had resisted.”

I. Introduction

We wish to commend the complex, timely, and important undertaking of the Commission in developing the EU Proposal for Regulation of Artificial Intelligence (the "Regulation"). We appreciate large parts of the proposal. In this brief feedback, however, we draw on our expertise as a legal scholar and a philosopher respectively to highlight one area in which the Regulation fails to provide adequate protection: manipulative influences on thought and behaviour through AI systems.

Although the problem of manipulative influence has received less scholarly attention than other ethical, legal and social issues raised by AI systems, it is one of the main obstacles to the creation of an ecosystem of trustworthy AI governed by a fundamental rights framework. The Regulation acknowledges the problem posed by manipulative AI systems. However, it only regulates exceptional and peripheral cases while failing to capture many central cases, some of which caused grave public concern in recent years. These include microtargeted advertising as well as the abuse of trust in recommender systems.

We wish to draw attention to the fact that many manipulative influences via AI systems contravene and potentially violate the rights to freedom of thought and opinion (Articles 9 and 10 European Convention on Human Rights, ECHR, and Charter of Fundamental Rights and Freedoms, ECFR), and the right to mental integrity (Article 3.1. ECFR), and the right to private life and personal data (Article 8 ECHR, Articles 7 and 8 ECFR). The EU, as well as member states, are obliged to preserve and protect these rights. The Regulation could be crucial in ensuring fulfilment of this duty and should be oriented towards this end.

According to the Explanatory Memorandum to the Regulation a key objective of the Regulation is the creation of a comprehensive and future-proof "legal framework for trustworthy AI", as part of an "ecosystem of trust" based on "EU values and fundamental rights" that "aims to give people and other users the confidence to embrace AI-based solutions".[1] The Regulation for AI seeks to ensure that AI "is a force for good in society, with the ultimate aim of increasing human well-being".[2] Therefore, "rules for AI … affecting people in the Union should be human centric, so that people can trust that technology is used in a way that is safe and compliant with the law, including respect for fundamental rights".[3] These objectives can only be realized if the risk of manipulative influence via AI systems is substantially mitigated through the Regulation. Unfortunately, the current proposal falls short in this regard.

We therefore urge the Commission to address the risk of manipulative influence by adopting additional provisions safeguarding the abovementioned human rights. The problem should not be deferred to other existing or future instruments, as this would likely result in a

[1] Regulation, p. 3.

[2] *Ibid.*

[3] *Ibid.*

fragmented and incomplete regulatory landscape. The Regulation is the appropriate place to address dangers arising from AI technology. It is highly visible, attracting worldwide attention, and will shape the development of AI systems in the future, even beyond the EU. Without a clear and firm stance on manipulative influence, there is a large lacuna at its centre.

In the following, we identify areas of concern and submit suggested amendments. At the outset, we wish to acknowledge the great difficulties in formulating a legal framework for manipulative influence via AI systems. The modes in which AI systems can interact with persons are manifold and, with respect to future technologies, hard to foresee: from AI-recommender systems to robots in care or industry, from virtual assistants to virtual realities, from autonomous cars to targeted advertising. Even AI systems implanted in the human brain are currently under investigation.[4] In these novel forms of human-technology interaction, the roles of the AI and the person, as well as the nature and strength of potential influences exerted on the person's thought (including motivational states, affective states, opinions, beliefs, and decisions) and behaviour may vary greatly.

Moreover, the fundamental normative issues are largely unsettled. Regulating manipulative influences requires delineating permissible from impermissible influences. However, there is no single, widely endorsed philosophical account of manipulation.[5] Nor is there an established legal account of manipulative or otherwise undue influence that can be directly applied to influences through AI systems. The scope and limits of the implicated fundamental rights are equally unresolved. For example, there is significant unclarity concerning the scope of the rights to freedom of thought and mental integrity,[6] and almost no work applying it to AI-systems. Regulation thus requires some prior theoretical groundwork, and that groundwork will need to take into account the diversity influences exerted by AI-systems. There is also the problem that manipulative influence via AI systems may often be subtle and context-specific, such that it can only be captured by fine-grained rules, whereas regulations are by their nature abstract and general. This generates a tension between the generality of norms and the granularity of effects. Finally, as the psychological effects of

[4] See Surjo Soekadar, Jennifer Chandler, Marcello Ienca, Christoph Bublitz, 'On the Verge of Hybrid Minds', *Morals & Machines* 2021, 1: 31-42.

[5] See Christian Coons and Michael Weber (eds.), *Manipulation: Theory and Practice*. (Oxford University Press, 2014).

[6] See, Jan Christoph Bublitz's 'The Nascent Right to Psychological Integrity and Mental Self-Determination', in v. Arnauld, v. d. Decken and Susi (eds.), *The Cambridge Handbook of New Human Rights. Recognition, Novelty, Rhetoric* (Cambridge University Press, 2020), 387 – 403; and 'Freedom of Thought in the Age of Neuroscience: A Plea and a Proposal for the Renaissance of a Forgotten Fundamental Right', *Archiv für Rechts – und Sozialphilosophie* 2014, 100:1, 1–25; Susie Alegre, 'Rethinking Freedom of Thought for the 21st Century'. *European Human Rights Law Review* 2017, no. 3: 221–33.

many new forms of influence are poorly understood, their strength and modes of operation are not always easy to determine.

Despite these challenges, however, the regulation of manipulative influence via AI systems (henceforth sometimes 'manipulative AI influence') is urgently needed. Its absence would leave central aspects of the human person unprotected and fail to respect some of the most important fundamental rights.

However, some regulatory challenges might be turned into benefits. Because the lines between manipulative and non-manipulative influence are notoriously hard to draw in both law and ethics, there is a large grey area of normative uncertainty. Regulators should regard this as a space of opportunity. The unchartered territory leaves ample leeway for drawing the contours of a reasonable regulation respecting and realizing fundamental rights.

II. Manipulative AI Influence

The definition of AI in the Regulation is broad and includes machine learning methods, such as big data analysis, as well as search and optimization methods.[7] Systems based on these methods may influence persons in many ways, depending on the precise mode of interaction. Abstractly speaking, AI can optimize stimuli which persons receive ("interventions") and render those stimuli more effective in altering mental states or behaviour by dynamically adapting them to the current psychological state of recipients, to psychological weaknesses and vulnerabilities, as well as to situational factors. To do so, it can draw on data-sets from the general population, specific groups, or affected persons themselves. AI systems yield predictions about the effects of interventions on persons as well as information about how to optimize interventions with respect to their content, form, time and mode of presentation, and other factors. This can generate powerful—and manipulative—forms of influence.

As noted above, there is significant unclarity regarding when an influence counts as manipulative, and when it does not. In what follows, we will employ an account inspired by influential philosophical work on this topic.[8] We take it that an influence is manipulative—and thus ethically and legally problematic—if it either (a) weakens or entirely undermines a person's rational control over her thought or behaviour (manipulative in effect), or (b) influences a person's thought or behaviour via means that bypass—that is, do not engage—the person's rational control (manipulative in means).

We take rational control to be the capacity to decide what to think or do on the basis assessing the arguments, reasons and evidence for and against the thought or action, and to implement these decisions in one's actions. Rational control can be exerted through conscious

[7] Annex I.

[8] See Moti Gorin's 'Do Manipulators Always Threaten Rationality?' *American Philosophical Quarterly* 2014, 51(1): 51–61, and 'Towards a Theory of Interpersonal Manipulation', in Manipulation: Theory and Practice, edited by Christian Coons and Michael Weber (Oxford University Press, 2014).

deliberation, but also through automatic and subconscious processes, such as following one's intuition. We use 'thought' very broadly, to refer to all mental states, whether cognitive (e.g. beliefs and opinions), motivational (e.g. desires, preferences and intentions) or affective (e.g. emotions and feelings), to the mental processes via which they are produced (e.g. reasoning and perception), and to mental events (e.g. decisions).

On this view, AI influences are frequently manipulative and this, we will later argue, brings them squarely within the scope of the fundamental rights to freedom of thought, freedom of opinion and mental integrity. We wish to exemplify this by AI-based optimization of common and established forms of interaction. If these are manipulative, some future modes of interaction will likely be manipulative too.

Consider first influence that weakens people's control over their *behaviour*. This not only includes rare cases in which people fully lose control of their bodily movements (disorientation, hypnosis, shock, etc), but also influences with weaker but still worrisome effects. A phenomenon familiar to everyone is technological captivation of attention that keeps people from doing what they prefer to do.[9] For example, optimisation of social media sites and games may keep people playing, or online, for longer than they would, on reflection, prefer. In theoretical terms, one may distinguish first order volitions (e.g., the desire to get to keep playing a game until one passes the current level) and second order volitions (e.g. the desire that this desire to keep playing be weaker or less effective).[10]

Trade-offs between first- and second-order volitions are a common aspect of life. They become problematic when first-order volitions are so strong and entrenched that they cannot easily be altered or overridden by higher-order volitions. Many digital technologies appeal to and strengthen and entrench first order volitions in in. way that brings about just this result. They create temptations which people struggle to resist. And if people fail to resist because it is difficult for them to do so, control over their behaviour has been compromised.

AI systems are especially likely to weaken rational control over our behaviour as they can dynamically adapt the type, form, intensity, and patterns of stimuli in real time, in response to behaviour of persons as well as their inferred mental state (inferred, for example, though the level of engagement a person shows, her reactions to other stimuli, or reaction times). This allows AI systems to harness weaknesses in people's behavioural control, with the result that even individuals with strong self-control struggle to resist the behaviours that an influence is designed to promote.

Consider second weakening of rational control over *decisions*. AI-systems can influence decisions by a variety of means. Especially worrisome are those which avail themselves of the many human frailties and biases, for example, excessive susceptibility to priming effects

[9] See, for an in-depth discussion, James Williams, *Stand Out of Our Light: Freedom and Resistance in the Attention Economy* (Cambridge University Press, 2018).

[10] Harry Frankfurt, Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68, 1971, 5-20.

and emotionally charged stimuli. Adversarial machine learning – one of the most powerful AI methods – has been experimentally deployed to identify and target such vulnerabilities. To this end, AI builds models of human decisionmaking and its vulnerabilities, and then explores strategies to exploit such frailties with this model.[11] Such methods may further increase the effectiveness of AI-influences.

Another area of worrisome influence of AI systems on decisions are recommender systems, widely employed in online shopping and booking sites. These can influence decisions in several ways. One is by ordering of the presentation of options; most people will choose one of the top few options on a list. Another is by cultivating the gradual development of trust in and reliance on recommendations through repeated interactions. While recommender systems may regularly present the most suitable options for the person, they may, every once in a while, include options which were not selected on the basis of suitability but are instead, for example, paid promotions or are presented for other reasons. If users have come to trust and rely on the system, they may find it difficult to call those recommendations into question in deciding which option to take. The exploitation of habitual reliance in such cases is plausibly an instance of weakening rational control over decisions.

Until this point, we have been considering ways in which AI systems might reduce control, over either decisions or behaviour. (In many cases, both will be affected.) However, it is possible to imagine cases in which AI influences affect behaviours, decisions or other aspects of our mental life significantly, but without diminishing our control over them. Perhaps, for example, the influence affects emotional states over which we in any case lack rational control.

In these cases, the influence may be problematic because it *bypasses* rational control. Here, the problem is not that the influence produces behaviours, decisions, or other thoughts over which we lack rational control. Rather the problem is that we lack control over the influence itself. For example, a person may lack control over whether her emotion was strengthened or weakened by the influence.

Consider, for instance, a case in which an AI system, detects that a social media user has become frustrated, and thus prone to provocation, it may present a stimulus intended to induce hatred towards immigrants. Even if the feeling of hatred is not sufficiently strong to count as an impediment to rational control, this influence is problematic because, in targeting the person while frustrated, the influence may have bypassed the person's capacity to control her the formation of the hateful feelings.

[11] See Dezfouli/Nock/Dayan, 'Adversarial Vulnerabilities of Human Decision-Making', *Proceedings of the National Academy of Sciences* 117, 46 (2020): 29221–29228.

As a real-life example of this, it is reported that Cambridge Analytica was able to target inflammatory political messages to individuals "prone to impulsive anger or conspiratorial thinking".[12]

We have highlighted various ways in which influences exerted via AI systems may weaken or bypass rational control. These effects are not unique to AI, however the microtargeted nature of many AI influences increases the risk that rational control will be weakened or bypassed. It does this for two reasons.

First, since they allow a greater degree of individualization (i.e. finer grained targeting) than more traditional forms of targeting, microtargeting systems are more likely to harness *idiosyncratic* deficiencies in rational control. As an example, suppose that there are 100 strategies that a social media platform can employ to keep a person engaged on the platform. Suppose further that a typical person is afflicted by defects of rational control such that, for 5 of these strategies, they will be unable to rationally resist the influence; employing any of those five strategies will be manipulative since it bypasses rational control. It is also likely to be highly effective,[13] and thus desirable from the perspective of the social media platform. However, since different people exhibit different defects in rational control, they are vulnerable to different strategies of influence. Suppose that these defects in rational control are evenly distributed across people, so that any one influence can be expected to bypass rational control only for 5% of people.

In the absence of targeting, it is unlikely (5% chance) that the any given individual will be presented with an influence that bypasses her rational control. However, the more finely the platform can target its influences, the more likely it is to hit on the particular influences that evade rational control in a particular person.

Second, since AI-systems automate the targeting process, they allow individualized influences to be exerted on a larger number of people than would otherwise be possible; they enable *individualized mass communication*. As a result, microtargeting may greatly increase the number of targeted influences that can be addressed to a single person.

III. Current and Proposed Regulation

Influence exerted via AI systems is particularly likely to weaken or bypass rational control—and thus be manipulative—because stimuli can be finely tailored to the person and situation and are often repetitive.

[12] Christopher Wylie, *Mindf*ck: Inside Cambridge Analytica's Plot to Break America* (Random House, 2019), p. 120.

[13] S. C. Matz et al., 'Psychological targeting as an effective approach to digital mass persuasion', *Proceedings of the National Academy of Sciences* 114, 48 (2017): 12714–9.

Unfortunately, the provisions of the proposed Regulation are largely silent on manipulative influence, although the proposal purports to be animated by concerns about manipulation. Recital 15 notes that AI "can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices. Such practices are particularly harmful and should be prohibited because they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law and Union fundamental rights". Nevertheless, the Regulation only robustly regulates two specific and exceptional forms of manipulative influence—those that employ subliminal stimuli, and those that exploit the vulnerabilities of especially vulnerable groups.

Subliminal Stimuli

Article 5.1 (a) of the Regulation prohibits AI systems deploying "subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm". Subliminal techniques are usually understood to be interventions that recipients cannot perceive, for example because they are presented so quickly that they do not rise to conscious awareness. The Regulation does not define the term, but Recital 16 suggests that it uses it in the same way. Research shows that such stimuli can have very limited effects. It is right to ban them. Nonetheless, subliminal techniques were the fear of the mid 20th century and it is neither clear that they have ever been put to alarming use, nor that there are AI-related applications drawing on them. This prohibition, we suggest, is largely symbolic.

Moreover, one may wonder why the Regulation only prohibits subliminal stimuli that lead physical or psychological harm. As both terms remain undefined, they are to be understood as they usually are in law, as physical or psychological injuries, setbacks to health or biological or social functioning. Yet nonconsensual subliminal interventions that significantly alter thought or behaviour are plausibly ethically wrong—because they bypass rational control—even if they inflict no harm. As ethically acceptable uses of nonconsensual subliminal interventions are hard to conceive, they should be banned in-principle, with only tightly defined exceptions (perhaps, for example, allowing uses in specific forms of scientific research). (This motivates to Amendment 1, below.)

Vulnerable Groups

Article 5.1 (b) of the Regulation prohibits AI systems exploiting "any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour 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". Again, the restriction to physical and psychological harm does not seem warranted. Vulnerable people should also be protected against significant influences that bypass, weaken or undermine rational control, regardless of whether those influences can be expected to cause harm (see suggested Amendment 2, below).

A further question posed by this Article is why only specific group vulnerabilities should be relevant, rather than, for example, common vulnerabilities of human psychology. (This is, of course, not to deny that members of particularly vulnerable groups deserve special protection.)

Transparency for Emotion Recognition

Furthermore, Article 52.2 stipulates that persons exposed to an emotion recognition system should be informed about the operation of the system (see also Recital 70). “Emotion recognition system” is defined as a “an AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data” (Article 3.34). Biometric data stems from “physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person” (Article 3.33). Supposedly, this means that all physical, physiological or behavioural characteristics that *might* potentially be used to identify persons are relevant data regarding Art. 3.34.

Some manipulative AI influences will fall under this Article. If AI systems infer a person’s emotional states or intentions on the basis of biometric data, the person will need to be informed about the use of the system. However, this does not suffice to alleviate worries regarding manipulative influence since they need not be informed by inferences regarding emotional states or intentions, or indeed be based on biometric data. Consider, for example, the cultivation and exploitation of trust in and reliance on recommender systems mentioned above.

Beyond these provisions, the Regulation includes no substantial restrictions on manipulative influence via AI systems. Overall, then, while it prohibits some exceptional forms of influence, it leaves the central cases of manipulative AI influence unaddressed. Some influences are apparently so worrisome that they should be outright prohibited whereas others do not pose any problem at all. Such a strict binary approach to a complex problem is not optimal; a more nuanced approach would be preferable.

Other Instruments

One of the reasons for the absence of further provisions curbing manipulation may be the assumption that other instruments sufficiently address the issue. The Explanatory Memorandum asserts 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” (at 5.2.2).

However, this betrays a misunderstanding of the scope and aims of these instruments and the threats posed by AI systems. Of course, every concrete AI application needs to be legally examined individually, and much depends on technicalities. Nonetheless, none of the other instruments contain more specific or concrete rules that would capture the central cases of manipulative influence via AI systems. Therefore, specific rules about manipulation need to be developed in any case, and the Regulation would be an appropriate place for them.

1. GDPR

The EU Proposal confers the impression that the General Data Protection Regulation (GDPR) is sufficient to prevent manipulative AI influences. This is misleading. Breaches of data privacy and manipulative influence are interwoven, but distinct, types of wrong. From an ethical point of view, regulating the former does not render protection against the later superfluous. This is true also in the law—whereas invasion of privacy and manipulative influence may both infringe Article 8 ECHR, manipulative influence may also infringe Articles 9 and 10 ECHR. The GDPR primarily addresses invasion of privacy. The objective of the GDPR is the protection of personal data (Article 1.1. GDPR). A solid protection of personal data may deprive some AI systems of their source material, but it does so only contingently and will not cover all central instances of manipulative influence via AI systems.

For instance, it is not clear that the example cases discussed above necessarily rely on or make use of personal data. In some cases, data that do not allow identification of the particular person (e.g. "people who buy baby trolleys also buy Lego") may suffice. Indeed, a recent study of over 3 million facebook users found targeted advertisements were substantially more effective than untargeted advertisements (50% more effective in some cases) even when the targeting was based solely on (often non-personal) 'likes'.[14]

In other cases, data subjects may have consented to the processing of their data, as this is a necessary condition for the use of many digital services. Consent to the data processing may factually enable manipulative influences and ensure compliance with GDPR without constituting consent to those influences. Also, as long as the data is not sensitive pursuant to Article 9 GDPR (for example, a political opinion) there might be legitimate ways for data controllers to use them. In addition, data might be acquired through breaches of data laws. It would be naïve to assume that such breaches will not happen. For these reasons: Regulating data is not a sufficient safeguard against manipulative influences. They must be addressed in their own right.

2. Digital Service Regulation

The Digital Service Regulation, also published recently, contains reference to the problem of recommender systems that "play an important role in the amplification of certain messages… and the stimulation of online behavior" (Recital 62). It also mentions threats to "civil discourse" and "political participation" posed by them (Recital 63). To address these threats, it stipulates that advertisements be marked as such and that digital service providers must disclose the identity of the advertiser and inform users why they were targeted (Art. 24). Additionally, the AI system should give users several options for recommendations, at least

[14] S. C. Matz et al., 'Psychological Targeting as an Effective Approach to Digital Mass Persuasion', *Proceedings of the National Academy of Sciences* 114, 48 (2017): 12714–9.

one of which should provide recommendations without basing them on personal data (Art. 29). Moreover, a publicly available repository for advertisements must be created that shows who was targeted with which advertisement (Art. 30).

We support these measures, but they do not suffice to solve the previously mentioned problems of manipulative influence via AI systems. The option of turning off personal data use is laudable, but it does not advance the main aim of the Regulation, namely the creation of trustworthy AI system. A recommender system will be trustworthy only if it reliably guides the user towards options that advance the users' ends. A system that exploits habituated reliance on the system to promote options that were not selected on the basis of their fit with the users ends--but instead those of others who have paid to promote them--might be trusted, but is not trustworthy. The Regulation should thus strive to ensure that recommender systems do not employ manipulative influences of this kind.

3. Consumer Directive and Unfair Commercial Practices Directive

The Regulation mentions consumer protection such as the Consumer Directive, recently amended to cover digital services.[15] However, in general, consumer protection applies only to people in the role of a consumer, in their relation to a business ("trader"; see, for example, Article 3.1. Consumer Directive). Yet manipulative AI influences also occur in other contexts, for example, in business-to-business and person-to-person interactions. Moreover, political microtargeting may affect people in their role as citizens, not consumers. Such interactions are not covered by the Directive. Furthermore, the Consumer Directive does not include substantive provisions protecting consumers against manipulative influences.

The Unfair Commercial Practices Directive (UCPD)[16] prohibits, among other conduct, any commercial practice that "materially distorts or is likely to materially distort the economic behaviour with regard to the product of the average consumer whom it reaches or to whom it is addressed", Article 5.2 (b) UCPD.

This may complement the prohibitions of the Regulation. However, it would, for two reasons, be a mistake to assume that manipulative influences are sufficiently covered by the UCPD.

First, as with the Consumer Directive, the UCPD applies only to business-to-consumer "commercial transactions" (Article 3).

Second, the range of influences covered by Article Article 5.2 (b) is narrow. Article 2 (e) explains that "to materially distort the economic behaviour of consumers' means using a commercial practice to appreciably impair the consumer's ability to make an informed decision, thereby causing the consumer to take a transactional decision that he would not have taken otherwise". Thus, to fall under the prohibition contained in Article 5.2 (b), a practice has to impair the ability to make an informed decision. This is a demanding

[15] Directive 2011/83/EU (25 October 2011), as amended by Directives 2015/2302 (25 November 2015), 2019/2161 (27 November 2019).

[16] Directive 2005/29 EC (11 May 2005), as amended by the Directive 2019/2161 (18 December 2019).

standard, and one that would not necessarily be met by influences that are problematic not because they undermine or weaken rational control, but because they operate via means that bypass it. Given the absence of further clarifications regarding what qualifies as an informed decision, the standard is also imprecise.

Even with respect to subliminal stimuli, the standard imposed by the UCPD may not be met. Consider one of the historical cases associated with subliminal stimuli, the supposed increase of ice-cream sales in a cinema that flashed subliminal ice-cream advertisements at the audience.[17] Suppose a person watches a movie and now has the desire for an ice-cream. At the vendor, she considers the varieties available and decides for a medium-sized strawberry ice-cream which has a good price and contains not too much sugar. Has this customer made an informed decision? It seems so. Controlling the source of a given desire is not a necessary for the ability to make an informed decision. Plausibly, the customer in this case has made an informed, but manipulated, choice.

Thus, the text of Article 5.2 (b) UCPD does not clearly cover influences that are problematic because of the means that they employ, not their effects. It therefore cannot be assumed that it covers all worrisome AI influences. Of course, one would have to look into the case-law to evaluate concrete scenarios, and courts could surely find ways to apply Article 5.2 (b) UCPD to some of them. But this cannot form the basis for a comprehensive and future proof AI regulation. This is rather courts filling gaps arising from incomplete acts of legislation. Without doubt, courts will play an important role in rendering the Regulation more concrete and filling it with life. It is the genuine task of the regulator to develop the central normative contours of the Regulation.

Summary

Contrary to the claim made in the EU Proposal, the instruments surveyed in this section do not “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”, since they cover only influences employed within certain contexts (e.g. business-to-consumer interactions) and that are problematic in certain ways (e.g., because they inhibit informed decision-making). Referring—and deferring—this important issue to other instruments would result in a fragmented and incomplete regulatory landscape. The EU should develop a more systematic and comprehensive framework that addresses all manipulative influences exerted via AI systems.

An example of the regulatory problems that can arise without a firm guidance is provided by so-called ‘loot boxes’ in games. These are randomised awards that gamers can buy in the context of a gaming or a virtual environment. Psychologists warn that loot boxes can cause behaviour that resembles gambling addiction.[18] This suggests that they can undermine or

[17] Vincent Packard, *Hidden Persuaders* (Longmans, Green & Co, 1958).

[18] David Zendie, Rachel Meyer, Harriet Over, ‘Adolescents and loot boxes: links with problem gambling and motivations for purchase‘. *Royal Society Open Science*, 6 (2019), 190049.

weaken rational control by strengthening addictive desires. For a variety of reasons, however, loot boxes are not regulated at EU level, and are regulated at national level only in a few member states. This shows how the lack of systematic and comprehensive regulation of manipulative influences can result in regulatory failures even in relatively clear cases of manipulative influence.[19]

IV. A Fundamental Rights Approach

The European Convention and the European Charter enumerate the fundamental rights which bind and oblige the EU and member states. They must form the foundation for a legal framework for AI. A comprehensive fundamental rights approach to manipulative influence via AI systems would lay out the potentially implicated rights and would interpret them to render them applicable to the domain of regulation. It is notable that the Regulation does not fail to declare its commitment to fundamental rights; it mentions the term 80 times, in addition to mentioning particular fundamental rights. However, the ubiquity of references to fundamental rights is not matched by robust protections for them, and indeed the references often remain vague and generic. More often than not, the Regulation does not clearly state which rights are supposedly implicated by which measure and what this implies more precisely. Only a concrete and comprehensive assessment of implicated rights, their strengths and limits, as well as their relation to competing rights and countervailing considerations can provide the foundation for a Regulation living up to the obligation to preserve and protect these rights.

Admittedly, making such an impact assessment is challenging since many fundamental rights have not yet been rendered more precise with respect to novel threats posed by AI systems. But this is a task that anticipatory regulations must confront. Indeed, by drawing the contours of fundamental rights in grey areas pertaining to AI, regulators could advance the understanding of those rights.

Implicated Rights

We suggest that, with respect to manipulative influences, AI systems potentially interfere with the following fundamental rights:

1) The right to freedom of thought (Articles 9 ECHR, 10 ECFR). This right protects thought against interferences. The Regulation does not mention it once.

[19] See Anette Cerulli-Harms et al., 'Loot Boxes in Online Games and their Effects on Consumers, Particularly Young Consumers: Publication for the Committee on the Internal Market and Consumer Protection, Policy Department for Economic, Scientific and Quality of Life Policies (European Parliament: Luxembourg 2020), available at https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652727/IPOL_STU(2020)652727_EN.pdf (accessed 6 Aug, 2021).

2) The right to hold opinions without interferences, protected as part of freedom of expression in Articles 10.1. ECHR and 11.1 ECFR. The Regulation refers to freedom of expression four times, but chiefly with respect to the expressive dimension, not the protection of opinion against interference.

3) The right to mental integrity (accepted as part of Article 8 ECHR and Article 3.1. ECFR). It is not mentioned in the Regulation. The right to mental integrity is not well understood. But, in lose analogy to the right to bodily integrity, it captures adverse interferences with the mind of some gravity.

4) Lastly, the right to privacy or private life, which the Regulation mentions numerous times.

In our view, the Regulation is not sufficiently sensitive to the specific demands of the first three rights. Freedom of thought protects against interference with 'thoughts'. In addition, freedom of opinion protects subjective value judgments. The scope of these terms, including whether they comprise motivational states (such as desires and preferences), affective states (such as emotions and feelings), and beliefs that concern trivial matters, is not settled in the law. But the right plausibly captures many of the previously outlined manipulative influences via AI systems. Moreover, mental states not covered by freedom of thought and opinion fall under the (non-absolute) right to metal integrity.

We wish to draw attention to the fact that the internal sides of the freedoms of thought and opinion are considered absolute rights in international human rights law, i.e., infringements of them cannot be justified (Articles 18 and 19 Universal Declaration of Human Rights and Covenant on Civil and Political Rights). The same is true for Article 9 ECHR. As they are among the most important fundamental rights, they command heightened respect. Their absence from the Regulation indicates that they have not been duly taken into consideration.

Infringements

The EU and member states have the positive obligation to protect rightholders from infringements of these rights by third-parties.

In our view, significant manipulative influences—that is, influences which typically significantly influence thought or behaviour through means that bypass rational control, or in ways that significantly weaken rational control—constitute *interferences* with thought and thereby raise considerable worries with respect to freedom of thought – or at least with mental integrity. Though there is scope for disagreement regarding which *influences* on the mind constitutes *interferences* with it, it is very plausible that significant manipulative influences do qualify as interferences and thus—unless done with valid consent—infringe the right to mental integrity.

The Regulation arguably acknowledges this reasoning, or something close to it, at least insofar as it pertains to bypassing rational control. Subliminal stimuli are worrisome precisely because they bypass rational control. Stimuli are perceived and processed, but do not come to conscious awareness, and therefore, it is typically assumed, they are not subjected to the usual range of psychological checks and balances. And as they bypass rational control, the Regulation – correctly – bans them.

The same criterion–bypassing rational control – should also apply to significant influences that are supraliminal. Psychological research suggests that a range of methods bypass rational control, often by exploiting known psychological weaknesses in control. For instance, stimuli might be designed so that they are processed only peripherally (shallowly, not critically). Another strategy is to time stimuli so that they affect people in conditions in which their rational control is diminished (for example, because they are fatigued, sad, hungry, cheerful, or frustrated).

Regulating significant influences that bypass control would thus be a natural extension of principles that are arguably already implicit in the Regulation. Regulating significant influences that undermine or significantly weaken control, would be a further step, but one that is in-line with regulations in other areas. For example, regulations pertaining to gambling and addictive substances can plausibly be understood as intended to prevent the loss of rational control associated with the formation of strong addictive desires.

V. Regulatory Proposal

This analysis provides rough guidance for setting the contours of a Regulation that would protect rights to freedom of thought, freedom of opinion, and mental integrity. Thorny questions of interpretation will, of course, arise. For example, there is room for disagreement regarding what counts as a significant influence and a significant weakening of rational control. There is also room for fundamental disagreement regarding what counts as bypassing or weakening rational control. And pragmatic challenges, such as the need to ensure that regulations are enforceable and do not stifle beneficial technological progress, will also arise. But these problems are not excuses for failing to regulate such influences.

Rather, we suggest making use of the flexible, risk-based approach of the Regulation. Instead of banning AI systems that influence thought or behaviour, they should be classified as high-risk systems, which can then be subjected to a range of further, more specific rules—many of which will need to be developed in response to particular technological developments. This, of course, does not entail that all such systems are permissible. Some might not and should be banned, based on an assessments of their risks to fundamental rights. In addition, the most important rules that should apply to high-risk systems are (a) transparent and comprehensive information for people exposed to them about the nature and effects of these influences, and (b) a requirement that consent be obtained before people are exposed to them. Further duties of documentation and duties to conduct assessments of risks to the rights to freedom of thought, freedom of opinion and mental integrity are suggested.

Accordingly, we propose the Regulation should adopt a new category of high-risk AI systems (suggested Amendment 3, below):

This tentative formulation will no doubt need refinements, but it captures some key concerns. For instance, an AI system drawing on behavioral or emotional data to ensure vigilance in the operation of a car seeks—and if well designed is likely—to preserve rational control over one's driving behaviour; whereas a system drawing on the same information to keep people engaged on a social media platform that they, on reflection, would like to spend less time on, may reduce rational control. An AI system that identifies and exploits flaws and biases in reasoning, for example, to increase the desire to buy a product, is likely to bypass the capacity for rational control, whereas an AI system that instructs a Socratic bot to present arguments and evidence in a balanced way, plausibly, does not.

An advantage of classifying significant manipulative influences as 'high risk', rather than simply prohibiting outright, is that this allows for a nuanced regulatory approach that is also responsive to the interests and moral and legal rights of those who place AI systems on the market or make use of them in their services or products. Some such uses might, for instance, be protected by the right to freedom of expression. Given the diversity of potential manipulative AI influences, balancing the rights and interests of influencers with those of influencee's will, initially at least, need to be done on a case-by-case basis.

An amendment along these lines is necessary for the Regulation to observe the demands of fundamental rights, but also to realize the objective of developing trustworthy AI-systems which are taken up by citizens. While it may lead to impediments in the dynamic development of AI systems, it will primarily affect applications which in any case lie in an ethical and legal grey area. Their elimination is in the medium-term interest not only of the EU and its citizens, but also of industries and software developers wishing to create trustworthy AI in competitive markets.

Suggested Amendments

Amendment 1

Art. 5. 1 (a) should be replaced with:

"AI system that deploys subliminal techniques beyond a person's consciousness in order **to influence her thoughts or opinions,** or materially distort a person's behaviour **or decisions.**"

Amendment 2

Art. 5. 1 (b) should be replaced with:

"AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to **influence thoughts, opinions or weaken the**

behavioural control of a person of that group, or to materially distort the behaviour **or decisions** of a person pertaining to that group."

Amendment 3

A further class of high-risk AI should be defined:

"AI systems that typically significantly alter thought (including beliefs, opinions, desires, emotions, decisions and the mental processes that lead to them) or behaviour in ways which bypass or significantly weaken an average person's rational control are high-risk systems. There will be reason to think that a system bypasses an average person's rational control if it is typically highly effective in altering thought or behaviour even though most people report resisting, or wishing they had resisted, the influence. There will be reason to think that a system undermines an average person's rational control if it typically leads to people thinking or behaving in ways which they report they resisted, or wish that had resisted."