Cambridge Forum on AI: Law and Governance (2025), **1**, e20, 1–28
doi:10.1017/cfl.2025.4

Cambridge Forum

RESEARCH ARTICLE

Cognitive freedom and legal accountability: Rethinking the EU AI act's theoretical approach to manipulative AI as unacceptable risk

Taimur Aimen

Tilburg Institute for Law, Technology and Society (TILT), Tilburg University, Tilburg, Netherlands
Email: a.taimur@tilburguniversity.edu

(Received 1 November 2024; revised 15 February 2025; accepted 3 March 2025)

Abstract

This paper examines the profound challenges posed by manipulative artificial intelligence (AI) and critically evaluates the adequacy of the EU AI Act in mitigating these threats. Modern AI technologies possess the capability to influence human cognition and behaviour imperceptibly, thus endangering cognitive freedom, the fundamental right to autonomous thought. Although the EU AI Act classifies manipulative AI as an unacceptable risk and prohibits its deployment, its current framework, characterized by imprecise definitions and regulatory gaps, undermines its efficacy in holding entities accountable and safeguarding individuals. To address these deficiencies, this paper introduces an innovative analytical method that traces the origins of manipulation, enabling a systematic understanding of the harm. Central to this discussion is the expanded concept of cognitive freedom, which transcends conventional notions of thought rights to encompass protection from covert digital influence. Through illustrative case studies, such as the use of psychographic profiling in political campaigns, the paper elucidates how data-driven methodologies can be harnessed to subtly mould public perception and decision-making. The analysis further investigates digital design strategies, including targeted advertising and algorithmic curation, which constrain user autonomy and erode independent judgment. The paper advocates for a restructured EU AI Act that incorporates precise definitions, mandatory transparency and continuous oversight by independent, multidisciplinary bodies. Such enhancements would strengthen the detection and regulation of manipulative AI practices. By embedding cognitive freedom within legal protections and proposing real-time audits and comprehensive ethical assessments, this paper outlines a strategic pathway for preserving cognitive autonomy. This approach aims to mitigate the erosion of mental sovereignty and uphold the essential principles of independent thought and informed decision-making within the rapidly evolving digital landscape.

Keywords: cognitive freedom; manipulative AI; digital autonomy; EU AI Act

1. Introduction

The human mind has always been the final refuge of personal freedom, a sanctuary for unspoken beliefs, private reflections, and the raw essence of individual autonomy. However, in an age where artificial intelligence (AI) can silently shape perceptions and influence decisions beneath the veil of awareness, this sanctuary is under unprecedented threat (Cohen, 2012). Today, AI systems are not merely tools but silent architects of behaviour, wielding the power to manipulate without detection and to alter thought without consent (Floridi, 2014). As these technologies evolve, the question is

no longer just about protecting our data or identities; it's about safeguarding the integrity of our very consciousness. The European Union AI Act (EU AI Act) finalized in June 2024, seeks to address these looming risks by defining and prohibiting 'manipulative AI' practices (Regulation (EU) 2024/1689, Art. 5(1)(a)). This legislative move is part of the European Commission's broader strategy to foster trustworthy AI aligned with European values (European Commission, n.d.).Yet, when legislation confronts the complexities of cognitive intrusion, can it truly hold the line against these invisible encroachments? This article embarks on a critical exploration of this issue, probing the fragile boundaries between innovation and exploitation, and challenging whether our current legal frameworks are fit to defend the most fundamental of human freedoms – the freedom to think and choose independently.

In the 2024 EU AI Act, the classification of threats posed by AI has been broadly framed through a four-tier risk assessment model. The threat of manipulative AI (AI Act 2024, OJ L 1689, Article 5) has been classified as an 'unacceptable risk.' In response to the detection of unacceptable risk, the Act stipulates that such practices must be explicitly 'prohibited' as a strict industry standard. However, realistically, the prescribed prohibition cannot be practised if clear boundaries around cognitive behavioural manipulation are not defined. This creates abstractness in attributing responsibility to a specific party once a delict has been committed. There is no distinct proveable charge with components that must be present for prosecution. The main legal challenge that this presents is the viability of the protection for an average citizen if they are the subject of targeted cognitive manipulation leading to behavioural modification. Moreover, the liability for tech companies and data analytic firms remains in obscurity, which leaves an ever-increasing margin for them to evade accountability as the methods of algorithmic manipulation become even more advanced and opaque over time (Franklin, Ashton, Gorman & Armstrong, 2022).

Another important consideration is that if the Act fails to establish an effective accountability framework for addressing what it identifies as the most severe category of risk, this raises broader concerns about its capacity to effectively regulate lesser, yet still significant, risks. But it must be noted that this Act is meant to be part of a larger regulatory package which is for the European Commission to fill in with Guidelines, Delegated Acts and Commission Decisions. Although the AI Act is meant to be a general framework, it presents issues of interpretation which need to be addressed for its efficacy in regulating and overseeing innovation in AI along with curbing ethical risks which can have negative consequences for society at large.

The necessary prerequisite for creating sufficient legislative protection is the foresight of adequate redressal and the capability to trace the realistic outlines of the requirements that the legislation is intended to plug. This purposive approach to the protection of 'cognitive freedom' is the optimal way forward since it will provide a dissection of the harm that the law aims to remedy. The term 'cognitive freedom' is intentionally chosen in this paper as a conceptual advancement that merges and extends the established notions of freedom of thought and cognitive liberty. While freedom of thought is a well-recognized human right enshrined in various international frameworks, such as Article 18 of the International Covenant on Civil and Political Rights, it predominantly addresses protection from external coercion in belief formation and expression. On the other hand, the term cognitive liberty, often championed in neuroethics, emphasizes the individual's autonomy over their mental processes and the right to avoid unwanted intrusions or alterations to their cognitive states (Bublitz, 2015).

However, as the digital landscape evolves with the proliferation of sophisticated AI systems, these traditional constructs may not adequately capture the multi-dimensional threats posed to mental autonomy and self-determination. Cognitive freedom is proposed here as a more comprehensive and contemporary framework that encompasses both the right to freedom of thought and the protection of mental integrity. Unlike cognitive liberty, which focuses primarily on individual autonomy concerning neuro-interventions or mental manipulation, cognitive freedom broadens the scope to include an individual's right to form thoughts and beliefs in an environment free from covert influence, manipulation, or exploitation by digital technologies and algorithmic systems. This expanded

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

conception better reflects the emergent ethical challenges posed by manipulative AI, algorithmic choice architecture, and behavioural profiling, which target not only the integrity of mental states but also the conditions under which cognitive processes unfold. Cognitive freedom thus captures the dual necessity of safeguarding internal cognitive sovereignty and ensuring that external influences remain transparent, accountable, and respectful of mental autonomy. By employing this term, the aim is to signal a more holistic approach to the issue, one that not only emphasizes the inviolability of thought but also acknowledges the broader socio-technical landscape that increasingly shapes cognitive experiences.

Consequently, the discourse around manipulative AI calls for a reconceptualization of cognitive freedom in the digital age. Beyond protecting individuals from coercion, it is crucial to consider how digital environments and AI systems can erode mental autonomy by exploiting vulnerabilities in human cognition (Zarsky, 2016). Addressing these issues requires not only regulatory measures but also a deeper philosophical understanding of how these technologies intersect with fundamental human rights to preserve the integrity of thought in an increasingly algorithmic world.

It is indicated via the literature on the area that there is a missing piece in all existing laws (Ienca & Ignatiadis, 2020) which two main arguments propose to solve i.e., the creation of new rights (neuro rights) (NeuroRights Foundation, n.d.; Ienca, 2021) or finding protections within the present human rights law to specifically maintain cognitive freedom as a form of legislative barrier against the impact of manipulative AI (Financial Times, 2023). However, before the merits of either of the new or suggested solutions can be considered, it is imperative to map out the precise scale of the problem.

2. Freedom of thought as a human rights-based safeguard

The right to freedom of thought, as articulated in Article 18 of the Universal Declaration of Human Rights (UDHR), serves as a cornerstone of individual autonomy, intricately linked to the rights of conscience and religious belief (United Nations, 1948, UDHR). At the heart of the Right to Freedom of Thought lies its emphasis on mental inviolability, safeguarding unexpressed thought, enclosed within the forum internum,[1] while offering only qualified protections to publicly articulated thoughts, situated in the forum externum (Stenlund & Slotte, 2018).

While the initial conceptualization of freedom of thought was often linked to protecting individuals from coercion and religious persecution, this view is historically limited. Freedom of thought possesses an intrinsic value beyond these early concerns. René Cassin, one of the principal drafters of the UDHR, emphasized that this right serves as the foundation for all fundamental freedoms, including freedom of expression and conscience (Cassin, 1948). Cassin notably described freedom of thought as 'the origin of all fundamental rights,' underlining its foundational importance in the human rights framework.

Additionally, legal scholars have argued that freedom of thought has been historically neglected, requiring renewed attention in contemporary legal discourse. Bublitz suggests that advances in neuroscience and AI-driven cognitive influence necessitate a re-evaluation of this right as an independent safeguard against technological intrusions into mental autonomy (Bublitz, 2015). The travaux préparatoires of the UDHR reveal that deliberations surrounding Article 18(1) intentionally positioned freedom of thought as an absolute right, distinct from other cognitive freedoms such as freedom of expression, which can be subject to limitations (Office of the High Commissioner for Human Rights [OHCHR], 2021).

The concept of freedom of thought is generally regarded as comprising three fundamental attributes that collectively uphold the integrity of individual thought (United Nations General Assembly, 2021). First, it includes the essential freedom to withhold one's thoughts from disclosure, ensuring that individuals can choose whether or not to share their innermost reflections and beliefs.

[1]The inner sanctum of the brain which holds unexpressed thoughts and ideas.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Second, it encompasses the right to be free from retribution or punitive measures based on one's thoughts, providing a crucial layer of protection for individuals to think freely without fear of consequence (United Nations General Assembly, 2021, para. 25). Lastly, freedom of thought entails the protection against unauthorized or impermissible alterations of thought, safeguarding the authenticity and originality of an individual's mental processes (United Nations General Assembly, 2021, para. 25). Together, these attributes form a robust framework that supports and nurtures intellectual autonomy and expression.

2.1. Scale of interpretation

Legal and philosophical interpretations of the right to freedom of thought can be categorized as being restricted, balanced, and expansive in their approach. The restricted interpretation contends that freedom of thought should be limited to core beliefs that fundamentally shape an individual's identity, such as religious or ideological convictions (United Nations General Assembly, 2021, para. 25). This perspective confines the protection of thought to deeply held beliefs, excluding more casual or transient thoughts.

The balanced perspective, gaining traction in ECtHR jurisprudence, broadens protection to thoughts that, while not restricted to religious beliefs, are significant in shaping an individual's worldview. This view recognizes that political ideologies, moral values, and similarly profound intellectual commitments should also be encompassed within the protective scope of freedom of thought, as long as they meet a threshold of coherence and seriousness (Ligthart, Bublitz, Douglas, Forsberg & Meynen, 2022). ECtHR rulings have indicated that certain expressions of thought, such as political intentions, remain firmly within the private, inviolable domain of the forum internum (Ligthart, Douglas, Bublitz, Kooijmans & Meynen, 2020). Conversely, the expansive interpretation, championed by scholars like Bublitz, advocates for a comprehensive application of freedom of thought that includes all cognitive processes – imagination, emotions, fantasies, and even trivial thoughts (Bublitz, 2020). Proponents argue that the essence of mental freedom lies in protecting the full spectrum of thought, without imposing arbitrary limits based on perceived significance (Bublitz, 2020). This perspective asserts that any external interference in thought processes constitutes a violation of mental autonomy (Alegre, 2021).

Having outlined the different interpretations of freedom of thought, it becomes crucial to explore how these perspectives align or conflict with modern technological realities. The subsequent section investigates the profound impact of AI on cognitive autonomy, examining how these interpretations hold up against AI's ability to subtly shape, influence, or even infringe upon an individual's mental processes.

2.2. AI and cognitive autonomy

At the heart of freedom of thought is the principle of cognitive autonomy, safeguarding an individual's capacity to think freely and without undue external influence. This principle becomes increasingly critical in light of technological advances that intrude upon individuals' mental privacy. The rapid evolution of neurotechnologies and data-driven behavioural manipulation has raised alarms about the erosion of cognitive autonomy in ways that the drafters of earlier human rights treaties could not have anticipated. Technologies capable of predicting, influencing, or even altering thought patterns, such as subliminal messaging or algorithmic nudging as will be discussed in the following sections, pose direct threats to the sanctity of cognitive freedom (Susser, Roessler & Nissenbaum, 2019).

This landscape necessitates a re-evaluation of freedom of thought in response to these emerging challenges. While the initial focus of freedom of thought was on protecting individuals from coercion and religious persecution, modern forms of thought interference demand more nuanced protections. Cognitive manipulation targeting the subconscious requires legal frameworks to adapt, ensuring that

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

protections against impermissible thought alteration encompass these novel intrusions (Susser *et al.*, 2019, p. 12).

While the right to freedom of thought is firmly established in international law, its dimensions must be reevaluated in light of emerging technologies and evolving societal landscapes. The scope of this right whether interpreted through a restricted, balanced, or expansive lens continues to be a subject of active discourse. Nonetheless, the fundamental principle remains unchanged: the right to think freely and independently is a cornerstone of human dignity and autonomy. As our understanding of cognitive processes deepens and external threats to cognitive autonomy evolve, the imperative to safeguard the forum internum from interference becomes increasingly important.

3. The philosophical interpretation of manipulation

Manipulation can be understood through three fundamental aspects (Barnhill, 2022). The first is the use of deceptive or unethical tactics, where manipulation often relies on hidden or unfair methods to control an individual or situation. This element of secrecy or dishonesty may indicate its morally questionable nature, though not all forms of manipulation are inherently unethical. In many cases, manipulation operates in such a way that the target remains unaware of the influence being exerted (Barnhill, 2022, p. 66). The second aspect highlights the skilful and strategic nature of manipulation. It requires careful planning and a refined approach, often reflected in practices like advertising, which employs subtle tactics to shape behaviour. The third and final aspect is the manipulator's intent, which is typically self-serving (Barnhill, 2022, pp. 67–72). Here, the primary goal is to achieve personal gain, with the manipulator leveraging control to their advantage. This element of intent underlines manipulation's inherent focus on benefiting the manipulator, often at the cost of the individual being influenced.

Ruth Faden and Tom Beauchamp, who introduced a framework known as the 'substantiality model' of autonomy, outline three essential conditions that must be met for an action to qualify as substantially autonomous: intentionality, understanding, and the absence of external control (Faden & Beauchamp, 1986). In their examination of the concept of non-control, Faden and Beauchamp categorize influences into three distinct types: coercion, manipulation, and persuasion (Faden & Beauchamp, 1986, p. 339). They define manipulation broadly as any intentional and effective influence exerted on an individual, which is achieved by non-coercively altering the choices available to that individual or by changing their perceptions of those choices without relying on persuasive techniques.

According to their analysis, an action is deemed entirely non-controlled when it is free from any attempts at influence; if such attempts occur, they should not impede the individual's capacity to make a free choice (Faden & Beauchamp, 1986, p. 258). Faden and Beauchamp maintain that persuasion typically lacks controlling features, while coercion is fundamentally characterized by control. In contrast, manipulation can manifest as either controlling or non-controlling, contingent upon the level of influence exerted. They argue that manipulative tactics that exert control violate autonomy, whereas those that do not impose control remain compatible with it. This distinction holds significant ethical implications regarding the acceptable limits of manipulation within the context of the right to freedom of thought, which aims to protect an individual's autonomy over their mental processes (Faden & Beauchamp, 1986, p. 261). However, their analysis reveals a critical shortcoming: a lack of explicit criteria to differentiate effectively between controlling and non-controlling forms of manipulation. Additionally, Cohen highlights that a major limitation of AI regulation lies in the inability to clearly delineate between permissible persuasion and impermissible manipulation, particularly in commercial and political contexts (Cohen, 2023). She argues that without specific behavioural impact assessments, AI-driven influence techniques may remain undetected under current legal standards. The AI Act's reliance on subjective harm thresholds further complicates enforcement, as not

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

all manipulative AI systems produce immediate psychological distress, but may still undermine long-term cognitive autonomy. A regulatory approach that accounts for the gradual nature of AI-driven manipulation is necessary to close this enforcement gap.

As an example of this compounding effect, Cytowic explores how sensory overload and digital distractions contribute to users' susceptibility to manipulation (Cytowic, 2024). He explains that AI-driven engagement tactics capitalize on neural vulnerabilities, leveraging attention-hijacking mechanisms that reduce users' ability to critically assess information. This has significant implications for regulatory discussions on AI manipulation, as the ability to detect and resist persuasive technologies is not solely a matter of informed consent but also cognitive capacity. Additionally, Cytowic argues that AI-driven engagement strategies exploit cognitive fatigue by overwhelming users with continuous stimuli, reducing their ability to resist persuasive techniques. As attention spans decline due to prolonged exposure to algorithmically curated content, users become more susceptible to decision-making biases, reinforcing engagement loops that prioritize compulsive interaction over informed choice (Cytowic, 2024). Regulatory efforts, therefore, need to account for how AI interacts with fundamental neurological processes, particularly in environments where constant exposure to AI-curated content affects decision-making abilities over time. Therefore, the following section will introduce a comprehensive technique through various examples to track how manipulation may occur and be identified in instances of non-consensual cognitive control.

4. The EU AI Act's theoretical framework: Article 5 and the prohibition of manipulative AI

The EU AI Act is among the most comprehensive attempts to regulate AI, seeking to address its potential to manipulate human cognition and behaviour. At its core, Article 5 designates manipulative AI practices as 'unacceptable risks' and prohibits their deployment under specific circumstances (AI Act 2024, OJ L 1689, Article 5(1)(a)). This section examines the theoretical underpinnings of Article 5, emphasizing its alignment with principles such as cognitive freedom and autonomy, while also reflecting on the Act's potential gaps and ambiguities.

4.1. Framing manipulative AI as an unacceptable risk

Article 5 prohibits AI systems that manipulate individuals through subliminal techniques, exploit vulnerabilities, or exert undue influence leading to significant harm (AI Act 2024, OJ L 1689, Recital 28). Recital (28) highlights the intrinsic incompatibility of such practices with EU values, including human dignity and democratic principles. By explicitly linking manipulation to unacceptable risks, the Act provides a theoretical framework rooted in the protection of mental integrity and freedom of thought.

However, the breadth of this definition raises critical questions. For example, what constitutes a 'subliminal technique,' and where should the line be drawn between permissible influence and impermissible manipulation?

The term 'subliminal' refers to stimuli that operate below the threshold of conscious awareness, meaning individuals do not consciously recognize these influences, yet they can still shape decision-making processes, behaviours, and emotions (Merikle, 2000). Early research into subliminal messaging, such as James Vicary's 1957 advertising study, claimed that undetectable visual cues could alter consumer behaviour. While initial claims were disputed, contemporary research has substantiated that subliminal stimuli can exert measurable, lasting effects on cognition . In the context of AI, concerns have emerged regarding AI-driven subliminal techniques that exploit human cognitive biases without explicit awareness (Brooks et al., 2012). A 2024 Forbes report examined how generative AI could deploy subliminal cues in digital content, raising ethical concerns about the covert manipulation of users (Eliot, 2024). Given the evolving risks, Article 5(1)(a) of the EU AI Act categorizes AI systems that employ subliminal techniques as presenting an unacceptable risk, thereby

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

prohibiting their use when they materially distort behaviour in a manner likely to cause harm (AI Act 2024, OJ L 1689, Article 5(1)(a)).

Moreover, AI systems differ fundamentally from traditional subliminal messaging techniques used in advertising or media. Unlike fixed subliminal cues, AI models continuously adapt based on real-time behavioural feedback, creating a dynamic, personalized form of cognitive manipulation (AI Act 2024, OJ L 1689, Article 5(1)(a)). This raises concerns about user autonomy and informed consent, as individuals cannot reasonably opt out of influence they do not consciously perceive (Zuiderveen Borgesius, 2018).

The Act's prohibition of 'exploitation of vulnerabilities' further invites inquiry: does it address only specific vulnerabilities, such as age or disability, or should it also extend to universal cognitive biases? These questions reveal the complexities inherent in translating broad ethical principles into enforceable legal standards.

4.2. Vulnerability, autonomy, and recital (29)

Recital (29) offers additional insight into the theoretical approach of the AI Act, emphasizing the covert nature of manipulative practices and their reliance on exploiting vulnerabilities (AI Act 2024, OJ L 1689, Recital 29). While this recital appears to focus on specific groups, such as children, the elderly, or those with mental disabilities, it also implicitly acknowledges the broader spectrum of cognitive vulnerabilities that manipulative AI may exploit.

Sax and Helberger further emphasize that vulnerability is not a fixed trait but a dynamic condition shaped by digital interactions (Sax & Helberger, 2024). AI recommendation systems, for example, do not simply respond to user preferences but actively shape them by reinforcing behavioural patterns over time. This highlights the regulatory gap in the AI Act, as it does not address the fact that AI systems can construct vulnerability rather than merely exploit it. On the same point, Fineman's concept of universal vulnerability highlights that all individuals possess cognitive limitations that AI can leverage, even in the absence of traditionally recognized vulnerabilities (Fineman, 2010).

This raises important theoretical considerations. If manipulation operates on universal cognitive tendencies, does the Act's focus on specific vulnerabilities inadvertently exclude significant harm? For instance, techniques like algorithmic nudging or dark patterns may not target identifiable groups but could still undermine autonomy on a widespread scale. How far the Act's theoretical commitment to cognitive freedom extends in such cases remains an open question, leaving room for further exploration.

4.3. The framing of harm

Another key aspect of Article 5's theoretical framework is its treatment of harm. The Act emphasizes the prohibition of manipulative AI that causes 'physical or psychological harm,' but it remains silent on less tangible forms of harm, such as societal or economic impacts (Fineman, 2010). This focus aligns with traditional legal approaches, which prioritize direct and measurable harm, but it may also limit the Act's ability to address the broader consequences of manipulative AI. For example, could an algorithmic system that subtly reshapes public opinion or creates echo chambers fall outside the scope of the Act's prohibitions? Such scenarios challenge the sufficiency of the Act's current framework and invite deeper reflection on how harm should be conceptualized in the age of AI.

These gaps suggest the need for further refinement, both in terms of legal clarity and theoretical scope. Should the Act explicitly address societal and economic harms, or would this broaden its remit beyond practical enforceability? Similarly, how can regulators balance the need for precise definitions with the dynamic and evolving nature of AI technologies? These questions remain central to the ongoing evolution of the EU AI Act and its ability to safeguard cognitive freedom. Global frameworks such as the UNESCO, Recommendation on the Ethics of Artificial Intelligence (2022) and

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

the Council of Europe Framework Convention on AI (2023) provide complementary perspectives that could address some of these unresolved gaps. However, the Recommendation can be critiqued for its broad and non-binding nature, which may limit its practical enforceability and impact on AI governance. Similarly, the Council of Europe Framework Convention establishes legally binding obligations to ensure that AI systems respect human rights, democracy, and the rule of law, providing clarity on issues like undue influence and manipulation (Council of Europe, 2023, *Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law*). However, it has faced criticism for being drafted without sufficient civil society involvement and for potentially excluding the private sector from its scope, which could undermine its effectiveness in regulating AI applications comprehensively. These critiques highlight the challenges in developing international frameworks that are both inclusive and effective in addressing the complex ethical and societal issues posed by AI technologies. Regardless of the critique, these frameworks could serve as valuable references for the EU AI Act, offering insights into how to balance precise definitions with the dynamic evolution of AI technologies and how to address societal and economic harms without overextending the Act's enforceability.

While Article 5 provides a substantial starting point for addressing manipulative AI, its theoretical underpinnings leave certain areas unresolved. For instance, the absence of clear definitions for key terms like 'subliminal techniques' and 'undue influence' complicates enforcement and accountability (Yeung, 2017). Furthermore, the Act's reliance on physical and psychological harm as the primary metrics for risk assessment may fail to capture the full range of harms associated with manipulation.

Veale and Borgesius note that many AI-driven nudging techniques are not immediately harmful in an observable way but gradually reshape user behaviour, decision-making autonomy, and even political and economic participation (Veale & Borgesius, 2021). They further highlight that Article 5 of the AI Act lacks clear criteria for distinguishing permissible personalization from impermissible manipulation, thus enabling companies to defend practices as merely enhancing user experience rather than exerting undue influence. Neuwirth expands on this, arguing that real-time behavioural data processing allows AI systems to iteratively refine persuasive tactics, making it increasingly difficult to detect and regulate manipulation under existing legal frameworks. Neuwirth's analysis of prohibited AI practices under Article 5 underscores the ambiguities in enforcement mechanisms, particularly regarding subliminal AI systems and predictive AI models (Neuwirth, 2023b). He notes that while explicitly deceptive AI practices are banned, the Act remains unclear on the status of AI systems that operate within legal but ethically questionable frameworks, such as hyper-personalized algorithmic persuasion. This gap creates a legal grey area where AI developers can design systems that functionally manipulate users without triggering strict regulatory penalties. Addressing this requires not only stronger definitional clarity but also proactive enforcement mechanisms that target the cumulative psychological effects of AI-driven behavioural influence.

Moreover, the question of intentionality looms large. The distinction between intentional and unintentional manipulation is crucial in assessing the regulatory adequacy of the EU AI Act. While manipulation is often associated with deliberate influence, many AI systems produce behavioural shifts as an unintended consequence of their optimization processes, raising concerns about whether regulatory frameworks should focus solely on intent or also account for the broader systemic effects of AI-driven influence (Cao, 2024).

4.4. Intentional and unintentional manipulation by AI systems

The distinction between intentional and unintentional manipulation is central to evaluating the regulatory reach of the EU AI Act. While manipulation is often presumed to be a deliberate act aimed at influencing individuals for a specific objective, modern AI systems complicate this assumption. AI-driven decision-making systems may engage in manipulative practices not through explicit intent, but rather as an unintended consequence of algorithmic optimization. Zhong *et al.* apply behavioural

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

economics and psychology to assess the effectiveness of Article 5's prohibitions on manipulative AI (Zhong, O'Neill & Hoffmann, 2024). Their research highlights that while the AI Act acknowledges the risks of AI-driven cognitive influence, its reliance on subjective harm assessments overlooks the ways in which AI systems exploit predictable cognitive biases. They argue that AI regulation should move beyond a harm-based framework toward a predictive risk-based model that evaluates AI manipulation not only by its intended design but also by its real-world cognitive effects. Incorporating behavioural insights into AI regulation could improve enforcement strategies by identifying patterns of AI influence that subtly shape decision-making without immediate psychological distress. This raises critical legal and ethical questions regarding the attribution of responsibility and the adequacy of existing legislative safeguards.

Philipp Hacker introduces a crucial differentiation between intentional targeting, where AI systems are deliberately designed to exploit cognitive or emotional vulnerabilities, and unintentional targeting, where such exploitation emerges as an inherent byproduct of AI's optimization functions (Hacker, 2021). In cases of intentional manipulation, AI systems deploy tactics akin to 'Emotional AI,' where user behaviours and psychological states are actively detected and leveraged for strategic influence(Hacker, 2021, p. 14). This is particularly prevalent in targeted advertising, where AI refines content delivery based on inferred user susceptibilities. Such practices align with classical definitions of manipulation, as they operate covertly and systematically to shape decisions in a manner advantageous to the manipulator (Hacker, 2021, p. 16).

In contrast, unintentional manipulation arises when AI models, through their learning processes, identify and capitalize on human biases without an explicit directive to do so (Hacker, 2021, p. 20). For instance, recommendation algorithms may prioritize engagement metrics above all else, thereby steering users toward content that exploits their psychological predispositions, even if no human developer explicitly sought to achieve this outcome. This presents a fundamental challenge: if an AI system amplifies vulnerabilities without conscious human intent, should its outputs still be classified as manipulative? Hacker argues that the employment of unconstrained AI models, without proper safeguards, should be sufficient to attribute responsibility to the deploying entity, even in cases where manipulation was not an overt design objective (Hacker, 2021, p. 22).

This debate has direct implications for Article 5 of the AI Act, which prohibits AI systems that manipulate human behaviour 'in a manner that causes significant harm' (AI Act 2024 OJ L 1689, Recital 29). However, the regulation does not explicitly differentiate between intentional and unintentional manipulation, leaving open the question of whether liability should extend to firms deploying AI systems that inadvertently engage in manipulative practices. The legislative silence on this issue is particularly relevant in the context of digital platforms, where personalization algorithms often nudge users toward predetermined behavioural patterns without an explicit manipulative aim.

The challenge also lies in defining manipulation in a manner that captures both its intentional and unintentional forms. If the standard of manipulation is tied strictly to intent, then a significant portion of AI-driven influence may fall outside legal scrutiny(Bruegel, 2022). Conversely, if manipulation includes all cases where an AI system significantly distorts decision-making, regardless of intent, then the regulation risks becoming overly broad, potentially encompassing benign personalization mechanisms. Striking a balance between these poles is essential to ensuring that the AI Act remains both enforceable and effective in mitigating the societal risks posed by manipulative AI.

By distinguishing between deliberate and incidental manipulation, regulatory frameworks can better tailor enforcement mechanisms, ensuring that actors deploying AI are held accountable not just for intentional manipulative practices but also for failures to prevent harmful algorithmic behaviours. This discussion sets the stage for the following section, which explores how AI-driven manipulation manifests in real-world digital environments and the extent to which existing regulatory safeguards are equipped to mitigate its harms. By examining Article 5 and its theoretical underpinnings, this section has sought to illuminate the EU AI Act's approach to manipulative AI while highlighting areas for

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

further reflection. The following section will build on this foundation by exploring how manipulation operates in practice, tracing its trajectory from subtle influence to overt control.

5. Mapping the threat of manipulation

To effectively understand and address the nuanced process of thought manipulation, it is essential to reverse-engineer the harm, tracing it back chronologically to its source. This approach draws from the English common law tradition of establishing causation, a principle that underpins fault-finding by mapping the chain of events leading to an outcome (Moore, 2023). The method of reverse engineering, rooted in the causation principle, provides a compelling framework for evaluating thought manipulation. It allows us to systematically trace back from the point of influence to the initial breach, thereby clarifying how each step contributes to the erosion of mental autonomy. This approach ensures that responsibility can be attributed with rigour and that any intervention or legal redress can be grounded in a comprehensive understanding of how manipulation is operationalized. By adopting this method, we affirm that identifying and addressing cognitive intrusions requires more than recognizing their existence; it demands a structured, cause-focused analysis that exposes the full trajectory of manipulation, making ethical and legal boundaries unmistakably clear.

The reverse-engineering model outlined here conceptualizes the five steps as both constitutive elements and sequential processes, depending on the context. Each step can function independently to erode mental autonomy but often operates as part of a broader chain leading to harm. For example, while the harvesting of attention may occur without subsequent steps, it often serves as a gateway to deeper cognitive intrusions, such as decoding unexpressed thoughts or deploying manipulative design tactics. This flexibility allows the model to accommodate varying scenarios of manipulation, from isolated incidents to comprehensive strategies designed to control thought processes. By acknowledging this dual role, the model underscores the interconnected nature of these steps while leaving room for cases where only one or some of the steps may be present.

The process begins by envisioning a chain where a manipulated thought, which may or may not result in subsequent action, represents the final link. While the sequence suggests a linear progression, the framework does not imply that every instance of manipulation must involve all five steps; instead, each step represents a potential point of entry or escalation in the process of cognitive intrusion. This chain's formation starts with the breach of the forum internum, the inviolable inner sanctum of the mind where thoughts and beliefs are formed. Access to this cognitive space can often begin as simply as harvesting attention, which provides a foothold for deeper engagement. This access is not benign; it enables the second step, decoding or 'reading' the unexpressed thoughts and opinions an individual holds. The precision of modern psychographic profiling and behavioural analysis demonstrates that this is not a theoretical risk but a tangible reality.

The next step is to analyze how manipulation is operationalized, particularly through digital design techniques embedded in online practices. Dark patterns and subliminal AI techniques exemplify covert digital manipulation, influencing behaviour below the threshold of conscious awareness. Dark patterns, as defined by Brignull, refer to deceptive design strategies that exploit cognitive biases, such as misleading opt-outs or pre-selected choices that subtly push users toward predetermined outcomes (Bignull, 2023). Subliminal influence, studied extensively by Poetzl, laid the foundation for understanding how stimuli beneath conscious perception can shape thought patterns and decision-making processes. Poetzl's 1927 research on subliminal perception, particularly in relation to dreams, demonstrated that individuals exposed to visual stimuli outside their conscious awareness could later recall fragments of these stimuli in their dream states, suggesting that the subconscious mind retains and processes information even when it is not immediately accessible to conscious awareness (Poetzl, 1917). His work contributed to later psychological theories on implicit learning and unconscious influence, illustrating how external cues could subtly guide human cognition without direct realization.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

This phenomenon was further explored by Packard, who examined its implications in advertising and mass communication, highlighting how commercial and political messaging could exploit subconscious biases (Packard, 1957). AI-driven recommendation systems refine these methods by reinforcing biases while limiting alternative perspectives. While the EU AI Act classifies subliminal manipulation as an unacceptable risk, it does not sufficiently address AI-driven dark patterns, leaving room for covert digital persuasion to persist under the guise of engagement optimization.

Before turning to nudging, it is necessary to recognise that manipulative AI extends beyond behavioural steering to imperceptible cognitive influence. For instance, recommender systems are designed to subtly 'nudge' (Sunstein, 2015), individuals toward specific choices, influencing behaviour without overt coercion. Beyond these systems, the deployment of dark patterns represents another significant practice, where manipulative design strategies trick individuals into making decisions, they might not consciously choose, thereby undermining genuine autonomy (Gray, Kou, Battles, Hoggatt & Toombs, 2018). Each of these steps, whether isolated or sequential, contributes to the cumulative erosion of mental autonomy and the potential for significant harm. The intent behind these manipulative practices often extends beyond simple behaviour modification; it can lead to deeper cognitive effects, such as the suppression of dissenting thoughts and the erosion of the natural human capacity for independent thinking, fostering conformity to predetermined narratives (Haggerty & Ericson, 2000). The most alarming consequence of such pervasive influence is the potential elimination of free thought, which could detach individuals from objective reality and compromise their ability to make independent decisions based on logic and facts (Pariser, 2011). Inevitably this systematic conditioning has the potential to stifle critical, alternative thinking and, in more extreme cases, could evolve into the criminalization of dissenting opinions, imposing penalties for thoughts or beliefs that diverge from an accepted norm. This scenario raises significant ethical and social concerns, signalling a shift from influencing behaviour to controlling and standardizing thought, a trajectory that threatens the foundational principles of cognitive freedom and personal autonomy. The following sections will delve into the progression of manipulation, starting with the initial accessing of the mind through attention harvesting, followed by reading unexpressed thoughts, the act of manipulation itself, the suppression of dissent, and culminating in the potential criminalization of alternative thinking.

5.1. Accessing

Over the years, in the field of advertising, the ability to convince consumers to make certain decisions has been developed extensively over various mediums and platforms. Naturally, this process cannot be exercised without first getting the attention of the targeted individual. The word 'attention' in this context needs to be interpreted via the commonly used biomedical explanation by the American Psychological Association i.e., 'a state in which cognitive resources are focused on certain aspects of the environment rather than on others and the central nervous system is in a state of readiness to respond to stimuli (American Psychological Association, n.d.).'

Article 5 of the EU AI Act establishes a legal foundation by prohibiting AI systems that manipulate human behaviour through subliminal techniques, exploitation of vulnerabilities, or undue influence leading to significant harm. This provision explicitly frames manipulative AI as an 'unacceptable risk,' reinforcing the principle that human cognitive autonomy should remain protected from covert AI-driven influence. However, the Act does not always provide clear guidance on enforcement, particularly concerning AI-driven attention harvesting and cognitive intrusion (Wilczyński, Mieleszczenko-Kowszewicz & Biecek, 2024). Given that attention is an instinctive human response to external stimuli, control over it can be practised by manoeuvring the magnitude of the stimulus which as a consequence generates proportional attention. There is psychological evidence which illustrates how supernormal stimuli (Barrett, 2010) are used by advertisers to not just garner attention but also trick the viewer into watching and relating to the content being communicated on a deeper level by

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

creating loops of mimicry (Hendlin, 2019). For example, the use of fast-paced, high-contrast audiovisuals in advertisements that exaggerate human expressions and reactions, creates an amplified version of reality that compels viewers to engage and mimic the emotions presented, thereby deepening their connection to the content. In addition to the exploitation of primitive human instincts(Stel, Mastop & Strick, 2011), the audiovisuals of advertisements are also geared towards producing supernormal stimuli to get higher levels of viewer attention (Tushnet, 2010).

The concept of 'attention economy' (Williams, 2018) has originated in tandem with the idea that the very act of grasping someone's attention (Tran, 2016) has economic value within itself, making the competition to do so even more fervent. This is such a common practice that it has been identified on an industrial scale, wherein harvesting attention and then selling that to advertisers as a product by 'attention merchants' is seen as a marketable service (Wu, 2016). A well-documented example of this is the Cambridge Analytica firm, where the service that was being provided was centred on the premise that the data of Facebook users would be scraped to create psychographic profiles to push advertisements and political campaigns intended for targeted behavioural manipulation according to each individual's biases and psychological vulnerabilities. Since having the monopoly over individual attention is vital to the survival of a consumer base, systematic techniques of harvesting user attention are the goal of most social media companies to increase engagement of users with the relevant online platforms because more screentime ensures attention and more of an opportunity to push online content, including advertisements which may eventually lead to purchases (Yeung, 2017).

Even though attention may be a relatively abstract commodity, due to the value it holds, attentional metrics have been created to measure the success with which attention can be harvested for capitalistic gain (van der Ploeg et al., 2017). Through the study of these metrics, the measure of the attention-harvesting approaches can be analysed and further fine-tuned to be even more efficient. To bring notice to the importance of attentional privacy, it is vital to look at the system of exploitation created by the Big Data barons, as suggested by Yeung (Yeung, 2017, p. 119). It has been stated that even though our use of social media and online platforms may present itself as an autonomous choice, 'hyper nudges' can influence our interaction to increase engagement with these platforms themselves. Further, it was identified that:

> '…Big Data-driven decision guidance techniques can be understood as a design-based instrument of control, operating as a potent form of 'nudge.' The algorithmic analysis of data patterns dynamically configure the targeted individual's choice environment in highly personalised ways, affecting individual users' behaviour and perceptions by subtly moulding the networked user's understanding of the surrounding world. Their distinctly manipulative, if not straightforwardly deceptive, qualities arise from deliberately exploiting systematic cognitive weaknesses which pervade human decision-making to channel behaviour in directions preferred by the choice architect' (Yeung, 2017, p. 20).

In a study by Matz *et al*., it was demonstrated that advertising specifically tailored to individuals' psychological profiles had a substantial impact on altering their behaviour, thereby underscoring the inherent risks of such manipulative techniques (Matz, Kosinski, Nave, & Stillwell, 2017). Their findings reveal the extent to which personalized psychological targeting can influence decision-making processes, raising important ethical concerns about the potential exploitation of cognitive vulnerabilities in consumer contexts. Legislatively, the Unfair Commercial Practices Directive (European Union, 2005, Directive 2005/29/EC) addresses aggressive commercial tactics that may coerce consumer decisions but lacks detailed provisions on whether attention harvesting and targeted advertising qualify as sufficiently aggressive for prosecution. While the UCPD prohibits misleading and aggressive commercial practices, its enforcement is largely reactive, relying on consumer complaints and national

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

authorities' interpretations, which often fail to capture the subtle yet pervasive nature of algorithmic manipulation (Hacker, 2021) Unlike traditional deceptive practices that involve overt coercion, manipulative digital design strategies such as dark patterns and hyper-personalized persuasion fall into a regulatory grey area, where the burden of proof remains on consumers to demonstrate harm rather than on companies to justify their practices (Dobber et al., 2019). Similarly, the Digital Services Act 2022, intended to curb excessive commercial exploitation through targeted ads, fails to establish a comprehensive framework (TwoBirds, 2023, European Union, 2021) for protecting attentional privacy, offering absolute protection only for vulnerable minors. While Article 28 of the DSA prohibits targeted advertising based on profiling for minors, no comparable safeguards exist for adults, leaving a regulatory void where behavioural manipulation remains legally permissible (European Union, 2022). Moreover, the DSA primarily addresses transparency obligations and due diligence for very large online platforms (VLOPs) but does not explicitly classify attentional capture and exploitation as inherently harmful practices, despite growing concerns from consumer protection organizations (EDPS, 2022). This aligns with broader issues within the AI Act, as the Digital Services Act, while acknowledging the manipulative potential of targeted ads post-attention capture and exploitation for minors, omits comprehensive protections for broader demographics. This legislative gap results in an inherently exploitative digital marketplace (Bignull, 2023), reinforcing an asymmetrical power dynamic that benefits dominant commercial marketers over consumers (Helberger, Sax, Strycharz & Micklitz, 2022).

Aylsworth's exploration of autonomy and manipulation within the framework of persuasive advertising, though not directly linked to the right to freedom of thought, introduces a conceptual criterion that could be pertinent to this analysis. He asserts that in evaluating whether an influence constitutes manipulation, it is essential to consider whether the individual, upon critical reflection, would repudiate the desire induced by the manipulative process (Aylsworth, 2022). Aylsworth contends that manipulation occurs when the person if made fully aware of the mechanisms through which their desire was shaped, would reject it. This approach underscores the role of reflective self-awareness in determining the extent to which autonomy has been compromised, particularly when individuals are subjected to covert or persuasive techniques that subtly capture their attention and force limits on their capacity for informed decision-making (Aylsworth, 2022, p. 696).

In light of the concerns over the monopolisation of individual attentional privacy, two major sets of literary opinions exist. These are that either the control over individual attention should be protected as a negative right which would require those who exploit attention to refrain from doing so or a positive right (Puri, 2021) originating from the rights to privacy and bodily integrity (Chomanski, 2023). The policy response and a detailed outline of what a workable right to attentional privacy could be is a complex endeavour which may be difficult to create while balancing the commercial interests of advertisers and consumers. Furthermore, a similar concern has already been historically observed in the 1952 American case of Public Utilities Commission v. Pollak (Pub. Util. Comm'n of D.C. v. Pollak, 1952) wherein the dissenting opinion highlighted how forceful listening to another's ideas could prove to be a dangerous and effective weapon for propagandists. However, further commentary on this judgement elaborated on the possibility that a right to attentional privacy could have a destructive outcome and derail the entire infrastructure of commercial advertisements (Russo, 2009). Hence, the accessing of attention is an identifiable frontier that is currently unprotected which is the starting point of the deprivation of individuals over what may, consequently, dominate their thoughts.

The example of the accessing of the mind has been illustrated through the example of the breach of attentional privacy. In the context of social media, this may be practised via the use of feed curation, push notifications or aggressive advertisement techniques, where the user may feel compelled to engage further with each platform (Future of Privacy Forum, 2022). This is interlinked with the next step in the process of thought manipulation which is the reading of the thoughts to

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

further capture attention while using customized techniques, tailored to the vulnerabilities of each user.

5.2. Reading

The second step within the chain of manipulation begins after access to the inner sanctuary of the mind is secured. At this stage, data mining is initiated, capturing detailed records of an individual's predispositions, preferences, vulnerabilities, and cognitive biases (Cabena, Hadjinian, Stadler, Verhees & Zanasi, 1998). This process involves tracking online behaviour, interactions, and subtle digital cues that reveal personal inclinations (Wojnarska-Krajewska, 2021). The collected data is then aggregated and processed, transforming these fragments into comprehensive profiles. This form of datafication turns unaware individuals into detailed datasets primed for analysis.

These profiles are subsequently used for psychographic profiling, enabling manipulators to predict, influence, and guide thoughts and behaviours with tailored strategies, all without the individual's conscious awareness (Quach & Lee, 2021). Botes appears to concur with this perspective, emphasizing that the manipulative nature of persuasive technology hinges on whether 'the algorithm is exploiting some form of psychological, emotional, or behavioural vulnerability or weakness of the user for its subjective purpose, which may not align with the values and beliefs of the user' (Botes, 2023). This assertion highlights the significance of the algorithm's intent and its potential misalignment with the user's principles, further elucidating the complexities of manipulation in digital contexts.

Traditionally, psychographic profiling was used to explore individual behaviours in the context of collecting marketing data and creating a correlation with everyday practices to predict the demand and supply of products crudely for commercial purposes. The input of this data was either statistical data by observing purchasing trends or generalised market analysis through a competitional landscape (Solomon, 2004). However, the scale of data collection after the popularisation of social media and the digitising of databases has resulted amount of data being held by the digital spaces increasing manifold.

Even though in present times, open source technology has advanced enough to provide political consultancies with the ability to use fake audios for manipulating public opinion (Associated Press, 2024), historically, the UK-based consulting firm, Cambridge Analytica, is the most coherent case study to assess the extent to which individuals could be 'read' through one of the early forms of mass psychographic profiling done to sway political opinion. Prior to the legal notice, Cambridge Analytica was hired during the 2016 US elections as a consulting firm for the campaigns of two prominent Republican candidates (Smith, 2018). To ensure that the political messaging was effective, the firm resorted to using data science methodologies to find the exact segment of society to target. The target group was identified to be mostly Trump supporters or swing voters who still presented the opportunity to be influenced to vote for the Republican party (Lewis & Hilder, 2018). The firm primarily used Facebook to harvest data for accurate, targeted political campaigning. Reportedly, the data used to develop psychographic profiles was extracted through another app called 'This Is Your Digital Life App' (Hern & Cadwalladr, 2018) which linked the Facebook profiles of the users and their friends after the downloading of the said app and acted as a bridge for data scraping from Facebook databases. Upon the publicization of the leak, Facebook released a statement creating distance from the leak and upgrading its security policy (Facebook Newsroom, 2018).

The US Federal Trade Commission identified both Cambridge Analytica and the creator of the app, Kogan, (University of Cambridge) as the main actors behind the privacy breach of 'approximately 250,000–270,000 Facebook users who directly interacted with the app, as well as 50–65 million of the "friends" in those users' social networks.' (Federal Trade Commission, 2018). The Federal Trade Commission further found that the actual harvestation of the datasets was done vastly through

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

likes on Facebook posts and this activity was then analyzed via the OCEAN scale, 'a psychometric model that measures an individual's openness to experiences, conscientiousness, extraversion, agreeableness, and neuroticism' (Federal Trade Commission, 2018, p. 3, para. 8). These findings were then cross-referenced with data obtained through another survey communicated through the app, which asked questions about individual political allegiances and emotional responses to certain topics to train the algorithm to produce a precise output which would be able to show choice-focused projections of the voter pool.

During the testimonies, it was stated that the amount of data collected and the intended goal of behavioural micro-targeting through the algorithm was perhaps unrealistic due to the insufficiency of the available training data and the inaccuracy of the algorithm being used (United States Senate Committee on the Judiciary, 2015, *Testimony of Eitan D. Hersh*). In his expert testimony, Hersh claimed that the data used by Cambridge Analytica did not effectively influence the outcome of the election and that the hype around the scale of behavioural manipulation was blown out of proportion in mainstream media. He explained that;

> "Commercial data – such as information about purchasing habits or leisure interests – can also help campaigns with mobilization, but their use in the past has been limited. In Hacking the Electorate, I found that commercial data did not turn out to be very useful to campaigns. Even while campaigns touted the hundreds or thousands of data points they had on individuals, campaigns' predictive models did not rely very much on these fields. Relative to information like age, gender, race, and party affiliation, commercial measures of product preferences did not add very much explanatory power about Americans' voting behaviour" (United States Senate Committee on the Judiciary, 2015, *Testimony of Eitan D. Hersh*).

To understand how the data obtained via the Facebook databases was 'read,' it is important to know the methodologies used to generate psychographic profiles. This can be done by looking at the data mining techniques, which rely on the algorithmic conversion of large collections of data into specific useful outputs. Some of the popular types of data, mining techniques are association rules, classification, clustering, decision trees, K-nearest neighbour, neural networks and predictive analysis (Investopedia, n.d.). From the evidence in the testimonies during the legal inquiry and the expert opinion, the techniques most likely to be used after the first stage of data acquisition were classification, clustering and general organization. In the second stage, to assess the political orientation, the data points were collated and converted into profiles that when put through a predictive analysis could give an accurate presumption of how the individuals would vote in the upcoming election (Associated Press, 2024). This would logically also narrow down the voters who were unsure about which party to vote for, making them the perfect audience for targeted advertisements (Bangor University, 2020).

Referring to his prior work on the area, Hersh even went in so far as to say that such profiling could be used to mobilize and persuade potential voters and that such campaigning has been observed in different forms before. He further diluted the threat by mentioning that users engage with Facebook as political hobbyists and Facebook should not be treated as a platform that can be held responsible in a way that an official publication with editors would be. Similarly, in his written statement submitted to the House of Commons, Kogan stated that his algorithmic assessment of personality assessment was not accurate enough but continued to point out that '..the Facebook ads platform provides tools and capability to run targeted ads with little need for our work – in fact, the platform's tools provide companies a far more effective pathway to target people based on their personalities than using scores from users from our work' (Kogan (2018). Written evidence submitted by Aleksandr Kogan).

In assessing the implication of data mining upon opinion, it has been discussed that such an ability can easily be used to create an 'autonomy trap' (Zarsky, 2002-2003) where individuals can be secretly

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

pushed towards making a choice. So far back as 2000, the United States Federal Trade Commission cautioned Congress against the impacts of online profiling by illustrating how 'detailed profiles' (Federal Trade Commission, 2000) can be used to predict consumer behaviour while highlighting the lack of consent in the process along the following terms:

> "The most consistent and significant concern expressed about profiling is that it is conducted without consumers' knowledge. The presence and identity of a network advertiser on a particular site, the placement of a cookie on the consumer's computer, the tracking of the consumer's movements, and the targeting of ads are simply invisible in most cases" (Federal Trade Commission, 2000, p. 10).

Additionally, in the context of freedom of thought, within its jurisprudence, ECHR has already held that even the intention to vote for a particular political party constitutes a mental process inherently situated within the private realm of an individual's internal cognition (Russian Conservative Party of Entrepreneurs and Others v. Russia, 2007, para. 76; Georgian Labour Party v. Georgia, 2008, para. 120). This distinction reinforces the notion that personal political preferences, such as voting intentions, are protected as part of the inviolable sphere of thought, safeguarded from external interference or undue influence. By characterizing these decisions as confined to the internal domain, the Court emphasizes the paramount importance of protecting the autonomy of thought regarding democratic participation. Even though very collection and the hidden processing of data is a violation of data protection and privacy principles, this goes to show that the implications of Facebook's behavioural modelling cannot be understated especially after the last claim in Kogan's House of Commons testimony which can be a hint toward the technological potential of behavioural manipulation.

5.3. The act of manipulation

Within the chain of causation, the most technical and layered element of the offence is manipulation itself since it operates with secrecy and lack of express consent. The current version of the EU AI Act prohibits manipulation and textually defines manipulative AI as 'AI systems that deploy harmful manipulative subliminal techniques' (AI Act 2024, OJ L 1689, Article 5(1)(a)). While the AI Act prohibits AI systems that manipulate users through subliminal techniques or exploit cognitive vulnerabilities, its provisions remain vague, leaving open questions about how manipulation is identified and assessed in practice. Smuha *et al.* argue that without a precise framework distinguishing permissible persuasion from impermissible manipulation, enforcement of the AI Act's prohibitions will remain highly subjective and inconsistent (Smuha *et al.*, 2021). The current broad prohibitions in Article 5 fail to specify the degree of influence required for AI systems to qualify as manipulative, potentially allowing subtle yet effective persuasion techniques to persist under the guise of engagement optimization.

AI models optimized for engagement often influence user behaviour in ways that extend beyond their intended functionality. Furthermore, Smuha *et al.* also notes that AI models optimized for engagement may unintentionally manipulate users by reinforcing biases and shaping behaviour in ways that were not explicitly programmed by developers. Similarly, Veale and Borgesius add that Article 5 does not establish clear criteria for when AI-driven personalization becomes manipulation, leaving room for firms to argue that they are merely enhancing user experience rather than exerting undue influence (Veale & Borgesius, 2021). Neuwirth proposes a tiered regulatory model that assesses AI influence based on its level of cognitive impact, arguing that AI models with significant behavioural effects should be subject to heightened scrutiny, regardless of intent (Neuwirth, 2023b). While this approach acknowledges the growing risks of manipulative AI, it risks overregulating systems that shape behaviour without deceptive or coercive intent. By assessing impact without

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

accounting for intent, context, or proportionality, Neuwirth's model does not adequately distinguish between ethical persuasion and illegitimate manipulation.

The AI Act's limitations in defining manipulation, addressing vulnerability, and distinguishing between deliberate and incidental influence create a significant regulatory gap. Strengthening its provisions requires moving beyond general prohibitions and developing concrete risk assessment mechanisms to evaluate the broader implications of AI-driven behavioural manipulation. Susser *et al*. have thoroughly investigated the phenomenon of online manipulation, identifying three primary attributes that define manipulative practices. They assert that such practices are characterized by their: (1) hidden nature, (2) exploitation of cognitive, emotional, or other vulnerabilities in decision-making, and (3) targeted approach (Susser *et al*., 2019, p. 26). A key point in their analysis is the emphasis on the covert aspect of online manipulation, which significantly obstructs the manipulatee's capacity for conscious awareness of the manipulative tactics at play. They argue that this hidden quality is essential for the very establishment of manipulation, as it underpins the effectiveness of these practices. Moreover, Susser *et al*. highlight that the exploitation of vulnerabilities, along with the targeted nature of manipulation, serves as potential mechanisms for enacting manipulative practices (Susser *et al*., 2019, p. 27). They suggest that these factors can exacerbate the severity of manipulation, functioning as aggravating elements in the overall assessment of such actions with a clear power asymmetry between the manipulator and the manipulatee Fineman presents a compelling vulnerability theory, positing that all individuals are fundamentally due to their embodiment, a notion referred to as universal vulnerability. This theory acknowledges that while every person is perpetually susceptible to harm, the extent of these vulnerabilities varies based on individual circumstances, which she describes as a particular vulnerability (Fineman, 2010).

In the context of the digital landscape, Sax and Helberger argue that vulnerability is not a fixed state. Instead, engagement in digital interactions can lead individuals to fluctuate between different levels of vulnerability. They define digital vulnerability as 'a universal state of defenselessness and susceptibility to (the exploitation of) power imbalances that result from the increasing automation of commerce, datafied consumer-seller relations, and the inherent architecture of digital marketplaces.' They contend that this form of vulnerability is the norm rather than an exception (Helberger *et al*., 2022). When considering online interactions, it is reasonable to assume a pervasive state of vulnerability due to the significant power disparities between digital platforms and users, coupled with the fluid nature of these vulnerabilities, which can range from static to contextual (Sax & Helberger, 2024, p. 11, European Commission, 2021). Although all individuals exhibit vulnerability in online environments, the degree of this vulnerability is influenced by various factors, including age, cognitive capacity, educational background, digital literacy, and social status.

However, manipulation does not always require the presence of an identifiable vulnerability; rather, it may function independently as a coercive influence that alters decision-making processes without overtly exploiting cognitive or emotional weaknesses. While vulnerabilities whether universal or contextual, may increase susceptibility to manipulative practices, manipulation itself is a broader phenomenon that can operate even in the absence of traditional markers of vulnerability. The AI Act acknowledges this duality through two distinct prohibitions under Article 5: first, the ban on AI systems that deploy subliminal manipulative techniques (AI Act 2024, OJ L 1689, Recital 29) and second, the prohibition against AI systems that exploit vulnerabilities arising from physical or mental disabilities (AI Act 2024, OJ L 1689, Recital 30). This differentiation suggests that manipulation and vulnerability, while often interconnected, are not necessarily dependent on one another. Subliminal manipulation, for example, may affect individuals regardless of pre-existing vulnerabilities, whereas targeted exploitation presupposes an asymmetry between the manipulator and the manipulatee based on an identified weakness (Smuha, 2023). Understanding this distinction is essential in assessing the scope of legal protections under the AI Act and determining whether its current provisions are sufficient to mitigate both general manipulative risks and targeted exploitations.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

An important legislative critique is that Article 5 of the EU AI Act prohibits AI systems from exploiting specific vulnerable groups, such as individuals with physical or mental disabilities. This could be interpreted as a safeguard against tailored manipulation through psychometric profiling, assuming that such profiling can expose personal biases and vulnerabilities. However, the extent to which this provision applies to AI-driven profiling remains unclear, as the Act does not explicitly address this intersection. This reasoning has not yet been adopted, and the extension of the prohibition might not even be a preferred approach due to the widely accepted manipulative marketing strategies (Petropoulos, 2022), but it would eliminate opportunities for individual exploitation regardless of the intensity of vulnerability that one may pose. Both the UNESCO Recommendation and the Council of Europe Framework Convention explicitly address the risks of manipulative AI systems. UNESCO emphasizes the importance of protecting cognitive liberty and mental autonomy against covert AI-driven influence, while the Council of Europe highlights the need for transparency and accountability in AI systems to prevent deceptive practices. These frameworks reinforce the ethical imperative of safeguarding individuals from AI systems designed to exploit cognitive vulnerabilities.

The commentary on the AI Act has also identified potential gaps in its protection against subliminal manipulative AI. Feedback from the European Consumer Organization noted concerns about whether the Act sufficiently addresses the wide-ranging damages that manipulative AI could cause (Neuwirth, 2023a, p. 144). While Article 5 prohibits AI practices leading to physical and psychological harm, it is less explicit about societal and economic harms, which remain underexplored in the text. Recitals (28)–(31) hint at broader implications, such as threats to public discourse and democratic integrity, yet these are not directly addressed in the enforceable provisions. However, the harm caused by manipulation has had cross-sectoral harms as seen in the case study of the Instagram algorithm worsening body image issues and the heightening of anxiety and depression levels amongst teen girls to produce more user engagement (Wells, Horwitz & Seetharaman, 2021). Burr and Cristianini, in their 2019 study, propose that beyond the capabilities of diagnosis, prediction, and persuasion, there is the potential to exert control over psychological traits and mental states, such as influencing conditions like depression. This issue is highly relevant to the right to freedom of thought. They cite a 2014 experiment on Facebook, which revealed that users' emotions could be manipulated by adjusting the positive or negative tone of content in their newsfeeds. This experiment underscores the ability of digital platforms to control cognitive and emotional processes, raising critical concerns about safeguarding mental autonomy in the context of freedom of thought though the given instance can be limited to simple psychological harm, it does have a larger societal impact, ranging from the setting of unrealistic beauty standards to the development of popular culture (Burr & Cristianini, 2019).

To pin down manipulation, a few additional factors must also be identified (Noggle, 2020). These features have been noted to be

> "non-rational influence where the manipulator tries to bypass or weaken a person's deliberative decision-making capacities. Another is that manipulation requires the use of trickery and deception, often through hidden means, to get someone to behave in a certain way. The third is that it entails using some degree of pressure to do as the manipulator wants, for example, through emotional blackmailing. Lastly, it is generally not guided by the target's interests, goals and preferences, but only the manipulator's" (Future of Life Institute, 2022).

One of the more easily identifiable manifestations of manipulation is the direction of an action based on the choice architecture which nudges an individual to make certain choices (Thaler, Sunstein & Balz, 2010) in the digital environment because of the very format of the user interaction, as arranged by the creator of that environment. This establishes a clear power asymmetry. Dark patterns present a significant form of digital manipulation that operates both overtly and subliminally. These deceptive design techniques intentionally guide users toward making decisions they might not have otherwise

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

chosen, often by exploiting cognitive biases and behavioural tendencies. Herman argues that the EU's regulatory response to dark patterns, as seen in recent legislative efforts, remains insufficient in curbing their pervasive use, particularly in commercial and algorithmic environments (Herman, 2024). The Digital Services Act acknowledges dark patterns in Recital 67, emphasizing the need for robust safeguards against deceptive interface designs that distort user autonomy. While dark patterns share conceptual similarities with nudging, they differ in their intent, whereas nudges may steer decisions subtly but transparently, dark patterns deliberately obscure alternatives, leading users toward outcomes that serve the interests of the manipulator. Bongard highlights how the intersection of dark patterns and AI-driven recommendation systems exacerbates this issue, reinforcing engagement loops that subtly influence consumer behaviour (Bongard-Blanchy, Rossi, Bernhaupt, Lallemand & Sauer, 2021). Given their impact, dark patterns should not merely be an adjacent issue in discussions of manipulation but rather a central concern when assessing AI's role in shaping decision-making environments. By addressing dark patterns explicitly within the AI Act, regulators could bridge a significant gap in existing frameworks that fail to fully account for covert, design-based manipulation techniques. Similarly, this nudging may also be done through targeted behavioural advertisements and algorithmic echo chambers created via recommender systems which restrict those targeted with the free flow of information for the formulation of their independent opinions and perception of reality (Pazzanese, 2017).

An example of choice architecture is dark patterns, which is a manipulative user interface, created to trick individuals into nonconsensual activities. In response to consumer awareness and concern of dark patterns, it was found that people can have 'dark pattern-blindness' which obscures their ability to see the manipulative design that they are subjected to Bongard-Blanchy *et al.*, 2021) Additionally, it is believed that coercion is not a concealed influence, whereas manipulation is, and that nudges only have an effect on people when they are not aware of the influence being applied to them (Feinberg, 1986). Significant differences exist between the suggested designs according to the study's findings on dark pattern identification. Most users identified the high-demand/limited-time message and 'confirm-shaming,' but few identified the pre-selection nudge, coerced permission, and dark patterns based on deception techniques (e.g., trick questions, loss-gain framing, and hidden information). These results may imply that some dark patterns are inherently harder to identify, even while they only address a particular use of the dark pattern and cannot be applied to the category (Feinberg, 1986, page 126). Considering how manipulative practices in digital spaces have advanced and the lack of explicit regulation leads to the question of whether or not consent can be extracted from the user without their freedom to develop well-founded, explicit permission. This means that since individuals in digital spaces are already being manipulated and pushed into making choices that they may not even be aware of, it is a suppression of their ability to independently make decisions.

5.4. Suppression

As already established in the previous sections, manipulative practices and design create an environment which baits users to either not pay heed to what they are doing due to being subconsciously guided onwards making a certain decision or making a decision which may have already been implanted in their minds due to hyper exposure, subliminal messaging or the digital choice architecture (Thaler *et al.*, 2010). Following the steering of thoughts and decisions, suppression can be portrayed as of negative activity which is the elimination of an original thought, dissenting opinions or critical thought. The format that choice architectures employ to enforce the suppression can be witnessed by the employment of upload filters as a method for content moderation (Cobbe, 2021). Originally, these filters are presented as a protective mechanism to uphold community guidelines (Simon, 2020), however in reality they can be used as a form of suppression too. For example, a platform user may be not allowed to upload certain content from their profiles which may feature certain restricted content, topics or ideas.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Another under-recognized form of thought suppression can be the use of echo chambers which tend to present content in a way that the consumer may be enabled into believing something as being an objective truth while being oblivious to alternate knowledge (Jamieson & Cappella, 2008). This false formation of thoughts interferes with the ability to independently decide what the truth may be for everyone. Once the false formation of thoughts becomes commonplace, the power of determining one's agreement or dissent with an ideology or the situation can eventually cause the blockade of original critical thinking.

Suppression of free thought can have dire consequences because modern societies are built around the idea of autonomy and independent decision-making. Since it has already been established that the extraction of fundamental and intimate data is being done by large online platforms and used to run 'surveillance capitalism' (Zuboff & Schwandt, 2019), this system can be used not just for ulterior commercial purposes but also a systematic large-scale suppression of independent cognition. The breach of cognitive freedom through the suppression of free thought runs the risk of being a political project as much as a capitalist project with the suppression of thoughts being used to eliminate the power of making autonomous decisions and evaluating important choices such as voting or purchasing certain items(Metzinger, 2013).

5.5. Criminalization

The concept of criminalizing thoughts, famously explored in George Orwell's novel 1984, presents a cautionary tale where the suppression of dissent is taken to an extreme. In this fictional society, the ruling party penalizes unspoken ideas to prevent opposition (Orwell, 1949). While this may seem dramatic, there are real-world developments that hint at the potential for such practices.

For instance, in China, EEG headbands have been used in classrooms to monitor students' concentration levels. Although not an outright example of manipulative AI, it demonstrates how technology can assess cognitive states and track thought processes (Chen *et al*., 2023; Independent, 2019).

China's approach to AI regulation extends beyond experimental technologies such as EEG headbands and includes a structured legal framework to oversee the development and deployment of AI-driven cognitive systems. The Algorithm Recommendation Regulation, (Cyberspace Administration of China, 2022, March 1) which came into force in March 2022, mandates transparency in AI-generated recommendations and places restrictions on algorithmic practices that could induce addiction or excessive reliance on automated decision-making. In addition, the Deep Synthesis Regulation, implemented in January 2023, imposes requirements on AI-generated content, ensuring that synthetic media is clearly labelled and preventing deceptive applications of generative AI (Latham & Watkins, LLP, 2023). The Generative AI Regulation, which took effect in August 2023, further refines China's governance approach by addressing the ethical implications of AI models capable of influencing user perception and behaviour. These regulatory efforts highlight China's recognition of the risks associated with AI-driven cognitive manipulation and the need for safeguards to prevent unintended or intentional misuse. However, they also raise concerns about potential overreach, particularly if regulatory frameworks are leveraged not only for consumer protection but also for state control over information and behavioural influence. With advancements in Brain–Computer Interfaces (BCIs) incorporating brain coding capabilities, (Drew, 2023) there have been warnings that the neuro data generated could potentially be used for sophisticated tracking of unexpressed emotions (Farahany, 2023). This raises questions about where such capabilities could lead if not regulated, particularly if states decide to leverage these technologies to preemptively suppress dissent. The bridge from private manipulation to state-led criminalization is not as improbable as it might seem (Morris, 1965). While in Europe, the separation between market-driven manipulation and state control is more pronounced, the crossover potential exists when governments adopt technologies developed in the private sector for state purposes (Floridi & Cowls, 2019). This could blur the lines between influencing behaviour and policing thought.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Beyond privacy concerns, the ability to think freely is vital for creativity, innovation, and intellectual growth. Predictive policing refers to the use of AI to forecast criminal activity before it occurs (RAND Corporation, 2013). While traditional methods rely on historical crime data and geospatial analysis to predict high-risk areas, modern approaches increasingly focus on individual-level risk assessment through non-invasive digital tracking techniques. Unlike clinical neuro prediction tools that analyze brain activity, the focus here is on non-clinical and non-invasive methods, narrowing the scope to the analysis of social media activity, browsing patterns, and behavioural biometrics to infer the likelihood of criminal behaviour.

One widely used technique is sentiment analysis and keyword tracking on social platforms, where AI systems monitor online discourse to detect early signs of radicalization, cybercriminal intent, or violent extremism. These models assess language patterns, engagement with specific communities, and changes in discourse that align with known risk factors for criminal behaviour. AI-driven models can flag users who show progressive shifts in rhetoric, allowing law enforcement agencies to intervene preemptively. However, critics highlight concerns over false positives, the erosion of online privacy, and potential biases embedded in AI algorithms (Engemann, 2024, November 20).

Another method involves behavioural biometrics, which tracks user-specific digital habits such as keystroke dynamics, mouse movement patterns, and browsing behaviours (Chén, 2022, July 1). These data points can be used to develop individual behavioural profiles, detecting deviations that might indicate fraudulent activity, hacking attempts, or preparation for cybercrime (Berk, 2021). Unlike sentiment analysis, which relies on linguistic data, behavioural biometrics focuses on the subconscious digital behaviours of users, making it more difficult for individuals to intentionally mask their activity.

Finally, browsing pattern analysis examines users' interactions with specific online content, such as engagement with hacking forums, illicit marketplaces, or known extremist sites (Mugari & Obioha, 2021). AI-driven models aggregate these data points and apply risk assessment scores to individuals, determining whether their digital footprint aligns with established profiles of criminal intent. While proponents argue that such tools enable proactive intervention, critics caution that predictive algorithms inherently rely on probabilistic assessments rather than direct evidence of wrongdoing, raising concerns about preemptive criminalization and the presumption of innocence (Alikhademi *et al*., 2021).

The increasing reliance on AI-driven non-invasive neuro-prediction tools presents a challenge for legal and ethical frameworks, particularly in balancing public safety with individual privacy rights. As these technologies continue to evolve, the question remains whether they serve as preventative tools or if they introduce new forms of digital surveillance that blur the line between potential risk and actual criminal intent.

Concerns about cognitive surveillance and the regulation of brain-related AI technologies are not limited to China. In the United States, Neuralink has pioneered the development of implantable BCIs, enabling direct interaction between neural activity and external systems (Neuralink, 2019). The company has conducted human trials demonstrating that individuals with paralysis can control external devices using only their thoughts. While such advancements hold significant potential for medical applications, they also raise ethical and regulatory challenges regarding mental privacy, autonomy, and the risk of cognitive data being exploited for non-medical purposes (Landis, 2024). With neural data becoming a valuable resource, questions arise as to whether states or private entities could leverage this technology to monitor or influence individuals' cognitive processes.

Recognizing these risks, Colorado has become the first jurisdiction in the United States to enact a law specifically protecting neural data. The Colorado Brain Data Bill (Colorado House Bill 24-1058, 2024), passed in 2024, amends the state's privacy laws to classify neural data, information generated by measuring brain activity, as sensitive personal data. This means that companies or entities must obtain explicit consent before collecting, storing, or using such data. The bill sets a precedent for future regulations aimed at safeguarding cognitive autonomy as BCIs and neurotechnologies continue to

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

advance (Zhang). While this legislation represents a step toward stronger protections, its effectiveness will depend on enforcement mechanisms and how emerging neuroscience applications evolve. As states and nations debate the implications of neurodata regulation, legal frameworks must ensure that advancements in AI and neuroscience do not lead to intrusive monitoring or undue influence over thought and expression.

Criminalizing certain ideas risks deterring people from expressing unconventional thoughts or exploring new concepts, which could stifle societal progress. Additionally, awareness that specific thoughts are monitored or penalized could impact mental well-being, fostering self-censorship, anxiety, and a climate of fear and distrust (Haggerty, 2006). While the notion of thought-criminalization may seem exaggerated, it serves as a reminder that maintaining clear ethical boundaries and regulatory oversight is essential in the evolution of cognitive technologies. The lighter lesson here is that protecting freedom of thought means embracing the quirky, unconventional, and even the outright oddities of human thinking and that's something worth preserving.

6. Conclusion

While the proposed EU AI Act marks a step forward in recognizing and addressing the risks posed by manipulative AI, significant gaps remain that threaten its overall efficacy. To prevent a future where cognitive manipulation becomes commonplace, a more detailed framework must be integrated into the Act. This framework should include precise definitions of cognitive manipulation and behavioural influence, delineating the boundaries of acceptable and unacceptable AI practices (Floridi, 2018). For example, the Act could incorporate a provision mandating transparency obligations for AI systems that interact with human cognition. Such a provision would require AI developers to disclose the design and intent behind algorithms, particularly when those systems use psychological insights to influence user behaviour.

Recently, the European Commission issued guidelines aimed at addressing and clarifying some of the enforcement ambiguities within the AI Act (European Commission, 2025). These guidelines specifically address harmful manipulation among other unacceptable AI practices, effectively acknowledging the ambiguities and definitional gaps highlighted in this article. In line with the article's recommendations for regulatory improvements, the Commission's guidance provides much-needed clarification by detailing the scope of prohibited practices and illustrating how manipulative techniques can be identified (European Commission, 2025). This development validates the critiques raised herein, demonstrating institutional recognition of the critical need for clearer boundaries and interpretative criteria. Nevertheless, as the guidelines themselves are non-binding and rely on adoption by Member States, the longstanding challenges of enforceability, consistency in interpretation across jurisdictions, and practical application persist. Indeed, as noted by the Commission, these measures are designed to 'ensure the consistent, effective, and uniform application' of the Act (European Commission, 2025), yet their practical impact ultimately depends on uniform interpretation and enforcement by national authorities. Hence, while the guidelines represent meaningful progress and align with this article's regulatory improvement recommendations, they simultaneously underscore that protecting cognitive autonomy effectively will remain an ongoing regulatory endeavor.

To strengthen the Act's application, it could also incorporate dynamic ethical vetting processes for AI systems legally classified as 'high-risk AI systems' under the EU AI Act. This would involve comprehensive 'Cognitive Impact Assessments,' (CIA) uniquely designed to identify manipulative tendencies, assess psychological impacts on users, and align with established human rights principles. Such assessments would not be static documents but evolving checkpoints monitored throughout an AI system's lifecycle (Rahwan, 2018). To oversee this, an independent, multidisciplinary *Council for Cognitive Integrity* could be established, blending expertise from ethics, neuroscience, law, and technology. This council would proactively audit systems, propose real-time adaptations, and maintain an ongoing dialogue with developers to ensure cognitive autonomy is respected as technology evolves.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

By weaving adaptability and continuous oversight into the Act, the EU would create a living safeguard that evolves in tandem with advancements in AI.

The proposed *CIA* must provide a structured, pre-deployment evaluation framework that not only measures cognitive influence but also assesses intent, proportionality, and user autonomy. Unlike Neuwirth's rigid tiered model, the discussed *CIA* ensures that AI systems with significant influence are not automatically presumed harmful, but are instead evaluated on a case-by-case basis with clear thresholds for determining when influence crosses into manipulation. By integrating a proactive risk assessment, rather than relying on post-hoc scrutiny, this model offers a more precise and enforceable regulatory mechanism, preventing both underregulation of harmful AI and overregulation of ethically permissible AI applications.

In comparison, it is vital to see the available impact assessments provided in the current regulatory landscape. The phrase 'high risk' refers to AI systems legally classified as such under the EU AI Act. These include applications such as biometric identification, critical infrastructure management, and systems that significantly impact human rights. While the Act mandates Fundamental Rights Impact Assessments (FRIAs) for such systems to evaluate broader human rights concerns, the proposed *CIAs* would serve a complementary role. *CIAs* would provide a focused evaluation of the psychological implications and manipulative capabilities of AI systems, particularly their potential to manipulate cognition or compromise mental autonomy. By addressing this specific dimension, *CIAs* would fill a critical gap not currently covered by FRIAs, ensuring that high-risk systems are evaluated comprehensively across both fundamental rights and cognitive integrity.

Additionally, provisions for real-time auditing and traceability should be stipulated to monitor how data is used for psychographic profiling and predictive modelling. Clear penalties for non-compliance must be outlined to serve as a deterrent against breaches that manipulate individuals' mental processes. It must be noted that the EU AI Act's Article 27 introduces fundamental rights impact assessments as a mechanism to protect against cognitive manipulation, its framework falls short of providing comprehensive safeguards for cognitive freedom. The Act's approach is limited in three crucial ways: first, it restricts mandatory assessments to public authorities and entities providing public services, leaving significant gaps in private sector deployment; second, its assessment criteria focus broadly on fundamental rights without specific attention to the unique characteristics and vulnerabilities of cognitive manipulation; and third, it lacks detailed methodological guidance for identifying and measuring cognitive impacts. In contrast, this Article's proposed CIA framework offers several key advantages: it provides a systematic method for tracing the origins of manipulation, enables more precise identification of cognitive risks through the reverse-engineering approach, and establishes clear criteria for evaluating impacts on mental autonomy. Furthermore, while Article 27(2) requires updates only when 'elements' change, the *CIA* framework introduces continuous monitoring mechanisms specifically designed to detect subtle forms of cognitive manipulation that may emerge over time. To strengthen the EU AI Act's theoretical approach, future amendments should consider incorporating these more specialized assessment tools and methodologies for evaluating impacts on cognitive freedom. In comparison with the Digital Markets Act (Article 15) and the Digital Services Act (Article 37), it can be seen that they already establish provisions for auditing profiling systems, these frameworks primarily focus on transparency and compliance rather than the deeper cognitive impact of AI-driven profiling. The DMA primarily requires gatekeepers to provide explanations for ad targeting and profiling processes, while the DSA imposes auditing obligations on VLOPs to ensure accountability in their algorithmic systems. However, these provisions do not specifically assess the psychological influence of AI on user decision-making and autonomy, nor do they impose targeted evaluations of manipulative cognitive techniques. Given this gap, additional safeguards, such as real-time auditing mechanisms tailored specifically to cognitive impact, could complement existing regulatory measures. Integrating such oversight into the AI Act would ensure that audits are not merely compliance exercises but also proactive measures that identify and mitigate AI's manipulative potential beyond what is currently required under the DMA and DSA. This structured approach,

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

combined with enhanced transparency, ethical assessments, and independent oversight, would provide a strong safeguard against the risk of AI being used to erode cognitive freedom and human agency.

Protecting cognitive freedom should become a primary pillar of the EU's digital legal framework. Incorporating principles akin to 'neuro rights,' which focus on the preservation of mental privacy and integrity, could further solidify the commitment to protecting citizens against invasive AI technologies (Ienca & Andorno, 2017, Berger & Rossi, 2023). As explored in this Article, the method of reverse engineering thought manipulation offers a compelling approach to understanding and proving causality, making it an essential tool for future policy measures. Through these targeted provisions and strategic adjustments, the EU AI Act could not only address present challenges but also future-proof against emerging risks in the rapidly evolving landscape of AI.

Funding statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The author declares no competing interests that could be perceived to exert an undue influence on the content or publication of this article.

References

Alegre S. (2021). Regulating around freedom in the "forum internum". *ERA Forum*, *21*(4), 591–604. https://link.springer.com/article/10.1007/s12027-020-00633-7

Alikhademi, K., Drobina, E., Kanich, C., Maiorca, D., Malloy, T., Meingast, M., … **Zheleva, E.** (2021). A Review of Predictive Policing from the Perspective of Fairness. *Artificial Intelligence and Law*, 29, 1.

American Psychological Association. (n.d.). APA dictionary of psychology. Retrieved July 11, 2024, from https://dictionary.apa.org/attention

Associated Press. (2024, August 22). Company that sent fake Biden robocalls in New Hampshire agrees to $1m fine: Case is seen by many as unsettling example of how AI might be used to influence groups of voters and democracy. The Guardian.

Aylsworth, T. (2022). Autonomy and manipulation: Refining the argument against persuasive advertising. *Journal of Business Ethics*, *175*(4), 689–696.

Bangor University. (2020). Psychological operations: Manipulating human behaviour. Retrieved March 14, 2024, from https://research.bangor.ac.uk/portal/files/35063914/2020_Psychological_Operations.pdf

Barnhill, A. (2022). How philosophy might contribute to the practical ethics of online manipulation. In F. Jongepier & M. Klenk (Eds.), *The philosophy of online manipulation* (pp. 66–92). Routledge.

Barrett, D. (2010). *Supernormal stimuli: How primal urges overran their evolutionary purpose*. W.W. Norton & Company.

Berger, S., & Rossi, F. (2023). AI and Neurotechnology: Learning from AI Ethics to Address an Expanded Ethics Landscape. *Communications of the ACM*, *66*(3), 26–28.

Berk, R. A. (2021). Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. *Annual Review of Criminology*, *4*(1), 209–237.

Bignull, H. (2023). *Deceptive patterns: Exposing the tricks tech companies use to control you*. Testimonium Ltd.

Bongard-Blanchy, K., Rossi, P., Bernhaupt, R., Lallemand, C., & Sauer, G. (2021). "I am definitely manipulated, even when I am aware of it. It's ridiculous!"—Dark patterns from the end-user perspective. In Proceedings of the Designing Interactive Systems Conference 2021 (pp. 763–776). doi:10.1145/3461778.3462086

Botes, M. (2023). Autonomy and the social dilemma of online manipulative behaviour. *AI and Ethics*, *3*(1), 315–321.

Brooks, S. J., Savov, V., Allzén, E., Benedict, C., Fredriksson, R., & Schiöth, H. B. (2012). Exposure to subliminal arousing stimuli induces robust activation in the amygdala, hippocampus, anterior cingulate, insular cortex and primary visual cortex: A systematic meta-analysis of fMRI studies. *NeuroImage*, *59*(3), 2962–2973.

Bruegel (2022). The dark side of artificial intelligence: Manipulation of human behaviour. Retrieved May 15, 2024, from https://www.bruegel.org/blog-post/dark-side-artificial-intelligence-manipulation-human-behaviour

Bublitz, C. (2015). Cognitive liberty or the international human right to freedom of thought. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics* (pp. 1309–1333). Springer. doi:10.1007/978-94-007-4707-4_166

Bublitz, J.-C. (2020). The nascent right to psychological integrity and mental self-determination. In A. von Arnauld, K. von der Decken & M. Susi (Eds.), *The right to mental integrity* (pp. 111–130). Cambridge University Press.

Burr, C., & Cristianini, N. (2019). Can machines read our minds? *Minds and Machines*, *29*(3), 461–481. doi:10.1007/s11023-019-09509-w

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering data mining: From concept to implementation*. Prentice Hall.

Cambridge Analytica, Retrieved November 7, 2023, from https://cambridgeanalytica.org/
Cao, Y. and **others** (2024). The Era of Artificial Intelligence Deception. *Information*, *15*(6), 299. https://www.mdpi.com/2078-2489/15/6/299
Cassin, R. (1948). 'Travaux Préparatoires of the Universal Declaration of Human Rights', 1427 UN Doc A/76/380.
Chen, J., Qian, P., Gao, X., Li, B., Zhang, Y., & Zhang, D. (2023). Inter-brain Coupling Reflects Disciplinary Differences in Real-world Classroom Learning. *NPJ Science of Learning*, *8*(1), 11.
Chén, O. Y. (2022). Uniting machine intelligence, brain and behavioural sciences to assist criminal justice. arXiv preprint arXiv:2207.01511.
Chomanski, B. (2023). Mental integrity in the attention economy: In search of the right to attention. *Neuroethics*, *16*(8).
Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. *Philosophy and Technology*, *34*, 739–766.
Cohen, J. E. (2012). *Configuring the networked self: Law, code, and the play of everyday practice*. Yale University Press.
Cohen, T. (2023). Regulating manipulative artificial intelligence. *SCRIPTed: Journal of Law, Technology and Society*, *20*(1), 203–242.
Colorado House Bill 24-1058 (2024) Protect privacy of biological data. Retrieved August, 2024, from https://leg.colorado.gov/bills/hb24-1058
Council of Europe. (2023). Study on the impact of artificial intelligence systems, their potential for discrimination, and positive action to promote equality, including gender equality. https://rm.coe.int/study-on-the-impact-of-artificial-intelligence-systems-their-potential/1680ac99e3Council of Europe
Cyberspace Administration of China, 'Internet Information Service Algorithmic Recommendation Management Provisions' (Stanford DigiChina, 1 March 2022)
Cytowic, R. E. (2024). *Your Stone Age Brain in the Screen Age: Coping with Digital Distraction and Sensory Overload*. Cambridge: MIT Press.
Dobber, T., Ó Fathaigh, R., & Zuiderveen Borgesius, F. J. (2019). The regulation of online political micro-targeting in Europe. *Internet Policy Review*, 8(4). doi:10.14763/2019.4.1440
Drew, L. (2023). The rise of brain-reading technology: What you need to know. *Nature*, *623*(7986), 241–243.
Eliot, L. (2024, October 20). Generative AI and subliminal messaging. https://www.forbes.com/sites/lanceeliot/2024/10/20/generative-ai-and-subliminal-messaging/
Engemann, A. AI and predictive policing: balancing technological innovation and civil liberties (MJLST, 20 November 2024)
European Commission. (2020). Questions and answers on the Digital Services Act. Retrieved June 22, 2024, from https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348
European Commission. (2021, August 5). Feedback from: BEUC - The European Consumer Organisation. Retrieved July 15, 2024, from https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665432_en
European Commission. (2021). Guidance on the interpretation and application of Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market. *Official Journal of the European Union*, C 526, 1–129.
European Commission (2025) Commission Publishes Guidelines on Prohibited Artificial Intelligence (AI) Practices as Defined in the AI Act. European Commission. Retrieved February 23, 2025, from https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act
European Commission. (n.d.). European approach to artificial intelligence. Retrieved February 24, 2025, from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
European Data Protection Supervisor. (2022). Annual Report 2022. https://www.edps.europa.eu/2022-edps-annual-report/en/index.htm
European Union. (2005). 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.
European Union. (2022). 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).
Facebook Newsroom. (2018, April 8). Facebook recently announced a major update to News Feed; here's what's changing. https://about.fb.com/news/2018/04/inside-feed-meaningful-interactions/.
Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.
Farahany, N. A. (2023). *The battle for your brain: Defending the right to think freely in the age of neurotechnology*. Penguin Press.
Federal Trade Commission. (2000, June). Online profiling: A report to Congress. Retrieved May 16, 2024, from https://www.ftc.gov/sites/default/files/documents/reports/online-profiling-federal-trade-commission-report-congress-part-2/onlineprofilingreportjune2000.pdf
Federal Trade Commission. (2018). *Complaint in the matter of Cambridge Analytica LLC*.
Feinberg, J. (1986). *Harm to self: The moral limits of the criminal law*. Oxford University Press.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Financial Times. (2023). We don't need new 'neurorights' — We need to know the existing law. Retrieved May 7, 2024, from https://www.ft.com/content/e8fcb5f2-94a2-4b2f-94f5-bc6e27d7c136

Fineman, M. A. (2010). The vulnerable subject and the responsive state. *Emory Law Journal*, *60*, 251–269.

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Floridi, L. (2018). *The ethics of artificial intelligence*. Oxford University Press.

Floridi, L., & Cowls, J. (2019). The ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *29*(4), 689–707.

Franklin, M., Ashton, H., Gorman, R., & Armstrong, S. (2022). Missing mechanisms of manipulation in the EU AI Act. The International FLAIRS Conference Proceedings, 35. Retrieved June 4, 2024, from https://journals.flvc.org/FLAIRS/article/view/130723

Future of Life Institute. (2022, August). Manipulation AI Act. Retrieved January 17, 2024, from https://futureoflife.org/wp-content/uploads/2022/08/FLI-Manipulation_AI_Act.pdf

Future of Privacy Forum. (2022). The future of manipulative design regulation. Retrieved July 19, 2024, from https://fpf.org/blog/the-future-of-manipulative-design-regulation/

Georgian Labour Party v. Georgia, App No 9103/04, European Court of Human Rights (ECtHR), 8 July 2008

Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14.

Hacker, P. (2021). Manipulation by Algorithms: Exploring the Triangle of Unfair Commercial Practice, Data Protection, and Privacy Law. *European Law Journal*, *29*(1–2), 142–175.

Haggerty, K. D. (2006). Tear down the walls: On demolishing the panopticon. *Theoretical Criminology*, *10*(3), 277–299.

Haggerty, K. D., & Ericson, R. V. (2000). The surveillant assemblage. *The British Journal of Sociology*, *51*(4), 605–622.

Helberger, N., Sax, M., Strycharz, J., & Micklitz, H.-W. (2022). Choice architectures in the digital economy: Towards a new understanding of digital vulnerability. *Journal of Consumer Policy*, *45*(2), 175–198.

Hendlin, Y. H. (2019). I am a fake loop: The effects of advertising-based artificial selection. *Biosemiotics*, *12*(1), 131–151. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6582976/

Herman, J. (2024). Dark Patterns: EU's Regulatory Efforts. *Security & Privacy*, *7*(6), 1.

Hern, A., & Cadwalladr, C. (2018, May 6). Cambridge Analytica: How did it turn clicks into votes? The Guardian. https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie

Ienca, M. (2021). On neurorights. *Frontiers in Human Neuroscience*, *15*, 701258.

Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, *13*(1), 5. doi:10.1186/s40504-017-0050-1

Ienca, M., & Ignatiadis, K. (2020). Artificial Intelligence and Neurotechnology: Learning from AI Ethics to Address an Expanded Ethics Landscape. *Communications of the ACM*, *63*(11), 34–36.

Independent. (2019, January 9). China schools scan brains with concentration headbands. Independent. Retrieved February 8, 2024, from https://www.independent.co.uk/tech/china-schools-scan-brains-concentration-headbands-children-brainco-focus-a8728951.html

Investopedia. (n.d.). Data mining. Retrieved November, 2023, from https://www.investopedia.com/terms/d/datamining.asp#toc-data-mining-techniques

Jamieson, K. H., & Cappella, J. N. (2008). *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Kogan, A. (2018). Written evidence submitted by Aleksandr Kogan. Retrieved June 5, 2023, from https://www.parliament.uk/globalassets/documents/commons-committees/culture-media-and-sport/Written-evidence-Aleksandr-Kogan.pdf

Landis, B. (2024). Brain-Computer Interfaces and Bioethical Implications on Society: Friend or Foe? *St Thomas Law Review*, *36*(2), 178.

Latham & Watkins, LLP, 'China's New AI Regulations' (2023). Retrieved February 22, 2024, from https://www.lw.com/admin/upload/SiteAttachments/Chinas-New-AI-Regulations.pdf

Lewis, P., & Hilder, P. (2018, March 23). Leaked: Cambridge Analytica's blueprint for Trump victory. The Guardian.

Ligthart, S., Bublitz, C., Douglas, T., Forsberg, L., & Meynen, G. (2022). Rethinking the right to freedom of thought: A multidisciplinary analysis. *Human Rights Law Review*, *22*(4), ngac028.

Ligthart, S., Douglas, T., Bublitz, C., Kooijmans, T., & Meynen, G. (2020). Forensic brain-reading and mental privacy in European human rights law: Foundations and challenges. *Neuroethics*, *13*(1), 1–13.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48), 12714–12719. doi:10.1073/pnas.1710966114

Merikle, P. M., & Daneman, M. (2000). Conscious vs. unconscious perception. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 1295–1303). MIT Press.

Metzinger, T. (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, *4*, 931.

Moore, M. S. (2023). Causation in the law. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/causation-law/

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Morris, H. (1965). Punishment for thoughts. *Monist*, *49*(3), 342–376.
Mugari, I., & Obioha, E. E. (2021). Predictive policing and crime control in the United States of America and Europe: Trends in a decade of research and the future of predictive policing. *Social Sciences*, *10*(6), 234.
Neuralink (2019) An integrated brain-machine interface platform with thousands of channels. Retrieved July 6, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6914248/
NeuroRights Foundation. (n.d.). Retrieved May 6, 2024, from https://neurorightsfoundation.org/
Neuwirth, R. J. (2023a). *The EU's Artificial Intelligence Act: Regulating Subliminal AI Systems*. Routledge.
Neuwirth, R. J. (2023b). Prohibited Artificial Intelligence Practices in the Proposed EU Artificial Intelligence Act (AIA). *Computer Law & Security Review*, *48*(1), 105798.
Noggle, R. (2020). The ethics of manipulation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2020. Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/sum2020/entries/ethics-manipulation/
Office of the High Commissioner for Human Rights. (2021, October 20). Freedom of thought increasingly violated worldwide, UN expert warns. Retrieved March 21, 2025, from https://www.ohchr.org/en/press-releases/2021/10/freedom-thought-increasingly-violated-worldwide-un-expert-warns
Orwell, G. 1949 Nineteen Eighty-Four (Secker & Warburg).
Packard, V. (1957). *The hidden persuaders*. New York, NY: David McKay Company.
Pariser, E. (2011). *The filter bubble: What the internet is hiding from you*. Penguin Press.
Pazzanese, C. (2017, February 27). *In Europe, nationalism rising*. Harvard Gazette.
Petropoulos, F., Castle, J. L., Hendry, D. F., & Martinez, A. B. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. doi:10.1016/j.ijforecast.2021.11.001
Pötzl, O. (1917). The relationship between experimentally induced dream images and indirect vision. *Monograph No. 7, Psychological Issues*, *2*, 41–120.
Pub. Util. Comm'n of D.C. v. Pollak, 343 U.S. 451, 467 (1952).
Puri, A. (2021). The right to attentional privacy. *Rutgers Law Review*, *48*, 206–225.
Quach, X., & Lee, S. H. (2021). Profiling gifters via a psychographic segmentation analysis: Insights for retailers. *International Journal of Retail & Distribution Management*, *49*(3), 383–398. doi:10.1108/IJRDM-06-2020-0220
Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14.
RAND Corporation, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations (RAND Research Report RR233, 2013) https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf
Regulation (EU). 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L265/1
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) [2022] OJ L265/1
Regulation (EU). 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, OJ L 2024/1689
Russian Conservative Party of Entrepreneurs and Others v. Russia. App Nos 55066/00 and 55638/00. European Court of Human Rights (ECtHR, 11 January 2007.
Russo, A. (2009). An American right to an "unannoyed journey"? Transit radio as a contested site of public space and private attention, 1949-1952. *Historical Journal of Film, Radio and Television*, *29*(1), 1–21.
Sax, M., & Helberger, N. (2024). Digital vulnerability and manipulation in the emerging digital framework. In *Digital Fairness for Consumers* (pp. 11). BEUC, The European Consumer Organisation.
Simon, H. (2020). Hidden Persuaders: An Exploration of Subliminal Advertising Techniques and Their Impact on Consumer Behavior. *Journal of Marketing Research*, *57*(3), 456–472.
Smith, A. There's an open secret about Cambridge Analytica in the political world: It doesn't have the 'secret sauce' it claims. 2018. Business Insider
Smuha, N. A. (2023). Beyond the Individual: Charting the Societal Risks of AI. *EJRR*, *18*(4), 112.
Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: A response to the European Commission's proposal for an Artificial Intelligence act. *SSRN Electronic Journal*. doi:10.2139/ssrn.3899991.
Solomon, M. R. (2004). *Encyclopaedia of applied psychology*. Elsevier Academic Press.
Stel, M., Mastop, J., & Strick, M. (2011). The impact of mimicking on attitudes toward products presented in TV commercials. *Social Influence*, *6*(3), 142–153.
Stenlund, M., & Slotte, P. (2018). Forum internum revisited: Considering the absolute core of freedom of belief and opinion in terms of negative liberty, authenticity, and capability. *Human Rights Review*, *19*(4), 425–446.
Sunstein, C. R. (2015). *Choosing not to choose: Understanding the value of choice*. Oxford University Press.

https://doi.org/10.1017/cfl.2025.4 Published online by Cambridge University Press

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, *4*(1), 1–45.

Thaler, R., Sunstein, C., & Balz, J. P. (2010). Choice architecture. Retrieved October 11, 2023, from https://www.sas.upenn.edu/∼baron/475/choice.architecture.pdf

Tran, J. L. (2016). The right to attention. *Indiana Law Journal*, *91*, 1023–1082.

Tushnet, M. (2010). *Why the Constitution matters.* Yale University Press.

TwoBirds. (2023). DSA - Targeted advertising aimed at minors: A future ban? Retrieved January 17, 2024, from https://www.twobirds.com/en/insights/2023/global/dsa-publicite-ciblee-destinee-aux-mineurs-une-interdiction-a-venir

UNESCO, Recommendation on the Ethics of Artificial Intelligence (2022)

United Nations. (1948). Universal Declaration of Human Rights (adopted 10 December 1948, UN General Assembly Resolution 217 A(III)). Retrieved July 17, 2024, from https://www.un.org/en/about-us/universal-declaration-of-human-rights

United Nations General Assembly. (2021, October 5). Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed on freedom of thought (UN Doc A/76/380). Retrieved from https://undocs.org/A/76/380

University of Cambridge, Statement from the University of Cambridge about Dr Aleksandr Kogan. https://www.cam.ac.uk/notices/news/statement-from-the-university-of-cambridge-about-dr-aleksandr-kogan

van der Ploeg, M. M., Brosschot, J. F., Versluis, A., & Verkuil, B. (2017). Peripheral physiological responses to subliminally presented negative affective stimuli: A systematic review. *Biological Psychology*, *129*, 131–153. doi:10.1016/j.biopsycho.2017.08.051

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

Wells, G., Horwitz, J., & Seetharaman, D. (2021, September 14). Facebook knows Instagram is toxic for teen girls, company documents show. *Wall Street Journal*, Retrieved from https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-companydocuments-show-11631620739

Wilczyński, P., Mieleszczenko-Kowszewicz, W., & Biecek, P. (2024). 'Resistance Against Manipulative AI: Key Factors and Possible Actions', arXiv preprint arXiv:2404.14230

Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.

Wojnarska-Krajewska, E. (2021). Wybrane aspekty prawne scrapowania ze szczególnym uwzględnieniem informacji i danych publicznych. In M. Bernaczyk (Ed.), *Prace Naukowe Wydziału Prawa, Administracji i Ekonomii Uniwersytetu Wrocławskiego*. E-Wydawnictwo.

Wu, T. (2016). *The attention merchants: The epic scramble to get inside our heads*. Faculty Books. https://scholarship.law.columbia.edu/books/64

Yeung, K. (2017). Hypernudge: Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136. doi:10.1080/1369118X.2016.1186713

Zarsky, T. Z. (2002-2003). Mine your own business: Making the case for the implications of the data mining of personal information in the forum of public opinion. *Yale Journal of Law & Technology*, *5*, 1–38.

Zarsky, T. Z. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118–132.

Zhang, C. & **others**, 'Brain-to-Text Decoding: A Non-invasive Approach via Typing' (Meta AI, 2024) https://ai.meta.com/research/publications/brain-to-text-decoding-a-non-invasive-approach-via-typing/

Zhong, H., O'Neill, E., & Hoffmann, J. A. (2024). Regulating AI: Applying Insights from Behavioural Economics and Psychology to the Application of Article 5 of the EU AI Act. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(18), 20001.

Zuboff, S., & Schwandt, K. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & de Vreese, C.H. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, *14*(1), 82–96. doi:10.18352/ulr.420

Aimen Taimur is a human rights lawyer and is currently a PhD researcher at the Tilburg Institute for Law, Technology, and Society (TILT), Tilburg University. Her research explores the intersection of human rights law, AI regulation, and digital governance, with a specific focus on manipulative technologies and cognitive freedom.

Cite this article: Aimen T. (2025). Cognitive freedom and legal accountability: Rethinking the EU AI act's theoretical approach to manipulative AI as unacceptable risk. *Cambridge Forum on AI: Law and Governance* **1**, e20, 1–28. https://doi.org/10.1017/cfl.2025.4