Psychological Patterns and Article 5 of the AI Act:

AI-Powered Deceptive Design in the System Architecture and the User Interface

*Mark Leiser**

The article emphasises the urgency of addressing the risks posed by AI-powered deceptive design strategies intricately woven into online platforms. These 'psychological patterns' mislead users into making decisions contrary to their intentions, exploiting psychological vulnerabilities. The article also critically examines the complex interplay between AI-powered deceptive design and legislative responses, mainly focusing on Article 5 of the European Union's AI Act. It underscores the importance of safeguarding user autonomy in the rapidly evolving digital landscape. Thus, the article discusses the dynamics of psychological manipulation, the need for effective regulation via Article 5, and the critical importance of maintaining user autonomy amidst technological advances in AI.

Keywords: dark patterns, deceptive design, psychological manipulation, artificial intelligence, regulation, harm

I. Introduction

In an era characterised by the pervasive influence of digital platforms, the 'dark patterns' phenomenon has become a focal point of academic and legislative concern.[1] The term has transcended its original domain within user experience (UX) critique to become a rigorous debate within legal forums and policymaking institutions.[2] Although varied in form and function, these manipulative digital design strategies share the common goal of subverting user autonomy for the benefit of the platform or application, leading to a profound ethical dilemma within the digital economy. In technological advancements and artificial intelligence (AI), the emergence of deceptive design tactics is often overshadowed by the promise of innovation. Thus, the potential of AI has a nefarious aspect, mainly when intertwined with other forms of deceptive design strategies.

As the discourse of this article unfolds, it will illuminate the spectrum of deceptive design, from the overt to the insidious, and the emerging legislative responses to combat them. Section III acknowledges the legislative implications of dark patterns, the evolving nature of user experience manipulation, and the critical role of policy development in safeguarding user autonomy where digital interfaces are ubiquitous. This analysis is supported by contributions to policy development in the European Commission team responsible for implementing the Unfair Commercial

[*] Dr M.R. Leiser (Mark), Amsterdam Law and Technology Institute, VU-Amsterdam, The Netherlands. For correspondence: <m.r.leiser@vu.nl>.

[1] OECD, 'Dark commercial patterns' (OECD Digital Economy Papers, No. 336, OECD Publishing, 2022) <https://doi.org/10.1787/44f5e846-en>; Harry Brignull, 'Dark Patterns: Deception vs. Honesty in UI Design' (A List Apart, 1 November 2011) <https://alistapart.com/article/dark-patterns-deception-vs.-honesty-in-ui-design/> accessed 14 July 2023; See also H Brignull, 'Dark patterns: User interfaces designed to trick people' (UX Brighton Conference, Brighton, UK, 2010) <https://www.slideshare.net/harrybr/ux-brighton-dark-patterns> accessed 26 July 2023.

[2] European Data Protection Board, 'EDPB Adopts Guidelines on Art. 60 GDPR, Guidelines on Dark Patterns in Social Media Platform Interfaces and a Toolbox on Essential Data Protection Safeguards for Enforcement Cooperation between EEA and Third Country SAs' (EDPB, 15 March 2022) <https://edpb.europa.eu/news/news/2022/edpb-adopts-guidelines-art-60-gdpr-guidelines-dark-patterns-social-media-platform_en> accessed 14 July 2023; European Data Protection Board, 'Guidelines 02/2022 on the

Practices Directive[3] and collaboration with design ethicists[4], including Harry Brignull, the design ethicist who coined the term 'dark patterns'. It underscores the transition to more deeply embedded 'psychological patterns' in deceptive design. The ensuing discussion will navigate the legislative nuances of the AI Act[5], advocating for a future where user rights are not undermined by the digital tools that should empower them.

Section II explains the concept of 'dark patterns' and their evolution into 'deceptive design', highlighting the ethical and autonomy issues in the digital economy. Section II then transitions into a detailed exploration of how deceptive design has evolved from overt manipulative strategies to more sophisticated, AI-driven psychological patterns, emphasising the legislative challenges these pose, particularly under the EU's AI Act. This section delves deeper into psychological patterns, examining the subtle and complex ways digital platforms manipulate user experience and system architecture.

A critical examination follows this in Section III.2 on Article 25 of the Digital Services Act (DSA), particularly its limited capacity to regulate deceptive design practices and the inherent limitations of this regulatory approach. Shifting focus to the EU's approach for regulating artificial intelligence, Section IV discusses the AI Act in depth, focussing on its provisions, especially Article 5[6], and the ongoing debate around implementing AI technologies. This leads to a recommendations section for best interpreting Article 5 of the AI Act. These suggestions include advocating for precise definitions and broader protections to ensure more effective regulation.

The article concludes by summarising the main points raised throughout the discussion. It calls for more robust legislative measures to effectively address the challenges posed by AI and deceptive design, highlighting the need for more robust safeguards to protect user autonomy in the digital landscape. This comprehensive narrative critically analyses the intersection of AI technology, deceptive design, and legislative responses, highlighting the complexities and nuances in this rapidly evolving field.

II. From Dark Patterns to Deceptive Design

Dark patterns represent a contentious subject within user experience and interface design. They are broadly understood as deceptive design techniques utilised in digital platforms, such as websites and applications, that entice users into actions unintended by the user.[7] These actions can include making purchases, subscribing to services, or inadvertently providing consent for data collection.[8] In an era characterised by the ubiquitous influence of digital platforms, dark patterns have become a focal point of academic and legislative concern. Although varied in form and function, these manipulative digital design strategies share the common goal of subverting user autonomy for the benefit of the platform or application, leading to a profound ethical dilemma within the digital economy.[9]

In 'The Case for Regulatory Pluralism', I postulated that the term 'dark patterns' has competing uses.[10] First, it is a general term of disapprobation for any undesirable website design attributes that users may find objectionable or frustrating. Second, and more specifically, it denotes a coercive and manipulative design strategy employed by web designers to elicit a desired action from a user, often benefitting the service provider to the detriment of the user. This en-

Application of Article 60 GDPR' (EDPB, 14 March 2022) <https://edpb.europa.eu/system/files/2022-03/guidelines_202202_on_the_application_of_article_60_gdpr_en.pdf> accessed 14 July 2023.

[3] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'), OJ L 149/22.

[4] Directorate-General for Justice and Consumers (European Commission), Behavioural study on unfair commercial practices in the digital environment: Dark patterns and manipulative personalisation: final report (Publications Office of the EU 2022) DOI 10.2838/859030.

[5] 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.

[6] For this article, I have used the EU AI Act Draft Agreement leaked by Laura Caroli and Luca Bertuzzi. A copy of this version can be found online at <https://clairk.digitalpolicyalert.org/agreement/279/> accessed 7 February 2024.

[7] OECD (n 1).

[8] EDPB Guidelines (n 2) 5.

[9] CM Gray et al, 'End User Accounts of Dark Patterns as Felt Manipulation' (2021) 5(2) Proceedings of the ACM on Human-Computer Interaction 1-25.

[10] M.R. Leiser, '"Dark patterns": The case for regulatory pluralism between the European Union's consumer and data protection regimes' in E Kosta, R Leenes and I Kamara (eds), *Research Handbook on EU Data Protection Law* (Edward Elgar Publishing 2022) 240-269.

compasses a range of tactics, from hidden charges to misleading navigation and preselected options that do not align with user autonomy. An example of such practices can be observed in Ryanair's website's interface design, which includes pre-selected options for additional purchases and services the user may not intend to select.[11] This could lead to financial commitments not being part of the user's initial decision-making process.[12] These elements within the interface illustrate the subtle yet impactful way dark patterns can shape user interaction and decision-making. Unsurprisingly, the academic community recognised the need to critically examine dark patterns, advocating for a balance between business objectives and ethical design principles.[13] This includes an analysis of the techniques underpinning user decisions, the legal ramifications of manipulative design, and the development of guidelines that could inform the creation of more transparent and user-centric digital environments.

Together with Dr Cristiana Santos, we stratified dark patterns into varying degrees of subtlety and integration within digital interfaces.[14] With an argument rooted in a textual analysis of enforcement decisions, we argued that the most apparent category is visible dark patterns, characterised by their overt and explicit presence within the user interface.[15] These are typically identifiable by regulatory bodies or auditors, exemplified by tactics such as obscuring buttons or the use of pre-checked boxes that may lead users to opt into services or make purchases inadvertently.[16] Progressing to a more concealed level, darker patterns emerge as subtle and elusive, often with only realised *post facto* consequences.[17] Detecting these patterns requires rigorous examination by regulatory authorities and expert auditors, as they are not immediately apparent. Such patterns include the strategic concealment of information or the imposition of forced practices on users. The most insidious tier is the darkest patterns embedded within the system architecture of online services. These patterns are often driven by sophisticated algorithms or artificial intelligence, crafting personalised experiences that may nudge users subconsciously towards choices not in their best interest. This level of dark patterns represents a significant challenge, as it requires deep technical expertise to uncover and understand how these systems operate beneath the surface to influence user behaviour.[18]

Each level represents a unique challenge in pursuing ethical digital design, necessitating a multifaceted approach to governance and oversight. As the complexity of these patterns increases, so does the need for advanced tools and methodologies to protect consumers from covert manipulation and maintain the integrity of user autonomy in the digital space. Our hierarchical framework delineates a gradient from the visible to the deeply embedded mechanisms of user manipulation within digital interfaces. Our analysis highlights the varying levels of detectability and the corresponding challenges each presents for ethical digital design. The journey from the overt manipulation of user interfaces to the covert shaping of user experiences by algorithms signifies an escalating complexity in consumer protection and regulation.

III. Psychological Patterns

1. From Dark Patterns to Deceptive Design: The Evolution of Digital Manipulation and the Legislative Response

This section dissects the layered complexity of the darkest patterns, manifested across the very architecture of the system, encompassing algorithmic and AI-based practices that operate beyond the immediate perception of users. Herein lies the crux of the issue: the interplay between user autonomy and the manipulative potential of digital interfaces, a dy-

[11] M.R. Leiser and M Caruana, 'Dark Patterns: Light to be found in Europe's Consumer Protection Regime' (2021) 10(6) Journal of European Consumer and Market Law 237-251.

[12] M.R. Leiser and Wen-Ting Yang, 'Illuminating Manipulative Design: From "Dark Patterns" to Information Asymmetry and the Repression of Free Choice under the Unfair Commercial Practices Directive' (2022) Loy Consumer L Rev, 34, 484.

[13] Gray (n 9).

[14] M.R. Leiser and Cristiana Santos, 'Dark Patterns, Enforcement, and the Emerging Digital Design Acquis: Manipulation beneath the Interface' (2024, Forthcoming) EJLT.

[15] Leiser and Santos (n 14).

[16] C Matte, N Bielova and C Santos, 'Do cookie banners respect my choice?: Measuring legal compliance of banners from IAB Europe's transparency and consent framework' (2020 IEEE Symposium on Security and Privacy (SP), May 2020) 791-809. V Morel et al, 'Legitimate Interest is the New Consent--Large-Scale Measurement and Legal Compliance of IAB TCF Paywalls' (arXiv preprint arXiv:2309.11625, 2023).

[17] Leiser and Santos (n 14).

[18] Leiser and Santos (n 14).

namic in constant flux and challenging to regulate. [19]

While this article attempts to navigate this conundrum, the question arises of how the law might articulate provisions capable of encompassing and circumscribing such a nebulous set of practices. The European Union's DSA has emerged as a platform regulation designed to face these challenges. However, as this article elucidates, the Act's current provisions may already be insufficient to regulate the whole gamut of deceptive designs, which are not limited to the overt 'dark patterns' but extend into the more covert realms of 'darker' and 'darkest' patterns.[20]

2In the legal sphere, this phenomenon was traditionally viewed through the prism of data protection[21] – settings constructed such that users must opt out of data utilisation[22], byzantine cookie declarations[23], prolix privacy advisories[24], and features cleverly crafted to cajole the disclosure of an excess of personal information.[25] However, consumer protection agencies have since become vigilant, prompting an influx of amendments to guidelines on executing the Unfair Commercial Practices Directive and consumer legislation.[26] Legislation emanating from Brussels invariably encompasses provisions proscribing 'deceptive' or 'manipulated' design, which 'steers' or 'coerce' users, culminating in a substantive distortion of their decision-making faculties.[27]

This transition to 'deceptive design' reflects an evolving awareness within the design and legal communities of the sophisticated ways digital platforms engage users. The initial focus on data protection has expanded as consumer protection agencies intensify their oversight, leading to legislative developments to curtail deceptive designs that distort consumer decision-making. The shift in terminology from 'dark patterns' to 'deceptive design' underscores the evolution of these manipulative strategies and their increasing sophistication. As the European Union grapples with the rapid advancements in digital design, the AI Act emerges as a critical tool, potentially more attuned than the DSA, to address and regulate the full spectrum of psychological patterns woven into the fabric of system architecture. Considering Leiser and Santos' findings, there is a pressing need for legislation that can adapt to these intricate and evolving challenges, ensuring that the integrity of user decision-making is maintained in an increasingly digital world.

Designers are transitioning from mere user interface manipulation (dark patterns) to intentionally altering user experience (darker patterns) and, most ominously, to embedding the most pernicious patterns within the system's architecture (darkest patterns).[28]

2. Article 25 DSA's Limited Capacity to Regulate Deceptive Design

The Digital Services Act[29] represents a landmark legislative instrument within the European Union to regulate online intermediaries across the EU Single Market. Its scope includes various online entities such as internet service providers, search engines, domain registrars, hosting services, and various online platforms, irrespective of their geographical establishment.[30] The DSA's primary objectives are to bol-

[19] A Mathur et al, 'Dark patterns at scale: Findings from a crawl of 11K shopping websites' (Proceedings of the ACM on Human-Computer Interaction, 3 (CSCW), 2019) 1-32; M Nouwens et al, 'Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence' (Proceedings of the 2020 CHI conference on human factors in computing systems, April 2020) 1-13; C Bösch et al, 'Tales from the dark side: privacy dark strategies and privacy dark patterns' (Proceedings Privacy Enhancing Technologies, 2016(4), 2016) 237-254.

[20] Leiser and Santos (n 14).

[21] J Gunawan, C Santos and I Kamara, 'Redress for dark patterns privacy harms? A case study on consent interactions' (Proceedings of the 2022 Symposium on Computer Science and Law, November 2022) 181-194.

[22] Nouwens (n 19) 1-13.

[23] C Gray et al, 'Dark patterns and the legal requirements of consent banners: An interaction criticism perspective' (Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021) 1-18.

[24] Bosch et al (n 19).

[25] I Borberg et al, 'So I Sold My Soul: Effects of Dark Patterns in Cookie Notices on End-User Behavior and Perceptions' (Workshop on Usable Security and Privacy (USEC), vol 3, 2022).

[26] Under the horizontal EU consumer law acquis, dark patterns can be addressed by the Unfair Commercial Practices Directive, the Consumer Rights Directive, and the Unfair Contract Terms Directive.

[27] Art 25 Digital Services Act (proposal) prohibiting deceptive online interfaces; art 13(6) Digital Markets Act on user autonomy and decision-making; art 66(2)(a) Data Act (proposal) against coercing or deceiving users through digital interfaces.

[28] Leiser and Santos (n 14).

[29] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance) [2022] OJ L 277/1.

[30] Commission, 'Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines' <https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413> accessed 9 February 2024.

ster user protection, enhance transparency, and foster innovation, reflecting the EU Commission's commitment to ensuring comprehensive user rights protection across all online services.

A critical aspect of the DSA is its approach to 'dark patterns' - deceptive design strategies that manipulate user choices. Article 25(1) DSA specifically targets these practices, prohibiting platforms from using interface designs that mislead, manipulate, or impair users' abilities to make informed decisions. However, this article is limited to 'online platforms', leaving a gap in applicability to other entities employing such tactics. The DSA endeavours to safeguard users' 'decisional space', encompassing their autonomy, choices, and decision-making processes.

According to Article 25, legislators have circumscribed the interdiction to those stratagems not hitherto encompassed by the General Data Protection Regulation and the Unfair Commercial Practices Directive yet extant within the online interface.[31] Regrettably, any subterfuge of design that resides beneath the superficial layer eludes the strictures of Article 25. This regulatory approach is tantamount to dictating the aesthetic choices of a homeowner: dictating the palette of their walls or the fabric of their drapery while neglecting the governance of the foundational integrity of the domicile and the absence of codification for its structural assembly. Recital 67 expands on this, delving into the nuances of interface structure, design, and functionalities. It highlights the need for digital designers to comprehend these terms in the context of the DSA. The Act's provisions are broad and abstract, potentially creating legal uncertainties and allowing for flexibility in addressing emerging technologies and influence types.

Together with Dr Cristiana Santos, I hypothesise that while the DSA is a significant step towards regulating online intermediaries and protecting users from deceptive designs (dark patterns), its effectiveness is limited by specific gaps and ambiguities.[32] These include its limited scope of application to specific types of online platforms[33], potential overlaps with existing legislations like GDPR and UCPD, and uncertainties in addressing emerging technologies and sophisticated dark patterns. While offering flexibility, the DSA's broad and somewhat abstract provisions also pose legal clarity and enforcement challenges. For example, the Article 25 prohibition only applies to deceptive design in the online interface. The text suggests that the DSA's success in combating dark patterns and protecting user autonomy in the digital space may require further clarification, guideline development, and possibly future amendments to address these challenges effectively.

The shift from the term 'dark patterns' to 'deceptive design' signifies a recognition of the evolution in manipulative techniques: from mere interface trickery to sophisticated strategies that affect the user experience and system architecture. Suppose that my analysis with Dr Santos holds. In this case, current regulatory frameworks, such as the Digital Services Act[34], may already be insufficient to address the totality of current deceptive designs. This inadequacy suggests that 'dark patterns' are merely a subset of deceptive practices, and thus, a broader regulatory approach, such as that potentially offered by the AI Act, is required. [35]

3. Integrating AI into Deceptive Design Strategies

Integrating AI in deceptive design strategies is a nuanced and complex phenomenon. AI systems, by their nature, are capable of processing vast amounts of data, learning from this data, and making decisions or predictions based on it. This capability is instrumental in developing and implementing manipulative techniques that can subtly influence human behaviour. In the context of the AI Act, the focus is on those AI systems that leverage their advanced computational abilities to enact manipulation 'beyond consciousness' and inflict 'psychological harm'. For example, AI can be utilised to personalise user experiences in a way that subliminally influences decision-making processes.[36] This can be achieved through algorithms that analyse user data to identify vulnerabilities or tendencies and then exploit these

[31] Art 25(2) Digital Services Act.

[32] Leiser and Santos (n 14, s 2.2).

[33] Art 2 Digital Services Act.

[34] Digital Services Act, 2020, COM(2020) 825 final <https://rb.gy/tbiuh9> accessed 14 July 2023.

[35] AI Act (n 6).

[36] M Franklin et al, 'Missing Mechanisms of Manipulation in the EU AI Act' (The International FLAIRS Conference Proceedings, 2022), 35 <https://doi.org/10.32473/flairs.v35i.130723> accessed 9 February 2024.

for manipulative purposes.[37] Techniques such as micro-targeted advertising, personalised content feeds, or subtly altered user interfaces can guide user behaviour in a specific direction without their conscious awareness.[38]

The AI Act aims to regulate AI systems, mainly focusing on those that could cause harm. To determine whether a particular AI-enabled manipulative technique falls under the Act, several critical criteria must be considered:

1. Nature of the AI System: The Act applies to AI systems that can manipulate behaviour beyond the conscious awareness of individuals. This includes systems that employ advanced data analytics, machine learning, and pattern recognition to identify and exploit psychological vulnerabilities.
2. Method of Manipulation: The manipulation must be subliminal, which operates below the individual's conscious awareness threshold. This could include the use of invisible cues or the deployment of algorithms designed to influence behaviour subtly.
3. Potential for Harm: The AI system's manipulative techniques must have the potential to cause material distortion in behaviour, leading to physical or psychological harm. This criterion requires a causal link between the manipulation and the harm, necessitating a rigorous evaluation of the impact of these AI systems.
4. *Target of Manipulation:* The Act addresses explicitly manipulative practices that target vulnerable groups, such as individuals distinguished by age or disability. This focus reflects a heightened concern for protecting those more susceptible to manipulation.

By considering these criteria, it becomes possible to evaluate whether specific AI-driven psychological manipulation techniques fall within the scope of the AI Act. For instance, an AI system that uses subliminal messaging to exploit cognitive biases in a manner that could lead to addictive behaviours or mental distress would likely be encompassed by the Act, provided there is a demonstrable risk of harm.

The relationship between AI and deceptive design strategies hinges on the ability of AI systems to analyse and exploit human psychology in a manner that is both subliminal and potentially harmful. The AI Act seeks to regulate these practices by establishing clear criteria focusing on the nature of the AI system, the method of manipulation, the potential for harm, and the target. Understanding these criteria is pivotal in assessing the extent to which AI can influence psychological manipulation techniques and in determining the applicability of the AI Act to such techniques.

For Dr Santos and myself, this broader perspective of deceptive design encompasses a range of manipulative practices that are often not immediately visible to the user but are deeply embedded within the system's architecture, affecting not only the user experience but also the user's autonomy and decision-making.[39] These can be categorised into several types of manipulation, each with its unique impact on the user's interaction with the system, as outlined in the subsequent discussion. The classification of these methods is multifaceted and can be distilled as follows:

Sensory manipulation, as an umbrella term, refers to many techniques that subtly influence an individual's sensory perception, affecting their cognitive functions and behavioural outcomes.[40] These methods engage the senses in a manner that often bypasses conscious awareness, leading to a form of psychological manipulation. Subliminal messaging, for example, operates by presenting stimuli below the threshold of conscious perception, which can influence attitudes or actions without overt recognition by the individual. Background audio, another method, can include specific frequencies or rhythms that can alter mood or arousal states, potentially guiding decision-making processes. Colour psychology uses emotional and psychological associations with certain hues to induce specific states of mind or encourage actions. Collectively, these techniques could

[37] G Wagner and H Eidenmüller, 'Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping Preferences: Regulating the Dark Side of Personalized Transactions' (2019) 86 University of Chicago Law Review 593–595.

[38] A Simchon, M Edwards and S Lewandowsky, 'The persuasive effects of political microtargeting in the age of generative AI' (PsyArXiv, 2023); G Spitale, N Biller-Andorno and F Germani, 'AI model GPT-3 (dis)informs us better than humans' (2023) 9(26) Science Advances.

[39] Leiser and Santos (n 14).

[40] O Petit, C Velasco an C Spence, 'Digital sensory marketing: Integrating new technologies into the multisensory online experience' (2019) 45(1) Journal of Interactive Marketing 42-61; See also S Paek, DL Hoffman and JB Black, 'Shaping the sensory experience in digital environments: modality, congruency, and learning' (2021) Interactive Learning Environments 1-17.

amount to psychological manipulation, as they are employed to elicit predetermined responses that serve the interests of the manipulator, often without the conscious consent or awareness of the manipulated. This manipulation capitalises on the inherent vulnerability of the senses as gateways to the psyche, ultimately steering individuals in ways that might not align with their autonomous preferences.

Through the sophisticated deployment of adaptive algorithms[41], *algorithm manipulation* exerts a formidable influence over user experience and decision-making processes. Such algorithms, which constitute the backbone of personalised user interfaces, tailor content and user interactions based on individual behavioural data, subtly directing attention towards confident choices and away from others.[42] While ostensibly serving user preferences, this personalised curation can engender echo chambers, isolated information environments with limited exposure to diverse perspectives. Reinforcing pre-existing beliefs through algorithmic filtering can thus ossify ideological positions, leading to polarisation.[43] Moreover, latency manipulation, wherein the response time of digital interfaces is deliberately calibrated, can alter user engagement and decision-making patterns.[44] By varying the speed of content delivery, algorithms can influence the perceived attractiveness or importance of certain information, further steering user behaviour.[45]

The crux of the issue lies in the opaqueness of these algorithmic processes; users need to be made aware of the extent to which their digital environment is being shaped and their choices pre-configured. This covert orchestration can amount to psychological manipulation, as it exploits cognitive biases and behavioural patterns to mould user experience and decision-making, often aligning user actions with the strategic objectives of the algorithm's designers or deployers.

Behavioural conditioning operates at the fulcrum of psychological influence, manifesting as an intricate interplay of stimuli and responsive actions aimed at sculpting human conduct.[46] This paradigm is intricately woven into the fabric of reward systems and gamification, where the allure of incentives engineers a user's behaviour towards a predefined pattern or goal. The very essence of such mechanisms is to elicit and reinforce actions through the strategic dispensation of rewards, be it in tangible or virtual forms. Conversely, the paradigm also encompasses the domain of manipulative feedback, where the withholding or presentation of feedback is calculated to induce behavioural modifications, often subverting the user's awareness of being influenced. Such positive or negative feedback is not a mere reflection of performance but a tool to channel behaviour in a desired direction. In addition, subtle reinforcements, often indistinguishable in their immediacy, cumulatively contribute to the gradual recalibration of user habits. These reinforcements, adeptly integrated into the user's routine, leverage the incremental nature of behavioural adaptation, thereby establishing or altering user habits over time. This gradualism is the foundation of behavioural conditioning, ensuring that individuals are unaware of how their behaviours have been shaped externally. It resides in the nebulous terrain where the demarcation between benign guidance and coercive manipulation blurs. In this context, psychological manipulation denotes the orchestration of an individual's behaviour without their explicit consent or, at times, their knowledge, effectuating a shift in their decision-making processes and autonomy.[47] By exploiting cognitive biases and emotional triggers, behavioural conditioning can transcend the boundaries of persuasion and influence, culminating in a covert form of control that usurps the individual's volition. In this subtle usurpation, manipulation looms, calling into question the moral implications of such practices in shaping human behaviour.

System dependency and control are pivotal strategic design components that inexorably bind users to

[41] K Yeung, 'Algorithmic Regulation: A Critical Interrogation' (2018) 12 Regulation and Governance 505, 507.

[42] K Yeung, 'Hypernudge': Big Data as a mode of regulation by design' (2017) 20(1) Information, Communication & Society 118-136.

[43] HF de Arruda et al, 'Modelling how social network algorithms can influence opinion polarization' (2022) 588 Information Sciences 265-278.

[44] B Watson et al, 'Effects of variation in system responsiveness on user performance in virtual environments' (1998) 40(3) Human Factors 403-414.

[45] S Lewandowsky et al, 'Technology and Democracy: Understanding the Influence of Online Technologies on Political Behaviour and Decision-Making' (JRC Publications Repository, 2020) 45.

[46] BF Skinner, 'Operant conditioning' (1971) The Encyclopedia of Education, 7, 29-33.

[47] For variations of this concept, see RB Cialdini, *Influence: The psychology of persuasion* (vol 55, Collins Business Essentials, Harper Business 2007), 339.

a technological ecosystem.[48] This phenomenon is underpinned by the utilisation of network effects, where the value of a service increases commensurately with the number of its users.[49] Such an architecture is further entrenched by governing system updates' temporal sequence and substance. This often leaves users with little recourse but to adhere to the preordained trajectory of the platform's evolution. Intermittent connectivity, deliberately engineered breaks in service, can serve as a psychological lever, engendering a form of variable reinforcement that can increase user engagement like the mechanisms exploited by gambling industries. This intermittent reinforcement schedule is insidious, capitalising on the human predilection for pattern recognition and reward-seeking behaviour. As users navigate this engineered landscape, they may become unwitting participants in a larger schema of psychological manipulation, by which their autonomy is subverted by a calculated orchestration of their digital experience. The consequence is not merely a technologically mediated habituation but a deep-seated psychological dependence that can prove arduous to extricate oneself from, thus raising significant ethical considerations regarding the stewardship of digital autonomy and agency.[50]

Structural manipulation refers to the overarching design principles that govern the infrastructure of systems, potentially leading to the centralisation of power or informational asymmetries. It includes decentralisation versus centralisation, data collection and profiling, and the architectural mechanisms that lead to dependency and lock-in. This manipulation is inherently twofold, serving both as a scaffold for the organisation of systems and as a lever for the potential aggrandisement of dominion or the engendering of informational imbalances. At the heart of this concept lies the dialectic of decentralisation and centralisation, a spectrum that dictates the locus of control within a network. Decentralised systems are typified by a diffusion of power, with autonomous nodes operating independently, thereby mitigating single points of failure and resisting monopolistic control. In stark contrast, centralised systems consolidate authority within a command nucleus, often leading to heightened efficiency, but at the expense of a monopolistic power dynamic that can quash competition and innovation.

An integral aspect of structural manipulation is data collection and profiling. Systems structured around the greedy accumulation of data often segue into surveillance, profiling users and their behaviours to sculpt a detailed compendium of personal information. This data repository can be wielded to tailor and target content with uncanny precision, shaping perceptions and influencing decision-making processes in a manner that borders on the insidious. Architectural mechanisms embedded within systems architecture, such as proprietary standards or closed ecosystems, precipitate dependency and lock-in, chaining the user to a particular service or suite of products. This interdependence curtails choice, stifling the user's freedom to migrate to alternative systems without incurring significant costs or losses in data and functionality.

The psychological manipulation inherent in structural manipulation arises from the subtle yet pervasive influence these systems exert on user behaviour and thought patterns. By controlling the flow of information, shaping user interactions, and confining choices, these structures can surreptitiously influence the psyche, nudging users towards certain behaviours or decisions. Over time, repeated exposure to tailored content and constrained choices can recalibrate perceptions and priorities, often without the user's conscious awareness, leading to psychological conditioning. This manipulation, though structural in design, transcends the physical realm of systems and profoundly embeds itself in the cognitive landscapes of individuals, potentially altering not just actions but the way reality is construed.

Many of these techniques are used in conjunction with others. Consider a system predicated on ‘persuasion profiling’[51] or ‘hyper-nudging’[52], which surreptitiously aggregates behavioural data to ascertain the persuasive modalities most efficacious on an individual or demographic cohort, deploying deceptive patterns customised to their susceptibilities. Imagine a scenario where the system discerns a predilection

[48] This concept has been argued across a number of disciplines: in tech, see the collective works of E Morozov; S Zuboff, ‘The age of surveillance capitalism’ in W Longhofer and D Winchester (eds), *Social Theory Re-Wired* (3rd edn, Routledge 2023) 203-213; and J Bridle, *New dark age: Technology and the end of the future* (Verso Books 2018).

[49] M A Lemley and D McGowan, 'Legal Implications of Network Economic Effects' (1998) 86 Calif L Rev 479.

[50] J Lanier, ‘Agents of alienation’ (1995) 2(3) Interactions 66-72.

[51] M Kaptein, ‘Persuasion profiling: How the internet knows what makes you tick’ (2015) Business Contact.

[52] Yeung, ‘Hypernudge’ (n 42) 118.

for susceptibility to temporal urgency over other cognitive biases, resulting in an inundation of deceptive patterns exploiting this bias.[53] Extrapolating beyond cognitive biases, vulnerabilities such as dyslexia or dyscalculia could be targeted with intricate wording or numerically convoluted offers to precipitate actions of significant consequence, such as contractual commitments.

In his seminal text, 'Persuasion Profiling: How the Internet Knows What Makes You Tick', Kaptein contended that 'persuasion profiles' signify a fundamental evolution in improving the efficacy of online marketing strategies and are particularly crucial for entities engaged in the digital commerce sphere. The illustrative figure from Kaptein's research delineates a 'persuasion profile', with each horizontal line representing an 'influence principle' gleaned from Cialdini's 'weapons of influence'.[54] The horizontal axis measures the probability that each element influences the user's decision-making process positively or negatively.

Within a particular investigative study, Kaptein et al. enlisted many participants to undertake a 'susceptibility to persuasion scale' (STPS), from which individual persuasion profiles were deduced. Subsequently, these participants were involved in a dietary regimen, the objective of which was to reduce their snacking consumption between meals. Participants reported their snacking habits via SMS. Unknowingly, the SMS communications they received contained persuasive content, which varied according to the experimental condition. Those participants whose messages were consistent with their personal persuasion profiles were influenced more effectively than others – evidenced by a significant decrease in snacking intervals.

The findings of Kaptein et al, in their investigative study, bridge the theoretical concepts outlined in *Table 1* with practical evidence of the impact of system design on user behaviour. The research used an STPS to create individual persuasion profiles and showed that tailored persuasive content could significantly alter habits, such as reducing snack consumption. SMS messages, imbued with persuasive elements aligned with the user's profile, were a live application of the psychological techniques categorised in Table 1 below. The successful modulation of snacking behaviour through personalised messaging is a testament to the potency of system design informed by a deep understanding of the psychological categories that govern user interaction and response. This synergy between the persuasive techniques identified and applied in the study exemplifies the transformative potential of design that takes advantage of user-specific psychological profiles to guide behaviour subtly and effectively.

Thus, Table 1 classifies various psychological techniques in system architecture into five categories. These categories reflect the modality through which user experience and behaviour can be influenced by system design.

The comprehensive classification presented in *Table 1* elucidates how system design, through various psychological techniques, can influence user experience and behaviour, paving the way for a nuanced understanding of user-system interaction. Consider a system that uses adaptive algorithms to foster dependency or employs temporal manipulation to condition user behaviour. These systems might alter response times or functionality through updates, leading users toward certain behaviours under the guise of 'system improvements,' a tactic that falls within Algorithmic Manipulation and System Dependency and Control. Alternatively, a platform might manipulate latency to sway user decisions or grant disparate data access, an example of structural manipulation creating a tiered user experience that pressures individuals into compliance or payment for an enhanced service. These instances underscore the potential for system design to venture beyond the straightforward 'dark patterns' of user interface manipulation, delving into the intricate web of code and architecture where the DSA's reach may falter. However, the AI Act could provide a bastion against such covert tactics by regulating the underlying mechanisms that dictate user engagement, ensuring a safeguard against exploiting psychological vulnerabilities.

Consider a system employing adaptive algorithms to enhance user dependence or subconscious engagement or one that manipulates the cadence of exposure to messages or cues to elicit specific behaviours. A system calibrated to recognise the behaviour profile of a user, such as a preference for rewards over aversions to punishment, could variably adjust its re-

[53] Kaptein (n 51).

[54] RB Cialdini, 'Harnessing the science of persuasion' (2001) 79(9) Harvard Business Review 72–81.

Table 1. *Psychological techniques embedded in system design*

Psychological Technique	Sensory Manipulation	Algorithmic Manipulation	Behavioural Conditioning	System Dependency and Control	Structural Manipulation[a]
Data Collection and Profiling[b]		X			X
Algorithmic Filtering and Echo Chambers[c]		X			
Dependency and Lock-In[d]				X	X
Reward Systems and Gamification[e]			X		X
Intermittent Connectivity				X	
Latency Manipulation		X		X	
Decentralisation vs. Centralisation					X
Control over Updates				X	X
Network Effects[f]				X	X
Information Asymmetry[g]					X
Subliminal Messaging	X				
Background Audio	X				
Manipulative Feedback	X		X		
Colour Psychology[h]	X				
Subconscious Triggers	X		X		
Adaptive Algorithms[i]		X	X		

sponse times to induce desired actions or habit formation. A system could deploy updates at intervals dictated by the developer, altering functionalities without transparent communication, thereby steering users toward preferred actions or behaviours under the guise of 'system improvements'. A system could be intentionally engineered to increase response times during pivotal user decisions, potentially inducing doubt or coercing choices that align with the platform's objectives. A system could be designed to grant discriminatory access to data or features based on the user's status or subscription level, creating a hierarchy of influence and compelling the less privileged to capitulate to payment or compliance to achieve parity. Such asymmetrical access could create environments conducive to exploitation or manipulation. Each scenario delineated surpasses the implementation of an online interface with a 'dark pattern'. They require intricate coding within the system's architecture, a domain beyond the ambit of Article 25 DSA. However, the AI Act might offer a sanctuary for rectification.

Psychological Technique	Sensory Manipulation	Algorithmic Manipulation	Behavioural Conditioning	System Dependency and Control	Structural Manipulation[a]
Hidden Repetition[j]	X		X		
Temporal Manipulation		X			
Subtle Reinforcements			X		
Masked Penalties			X		

[a] N Helberger et al, 'EU consumer protection 2.0 Structural asymmetries in digital consumer markets' (BEUC, 2021) <https://www.beuc.eu/sites/default/files/publications/beuc-x-2021-018_eu_consumer_protection_2.0.pdf> accessed 9 February 2024

[b] N Sörum and C Fuentes, 'How sociotechnical imaginaries shape consumers' experiences of and responses to commercial data collection practices' (2023) 26(1) Consumption Markets & Culture 24-46.

[c] L Serafini, 'The old-new epistemology of digital journalism: how algorithms and filter bubbles are (re) creating modern metanarratives' (2023) 10(1) Humanities and Social Sciences Communications 1-9.

[d] B Artur, YM Ermol'ev and YM Kaniovskii, 'A generalized urn problem and its applications' (1983) 19(1) Cybernetics 61-71; PA David, 'Clio and the Economics of QWERTY' (1985) 75(2) The American economic review, 334, 332-337.

[e] P Rahmadhan et al, 'Trends and Applications of Gamification in E-Commerce: A Systematic Literature Review' (2023) 9(1) Journal of Information Systems Engineering & Business Intelligence.

[f] R Soeiro and AA Pinto, 'Negative network effects and asymmetric pure price equilibria' (2023) 22(1) Portuguese Economic Journal 99-124; S Miranda et al, 'Addiction to social networking sites: Motivations, flow, and sense of belonging at the root of addiction' (2023) 188 Technological Forecasting and Social Change, 122280.

[g] W-T Yang and M.R. Leiser, 'Illuminating Manipulative Design: From 'Dark Patterns' to Information Asymmetry and the Repression of Free Choice under the Unfair Commercial Practices Directive' (2022) 34 Loyola Consumer Law Review, 484.

[h] H ChangDa and A Bhaumik, 'Colour Psychology's Impact on Marketing, Advertising, and Promotion' (2023) 7(1) International Journal of Management and Human Science 24-32.

[i] Y Gui, D Li and R Fang, 'A fast adaptive algorithm for training deep neural networks' (2023) 53(4) Applied Intelligence 4099-4108.

[j] Taking advantage of, for example, knowledge about a traumatic event to repeatedly trigger a psychological reaction. C Caruth, Unclaimed experience: Trauma, narrative, and history (JHU press 2016); See also, M Ahmad et al, 'No safe place for war survivors: War memory, event exposure, and migrants' psychological trauma' (2023) 13 Frontiers in Psychiatry, 966556 <https://doi.org/10.3389/fpsyt.2022.966556> accessed 12 February 2024.

The subsequent analysis will explore the legislative ambiguities and potential within the AI Act.[55] It will propose a reframing of 'personality traits' (a term found in Recital 38 of the Proposal and the European Parliament's Mandate) within a broader lexicon of psychological constructs, advocate for precise definitions of manipulation techniques, and call for the reconceptualisation of 'informed decisions' within the digital milieu. Moreover, it will highlight the need for a holistic legislative approach that accounts for the deceptive design's multifaceted nature, transcending superficial interface design to consider the complex psychological manipulations embedded within system architecture.

In this scholarly exploration, the intricacies of Article 5 of the AI Act will be dissected in the next section, examining its capacity to regulate the covert and subliminal techniques that AI systems may employ to distort user behaviour. The 'psychological harm' concept, replaced with 'significant harm' in the legislative text, will be scrutinised for its legal and psychological implications. Furthermore, the discussion will address the technical and ethical challenges in defining and proving harm, particularly in subliminal manipulation and exploiting vulnerabilities.

IV. The EU's Approach to Regulating Artificial Intelligence

The ongoing debate on the ethical implementation of AI technologies has arrived at a pivotal moment with the presentation of the AI Act. The definition of prohibited practices that could inflict harm upon individuals is at the heart of this conversation, an issue that Article 5 aims to tackle. Ideally, this article should have profoundly acknowledged the legal and

[55] AI Act (n 6).

ethical intricacies of governing technologies with the power to impact human thought and action. However, it may need help from the courts to fully realise its intended purpose.

1. Article 5(1)(a) of the AI Act Proposal

Article 5(1)(a) of the Commission Proposal[56], European Parliament's Amendments,[57] and Council's Mandate[58] discussed prohibiting specific uses of AI systems that manipulate behaviour.[59] The differences between the three reflect variations in the scope and the specific conditions under which such AI practices are considered unacceptable. The Commission's proposal focused on using AI systems that deploy subliminal techniques beyond the consciousness of a person to materially distort a person's behaviour so that it could cause physical or psychological harm. The emphasis was on using techniques that are beyond the person's conscious awareness and have a significant negative impact on behaviour. However, the European Parliament's (EP) amendments expanded upon the Commission's proposal in several ways. First, it would have prohibited subliminal techniques and explicitly included 'purposefully manipulative or deceptive techniques'. This would have effectively broadened the scope to include any AI practices that intentionally distort a person's or a group of persons' behaviour. Second, the EP mandate addressed the impairment of the person's ability to make an informed decision, potentially leading to a decision that the person would not have made otherwise. This addition to the text introduced the concept of informed decision-making, suggesting protection against AI systems that could significantly impair this ability and cause harm. Furthermore, the EP mandate specifies that the harm can be to a group of persons, indicating a concern for collective harm, not just individual.

The Council Mandate's language closely resembled the EP Mandate's language with some nuances. It maintained the broader scope of prohibited AI practices, including those with the objective or effect of distorting behaviour. However, it changed the standard of harm caused from 'likely to cause' to 'reasonably likely to cause,' which could imply a different threshold for determining the potential harm of AI systems. This could have lowered the bar for what constitutes harm, broadening the range of AI practices that could fall under prohibition. Like the EP mandate, the Council mandate also recognised the effect on groups of persons.46F

The primary differences between the three mandates concerning Article 5(1)(a) are as follows:

- Scope of Techniques: The EP and Council Mandates included a broader range of manipulative techniques beyond subliminal ones, expanding the scope to purposefully manipulative or deceptive practices.
- Informed Decision Making: The EP Mandate uniquely emphasised the protection of an individual's or group's ability to make informed decisions, suggesting a particular concern for the cognitive autonomy of users.
- Standard of Harm: The Council Mandate modified the likelihood standard for harm, potentially broadening what AI systems could consider harmful behaviour manipulation.
- Collective Harm: Both the EP and Council Mandates explicitly mentioned the potential for collective harm to groups of persons, indicating an awareness of the broader social impact of such AI practices.

As it delineates two prohibitions, the Commission's proposal for Article 5(1)(a) and (b) should have served as a pillar against potentially abusive AI practices. First, there was a ban on AI systems that manipulat-

[56] 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 <https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-COM-Proposal-21-April-21.pdf> accessed 12 February 2024.

[57] European Parliament, 'Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' <https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf> accessed 12 February 2024.

[58] Council of the European Union, '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 - General approach' (25 November 2022) <https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-CZ-%E2%80%93-General-Approach-25-Nov-22.pdf> accessed 12 February 2024.

[59] The Commission's Proposal, the European Parliament's Adopted Amendments, and the Council's Mandate can be compared at the following link: <https://artificialintelligenceact.eu/wp-content/uploads/2023/08/AI-Mandates-20-June-2023.pdf> accessed 12 February 2024.

ed subliminal techniques to alter a person's consciousness, and second, there was the prevention of exploiting vulnerabilities due to age or physical or mental disability:

Article 5(1)(a) and (b): The following artificial intelligence practices shall be prohibited:

(a)the placing on the market, putting into service or use of an AI system that deploys 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.

(b)the placing on the market, putting into service or use of an 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 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.

These provisions are designed to mitigate the risk of psychological or physical harm that such technologies could pose.

However, Article 5(1)(a) from the finalised text states the following:

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective to or the effect of materially distorting a person's or a group of persons' behaviour by appreciably impairing the person's ability to make an informed decision, thereby causing the person to take a decision that that person would not have otherwise taken in a manner that causes or is likely to cause that person, another person or group of persons significant harm;

This means six things are now required to happen before the prohibition takes place:

– An AI system must be placed on the market, put into service, or used.
– The AI system must deploy subliminal techniques beyond a person's consciousness. Alternatively, it may use purposefully manipulative or deceptive techniques.
– The techniques used by the AI system should have the objective or the effect of materially distorting a person's or a group of persons' behaviour.
– The distortion must appreciably impair the person's ability to make an informed decision.
– As a result of the distortion, the person or group decides they would not have otherwise made.
– The decision caused by the distortion must cause or is likely to cause significant harm to the person, another person, or a group of persons.

1. A Matter of Interpretation

Both Article 5(1)(a) and (b) of the AI Act set out to prevent AI systems from using subliminal techniques that significantly manipulate behaviour and cause harm. The challenge lies in the vague legal language, particularly the term 'beyond consciousness,' which could lead to various interpretations. What qualifies as influence beyond the conscious mind is still being determined, complicating the enforcement of these prohibitions. To trigger these legal safeguards, the manipulation must be imperceptible to the conscious mind and strong enough to alter behaviour significantly.

This raises two primary issues: the first is the demarcation of the threshold where influence recedes from the conscious arena into the shadowy depths of the subconscious, and the second is the evidentiary burden to establish that such subliminal manoeuvring is causally linked to discernible harm. This ambiguity extends to practices which could subtly influence behaviour without the user's conscious awareness. Moreover, the finalised text fails to clearly define what constitutes psychological harm, leaving a wide margin for interpretation. This lack of clarity will hinder the practical application of prohibitions, which calls for a more precise framework to address the nuanced interplay between technological innovation and ethical considerations.[60]

Delineating what constitutes influence 'beyond consciousness' remains an elusive quarry. The finalised AI Act's current language suggests that, for the prohibition to be triggered, the manipulation must be undetectable by the conscious mind and po-

[60] See also M Franklin et al, 'Missing Mechanisms of Manipulation in the EU AI Act' (The International FLAIRS Conference Proceedings, 2022), 35 <https://doi.org/10.32473/flairs.v35i.130723> accessed 12 February 2024.

tent enough to skew behaviour materially. This presents a dual challenge. The liminality of 'beyond consciousness' thus beckons a multitude of interpretations. The finalised text of Article 5 refrains from offering a concrete definition of psychological harm, a concept not uniformly recognised across medical, psychiatric, psychological, or legal disciplines, leaving a substantial grey area concerning the extent and nature of harm that would warrant the prohibition's application.

The ambiguity surrounding the definition of 'beyond consciousness' and psychological harm in the finalised text of Article 5 leaves room for a spectrum of potentially harmful practices that may take time to become apparent to users. For example, biometrically targeted advertising, which might be arguably perceptible, may be subtly intertwined with the user's psychological fabric to the extent that eludes conscious awareness. The lack of a clear definition for psychological harm in the proposal exacerbates the uncertainty of what harm requires prohibition.

Transitioning from this conceptual grey area, *Table 2* offers a structured approach to understanding the latent impacts of AI by categorising psychological techniques that operate subliminally. Methodically breaks down the types of manipulation and the corresponding harms that can arise from such practices. By identifying specific outcomes such as impulsive behaviours or addiction, the table provides a clearer picture of the potential consequences of AI systems that exploit subconscious processes. This elucidation is crucial for creating more precise regulatory frameworks that can effectively address the challenges posed by AI, ensuring that systems are designed and operated with the user's psychological integrity in mind.

Thus, Table 2 outlines the intersection of AI-driven psychological techniques with their respective manipulation categories and the potential material harm that can arise when such techniques are employed beyond the user's consciousness. It serves as a resource highlighting the need for ethical considerations in designing and regulating AI systems to prevent covert manipulation and safeguard user welfare. In this vein, Article 5(1)(a) delineates prohibited practices where AI systems deploy subliminal techniques or exploit vulnerabilities of certain groups to materially distort behaviour to the extent of causing harm. However, this leaves an expanse of dubious ground untouched, wherein AI systems might navigate the peripheries of consciousness without breaching into the territory of tangible harm. This regulatory lacuna thus places an onus on the legal and technological communities to grapple with the subtleties of 'beyond consciousness', to establish more precise guidelines, and to ensure that AI systems operate within the realms of ethical acceptability, safeguarding the autonomy of individuals against the unseen currents of subliminal persuasion.

The final limb of the wrongful practice test under Article 5(1)(a) of the finalised text is related to the use of artificial intelligence systems that employ covert subliminal techniques to materially distort a person's behaviour in a manner that could inflict physical or psychological harm. The term 'psychological harm' has undergone lexical refinement to 'significant harm' in the Parliament's iteration, but both versions underscore the concern for manipulative techniques that elude a person's conscious awareness. An illustrative example of such manipulation might involve an AI system that detects user ennui within a digital experience and consequently emits an imperceptible sound designed to prolong engagement. This strategy falls under the ambit of sensory manipulation.

The imperative lies in assessing whether extant or emerging practices, such as those encapsulated by the taxonomy of dark, darker, and darkest patterns, are subsumed under this prohibition. Interpreting the provisions set forth by the finalised text requires a scrupulous examination of Article 5, revealing that the interdiction is not absolute. Instead, it is circumscribed by the caveat that manipulative practices must be both subliminal and causative of harm, physical or psychological. In the economic or political spheres, manipulating human behaviour by such means remains unprohibited unless it culminates in harm. The AI Act, as it stands, does not offer any noticeable guidance on the constitution of 'harm', a concept lacking a universal definition across the disciplines of medicine, psychiatry, psychology, and law.

A prerequisite for the applicability of the prohibition is the subliminal nature of the manipulation, a term that the finalised text should have elucidated with clarity. Should the interpretation be, that manipulation must elude our sensory perception, certain advertising practices that leverage biometric data to potentially harmful effects may evade coverage under the finalised text. The stipulation that the practice must engender harm introduces an additional

***Table 2.** Psychological techniques that operate 'beyond consciousness' and identify the potential material harm they may cause*

Psychological Technique	Category of Manipulation	Potential Material Harm
Subliminal Messaging[a]	Sensory Manipulation	Impulsive behaviours, unhealthy consumption habits, psychological distress through unconscious influence
Background Audio	Sensory Manipulation	Mood alteration leading to anxiety or stress, manipulation of consumer behaviour
Colour Psychology[b]	Sensory Manipulation	Environmental cues that cause stress or anxiety and impact mental well-being
Latency Manipulation[c]	Algorithmic Manipulation	Opt-out delays lead to privacy loss and manipulation into consent for unwanted terms.
Temporal Manipulation[d]	Algorithmic Manipulation	An altered perception of time influences user interactions, causing impatience or dependence.
Manipulative Feedback	Behavioural Conditioning	Addiction to platforms or services, over-reliance on technology, impacting mental health
Hidden Repetition	Behavioural Conditioning	Subconscious conditioning leads to changes in behaviour or habits, potential addiction, or overconsumption.
Dependency and Lock-In[e]	System Dependency and Control	Reduced autonomy, inability to switch services, leading to over-reliance and potential exploitation
Control over Updates	System Dependency and Control	Forced compliance, unwarranted data sharing, and exploitative changes in service terms
Information Asymmetry[f]	Structural Manipulation	Financial losses, unfair treatment due to lack of information
Adaptive Algorithms[g]	Algorithmic Manipulation & Behavioural Conditioning	Increased dependency, addictive behaviours, and decisions benefiting service providers at the expense of user well-being

[a] W Hofmann, M Friese and RW Wiers, 'Impulsive versus reflective influences on health behavior: A theoretical framework and empirical review' (2008) 2(2) Health psychology review 111-137.

[b] M Schweitzer, L Gilpin and S Frampton, 'Healing spaces: elements of environmental design that make an impact on health' (2004) 10 (Supplement 1) Journal of Alternative & Complementary Medicine, S-71; See also C Karthikeyan and R Joy, 'An exploratory study on colour psychology in marketing: A techno-leadership perspective' (2018) 8(9) International Journal of Research in Social Sciences 65-92.

[c] I Ayres, 'Regulating opt-out: an economic theory of altering rules' (2011) 121 Yale LJ, 2032; N Gerber et al, 'Don't accept all and continue: Exploring nudges for more deliberate interaction with tracking consent notices' (2023) 31(1) ACM Transactions on Computer-Human Interaction 1-36.

[d] M Obstfeld, 'Intertemporal dependence, impatience, and dynamics' (1990) 26(1) Journal of Monetary Economics 45-75.

[e] P Marchildon and P Hadaya, 'Understanding the impacts of increasing returns in the context of social media use' (2022) 35(3) Information Technology & People 1136-1169.

[f] C Bicchieri and A Chavez, 'Behaving as expected: Public information and fairness norms' (2010) 23(2) Journal of Behavioral Decision Making 161-178.

[g] Yeung, 'Algorithmic Regulation' (n 41).

layer of complexity since establishing a definitive threshold for harm and demonstrating a causal nexus between the manipulative AI practice and the resultant detriment will remain arduous.

2. Article 5(1)(b) of the AI Act

The second form of an illegal system involves exploiting vulnerabilities specific to groups distinguished by age, physical, or mental disability. To prove a violation of Article 5(1)(b), the following elements must be demonstrated:

(a) An AI system in question has been placed on the market, put into service, or used.

(b) This AI system exploits vulnerabilities of a person or a specific group of persons. These vulnerabilities are due to age, disability, or a specific social or economic situation.

(c) The exploitation has the objective to or the effect of materially distorting that person's behaviour or a person about that group.

(d) The distortion in behaviour is such that it causes or is reasonably likely to cause significant harm to the person targeted or another person.

Each of these components serves as a critical link in establishing the liability of an AI system under the prohibition outlined in Article 5(1)(b), ensuring that AI systems do not perpetuate harm by taking advantage of the most vulnerable populations. Here, the same stringent threshold is invoked, the distinguishing factor being the nature of vulnerability at play. The Parliament's rendition of the Proposal extended its reach to encompass additional vulnerabilities, such as known or predicted personality traits or socioeconomic conditions. However, this expanded ambit does not amount to a categorical prohibition. In isolation, an AI system's exploitation of individual vulnerabilities does not suffice to make it unacceptable; it must also materially distort the behaviour and be likely to cause harm.

3. Missing the Mark

The finalised text of Article 5 of the AI Act sidestepped a broad swath of manipulative techniques, which are addressed under the remit of other legislative frameworks, such as the DSA, data protection laws, and consumer protection statutes. However, the principles of free choice, transparency, and informed consent, as enshrined in the GDPR, are only sometimes clear-cut in the context of deceptive design. The UCPD confines itself to commercial practices that significantly distort, or are likely to distort, consumer economic behaviour. The DSA is poised to proscribe dark patterns, mandate transparency in online advertising, and strengthen advertising safeguards. However, as evidenced, it stops short of universally prohibiting manipulative techniques that operate *sub rosa*. Consequently, myriad AI-facilitated patterns elude explicit prohibition by legislative instruments.

Moreover, a leak of the compromise text from the Trilogue negotiations has sparked further debate and confusion. This text considers the classification of high-risk systems and posits that any system that engages in profiling should be deemed high-risk. However, the AI Act expressly excludes exploitative profiling systems from its purview. Therefore, the language remains a pertinent issue because the classification of profiling as 'high-risk' does not inherently diminish its potential for exploitation. Such a compromise could potentially erode the efficacy of Article 5, which, despite its vagueness and ambiguity, contains ambitious objectives.

Consequently, the next section will provide recommendations to refine the AI Act, advocating for precise definitions and broader protections that extend beyond personality traits to other psychological constructs. By critically examining the AI Act's current provisions and suggesting improvements, this dialogue contributes to the broader discourse on ensuring that AI serves humanity's interests without compromising the rights and well-being of individuals. This academic discourse is not merely an exercise in legislative analysis; it is a call to action for policymakers, technologists, and legal practitioners to advance a regulatory framework that can keep pace with the rapid evolution of digital technologies. It is an endeavour to ensure that the digital future remains a landscape where user rights are unequivocally respected and protected.

V. Recommendations

Legislating AI illuminates the challenge of delineating the parameters of its acceptable use in a world increasingly mediated by digital interfaces. The nuanced complexities of the provisions of the Act reflect the intricate balance between fostering innovation and protecting individuals from the insidious potentials of AI. As the discourse continues to evolve, the Act serves as a foundational document in the quest to navigate the ethical implications of AI, ensuring that technology serves humanity without compromising individual autonomy or well-being. The legislative obligation to restrict the ethical boundaries of AI within the nascent AI Act has surfaced many ambiguities, particularly within Article 5. This article forms a labyrinthine part of the legislative framework, seeking to curb AI systems from exercising manipulative influence over individuals. Considering the recommendations for greater specificity and more comprehensive protection, Courts should find a way to interpret Article 5(1)(a) and (b) of the AI Act as follows:

Article 5(1)(a): The following artificial intelligence practices are to be prohibited:

(a) Placement on the market, placement into service, or use of any AI system that employs techniques that exert an influence on individuals at a subconscious level to substantially alter behaviour in a manner that is likely to cause or has the potential to cause, physical or psychological harm to the person concerned or to others. This includes, but is not limited to, any form of stimuli not consciously registered by an individual that can lead to such a behavioural distortion.

(b) The placement on the market, putting into service, or utilisation of any AI system purposefully designed to exploit the particular vulnerabilities of certain groups of individuals, identified by age, physical or mental capacity, or other distinct attributes, with the intention or effect of substantially altering the behaviour of an individual within such a group, in a way that is likely to cause or has the potential to cause, physical or psychological harm to the person concerned or to others. This prohibition includes all forms of exploitative strategies that may capitalise on an individual's suggestibility, cognitive biases, or any other psychological characteristics that could render them more susceptible to manipulation.

To further clarify and strengthen the Act, the following additional provisions are recommended:

Clarify Ambiguous Concepts: The European Parliament's Mandate refers to 'personality traits' in Art. 5(1)(b). To ensure that the Act's stipulations are enforceable, personality traits must be precisely defined, drawing from established psychological research and AI industry standards. A clear definition would enable regulators and AI practitioners to understand precisely which traits are subject to protection under the law, thereby enhancing the Act's practicality and effectiveness.

Expanding Protection to Encompass Broader Psychological Constructs: The Act's narrow focus on 'personality traits' may overlook other significant psychological constructs influencing susceptibility to AI. By broadening the language to include a broader range of 'psychological traits', like suggestibility and nudgeability, the Act would offer more robust protection against the manipulation of individuals by AI systems. This would address various cognitive and behavioural characteristics that AI could exploit.

Dissecting and Defining Manipulation Techniques: The prohibition of specific manipulative techniques by AI systems in the Act is too vague, necessitating the establishment of explicit definitions for:

- Subliminal Techniques: These should be characterised as any attempt to influence that bypass conscious awareness, including non-perceptible stimuli.
- *Purposeful Manipulative Techniques:* Techniques should include deliberate design choices intended to alter behaviour covertly.
- *Deception: This* requires a definition that covers AI systems that misrepresent users about their functionality or performance, including overt lies and subtle misrepresentations.

Addressing AI's Influence on Human Preferences: The Act should concern itself with AI's influence on immediate behaviours and consider how AI may shape and manipulate human preferences over time. The Act must include provisions against AI systems designed to alter human preferences in ways that could lead to harm, ensuring that AI cannot covertly influence individuals' values and decisions.

Conceptualising 'Informed Decisions': The notion of 'informed decisions' posited in the European Parliament's mandate under Article 5(1)(a) should be adopted and interpreted to mean decisions made with a complete understanding of all relevant information, outcomes, and alternatives without any distortions introduced by AI systems. This would ensure that individuals retain autonomy over their choices, free from the underhanded influence of AI.

These revisions aim to enhance the Act's precision, applicability, and enforceability, ensuring a balance between promoting AI technologies and safeguarding individual psychological integrity and autonomy.

VI. Conclusion

In the framework of the AI Act, Article 5(1)(a) and (b) target specific malpractices in AI, such as subliminal techniques and exploitation of vulnerable groups, drawing a line where AI systems must not tread. However, this delineation needs to fully capture the subtleties of AI influence, leaving room for questionable activities that fall just short of these prohibitions. This gap in the legislation calls for a concerted effort from both legal and technological sectors to refine the concept of 'beyond consciousness' and to solidify the definition of harm, ensuring that the autonomy of individuals is protected from the more elusive forms of AI persuasion that may not result in immediate, tangible harm.

The transition from recognising these gaps to formulating solutions sets the stage for the AI Act's next steps. The Act must address the nuanced complexities of AI and its potential to influence human behaviour and preferences subtly. It calls for a more precise articulation of what constitutes manipulation and how it can be ethically contained. Recommendations for the Act include defining ambiguous terms like ‘personality traits’, expanding the scope to cover a broader range of psychological influences, and establishing clear guidelines for what constitutes subliminal manipulation and deception. These recommendations aim to ensure that AI is developed and implemented in ways that respect human dignity and agency and that informed decisions in the digital space are made without covert AI interference.

As we approach the next phase of legislative scrutiny, it is imperative to translate these recommendations into actionable regulations. By doing so, lawmakers and technologists can contribute to a legal framework that acknowledges the current understanding of dark patterns and anticipates the complexities of emerging AI technologies. The challenge is to mitigate not just the known knowns, but to prepare for the known unknowns and unknown unknowns of AI advancements, ensuring a future where digital integrity and human rights are inextricably interwoven and protected.

To map the reforms proposed in the text onto the psychological techniques outlined in Table 2 above, each recommendation is aligned with the relevant technique and category of manipulation that it would most likely impact and how it might mitigate the potential material harm. *Table 3* shows how the proposed reforms would map onto the psychological techniques provided.

The proposed reforms are designed to bring more clarity and coverage within the AI Act, particularly in areas where psychological techniques could potentially harm users. The recommendations focus on defining ambiguous terms, expanding protections to cover a broader range of psychological influences, and ensuring that users can make informed decisions free from covert AI interference. By addressing these areas, the reforms aim to mitigate the harms associated with each psychological technique and enhance the ethical use of AI.

Currently, the narrative of the Act resembles the philosophical musings once articulated by US Secretary of Defence Donald Rumsfeld:

> 'There are known knowns. These are things that we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.'[61]

In the domain of digital design and AI architecture, these concepts are equally resonant. The ‘dark patterns’ represent our 'known known', the identifiable and understandable patterns. However, as we explore the intricate layers of these systems, we encounter 'known unknowns', the complex and covert mechanisms that we recognise but may not fully understand. Beyond these lies the 'unknown unknowns', the nascent yet undetected patterns or manipulations so profoundly embedded in the technological fabric that they escape our current understanding. As advocates for user rights, our legislators must not be complacent, addressing merely superficial ‘dark patterns’. They must plunge into the technological abyss to unearth both the known and the unknown, wielding an appropriate measure of precaution. The commitment should ensure that each design stratum is suffused with transparency and ethicality and upholds the sanctity of user dignity. This venture, although formidable, is indispensable to safeguard the integrity of our digital future.

[61] DoD News Briefing - Secretary Rumsfeld and Gen. Myers; Presenter: Secretary of Defense Donald H. Rumsfeld, February 12, 2002 <https://web.archive.org/web/20160406235718/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636> accessed 12 February 2024.

Table 3. *Proposed reforms*

Psychological Technique	Category of Manipulation	Potential Material Harm	Impact of the Proposed Reform
Subliminal Messaging	Sensory Manipulation	Impulsive behaviours, unhealthy consumption habits.	Clarify 'Subliminal Techniques'; Define 'Personality Traits'; Conceptualise 'Informed Decisions'
Background Audio	Sensory Manipulation	Mood alteration leading to anxiety or stress	Define Purposefully Manipulative Techniques'; Address Influence on Human Preferences
Colour Psychology	Sensory Manipulation	Environmental cues that cause stress or anxiety	Define 'Deception'; Expand Protection to Broader Psychological Constructs
Latency Manipulation	Algorithmic Manipulation	Delay in opt-out leads to privacy loss	Clarify 'Subliminal Techniques'; Conceptualise the 'informed decisions'
Temporal Manipulation	Algorithmic Manipulation	Altered perception of time causing impatience	Define Purposefully Manipulative Techniques'; Address Influence on Human Preferences
Manipulative Feedback	Behavioural Conditioning	Addiction to platforms; Impacting mental health	Define 'Deception'; Expand Protection to Broader Psychological Constructs
Hidden Repetition	Behavioural Conditioning	Subconscious conditioning leading to habit changes	Define 'Purposefully Manipulative Techniques'; Conceptualise 'Informed Decisions'
Dependency and Lock-In	System Dependency and Control	Reduced autonomy, over-reliance	Expand protection to Broader Psychological Constructs; Address influence on human preferences
Control over Updates	System Dependency and Control	Forced compliance, data sharing	Clarify 'Subliminal Techniques'; Conceptualise the 'informed decisions'
Information Asymmetry	Structural Manipulation	Financial losses, unfair treatment	Define 'Deception'; Expand Protection to Broader Psychological Constructs
Adaptive Algorithms	Algorithmic & Behavioural	Increased dependency, decisions benefiting service providers	Define 'Personality Traits'; Clarify 'Subliminal Techniques'; Address Influence on Human Preferences