# What Is a Subliminal Technique? An Ethical Perspective on AI-Driven Influence

Juan Pablo Bermúdez
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
j.bermudez@imperial.ac.uk
ORCID 0000-0001-5239-2980

Rune Nyrup
*Independent Researcher*
nyrup.rune@googlemail.com

Sebastian Deterding
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
s.deterding@imperial.ac.uk

Laura Moradbakhti
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
l.moradbakhti@imperial.ac.uk

Céline Mougenot
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
c.mougenot@imperial.ac.uk

Fangzhou You
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
f.you22@imperial.ac.uk

Rafael A. Calvo
*Dyson School of Design Engineering*
*Imperial College London*
London, United Kingdom
r.calvo@imperial.ac.uk

***Abstract*— Concerns about threats to human autonomy feature prominently in the field of AI ethics. One aspect of this concern relates to the use of AI systems for problematically manipulative influence. In response to this, the European Union's draft AI Act (AIA) includes a prohibition on AI systems deploying subliminal techniques that alter people's behavior in ways that are reasonably likely to cause harm (Article 5(1)(a)). Critics have argued that the term 'subliminal techniques' is too narrow to capture the target cases of AI-based manipulation. We propose a definition of 'subliminal techniques' that (a) is grounded on a plausible interpretation of the legal text; (b) addresses all or most of the underlying ethical concerns motivating the prohibition; (c) is defensible from a scientific and philosophical perspective; and (d) does not over-reach in ways that impose excessive administrative and regulatory burdens. The definition is meant to provide guidance for design teams seeking to pursue responsible and ethically aligned AI innovation.**

***Keywords*— *Automated Influence, Online Manipulation, EU AI Act, Nudge, Autonomy, Mental Integrity, Ethical Risks of Artificial Intelligence, Dark Patterns***

## I. Introduction

Concerns about threats to human autonomy feature prominently in many statements on AI ethics ([1]–[4]). One aspect of this involves the threat of machine learning and other powerful AI techniques being used in problematically manipulative ways, that is, in ways that illegitimately influence the beliefs, values, decisions or behavior of individuals.

The desire to prevent problematic manipulation is also reflected in legislative initiatives, most prominently the European Union's forthcoming AI Act (AIA), which is currently being negotiated by the European Parliament and the Council of the EU. The AIA draft text contains two provisions that are explicitly motivated by a desire to prevent manipulative uses of AI, namely articles 5(1)(a) and 5(1)(b). In this paper we focus on 5(1)(a), which in the most recent draft proposal prohibits

> "the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness with the objective to or the effect of materially distorting a person's behavior in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm" [5].

Exactly what kinds of technologies might be covered by article 5(1)(a) remains uncertain. Part of this uncertainty relates to the term 'subliminal techniques' ([6]–[11]). The term does not have an established meaning in prior EU law, nor is it defined in the AIA. In cognitive science, the word 'subliminal' usually refers to stimuli that the person is unaware of having perceived, but which may nonetheless influence their behavior. Interpreted this way, the prohibition would have a rather narrow scope.

This is a sensitive issue. Many ethicists and legal scholars are concerned that the law might be interpreted too narrowly to offer meaningful protections ([6]–[9]). Meanwhile industry stakeholders argue that the costs of complying with an overly broad interpretation could negatively impact innovation. Organizations found to have breached this prohibition can be fined up to €30 million or up to 6 % of their global yearly turnover in the case of companies. Apart from the issue of fines, the costs of complying with article 5(1)(a) could be significant if 'subliminal techniques' were interpreted too broadly.

So far, most commentary on this issue has focused on how the legal text itself might be changed to address the issue of scope, for example by removing the term 'subliminal techniques' or supplementing it with a more precise definition ([6], [10], [11]). In this paper we pursue a different strategy. Starting from the current wording of the AIA, we propose a definition of 'subliminal techniques' which, we argue, (a) is a plausible interpretation of the legal text; (b) addresses all or

This work was made possible by funding from Huawei Technologies.

most of the underlying ethical concerns motivating the prohibition; (c) is defensible from a scientific and philosophical perspective; and (d) does not over-reach in ways that impose excessive administrative and regulatory burdens on companies and other organizations. We propose this definition as a guide for design teams seeking to pursue responsible and ethically aligned AI innovation. Our aim thus is not to pinpoint a definition that will suffice for legal compliance, but rather to support an ethically ambitious approach, which aligns with the motivations behind article 5(1)(a). Our definition provides a tool to assess the ethical risk of different influence techniques, and thus helps incentivize ethical design practices within the AI space.

Section II presents a standard, technical definition of subliminal techniques. Section III argues that such a definition cannot do the job required of it in the AIA. Section IV then specifies the desiderata that a useful definition of the term should fulfil and provides a psychological and philosophical background for building the definition. Section V presents the broader definition we defend and shows that it fulfils all the desiderata set out for it.

## II. The Narrow Definition of Subliminal Techniques

Subliminal techniques have long been studied in psychology and discussed in the marketing literature as a way to potentially influence consumer behavior [12]. To capture this concept, Still and Still introduce the notions of objective and subjective threshold ([13], see Fig1). Each perceptual stimulus has specific levels of intensity (e.g., loudness, brightness, duration). A stimulus is above the *objective threshold* if it is intense enough to have an effect on the agent's behavior (e.g., make the agent more likely to choose certain words than others). In contrast, a stimulus is above the *subjective threshold* whenever the agent is aware of having perceived it. This allows them to define subliminal stimuli as those stimuli that are above the objective and below the subjective threshold. If the agent reports having perceived the stimulus when asked, this suggests that the stimulus was intense enough to surpass the subjective threshold. On the other hand, if they report not having perceived it, but some measurable difference can be found in their behavior (compared to those not presented with the stimulus), this is evidence that the stimulus surpassed the objective threshold without surpassing the subjective threshold. It would, in other words, be evidence that the stimulus was "subliminal".

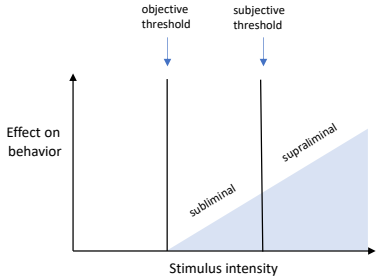


Figure 1. It is assumed that the stimulus will have greater effect on the agent as its intensity increases. A stimulus with intensity that surpasses the objective threshold but is still below the subjective threshold can affect the agent's behavior. If a stimulus intensity is above the subjective threshold, the agent becomes aware of it. (Adapted from [13, p. 458]).

Based on this discussion, one way to define subliminal techniques is as follows:

- *Narrow Definition:* Subliminal techniques aim at influencing a person's behavior by presenting a stimulus in such a way that the person remains unaware of the stimulus presented.

Subliminal techniques, in this sense, have been proposed as tools to facilitate technology adoption by transmitting information to the user without taxing scarce (meta)cognitive resources (e.g., working memory, attention) [13]. In a recent example, researchers used subliminal stimuli to influence driver behavior [14]. Sensors in and around a car recorded information about energy consumption, which was then conveyed to drivers through subliminal seat and seatbelt vibrations. In the routes where subliminal feedback was available, data suggested a significant improvement in driving economy, including for drivers who reported not being aware of the vibrations.

This case highlights the promise of these subliminal techniques, namely that they can be used to convey relevant information without taxing scarce cognitive resources. However, they also come with perils: if unaware of the vibrations and their influence, a driver may not be able to resist the influence on their behavior. This also seems to have motivated the European Commission's initial proposal for AIA: the briefings that accompanied this proposal mention a hypothetical example, where an inaudible sound is played in a truck driver's cabin, which pushes the driver to continue driving longer than is healthy and safe [7, p. 98].

While purely hypothetical, the example clearly illustrates *one* problematic way that AI could be used. Thus, it makes sense that AIA should seek to regulate techniques that rely on the subliminal presentation of information. However, as we will now argue, there are several reasons that the notion of 'subliminal techniques' ought not be interpreted as *only* covering techniques that rely on subliminal stimuli.

## III. Limitations of the Narrow Definition

### A. Legal Limitations

The most recent draft of the AIA, namely the 'General Approach' adopted by the Council of Europe, makes it clear that the motivation behind the "subliminal techniques" prohibition is intended to tackle a broader range of situations than those covered by the narrow definition. Specifically, Recital 16, which articulates the motivations behind Article 5(1)(a), singles out AI systems that

> "deploy subliminal components such as audio, image, video stimuli that persons cannot perceive as those stimuli are beyond human perception or other subliminal techniques that subvert or impair a person's autonomy, decision-making or free choices in ways that people are not consciously aware of, or even if aware not able to control or resist" [5].

The passage explicitly contrasts techniques using stimuli that persons cannot perceive with *other subliminal techniques*. The latter are characterized as impacting our autonomy in ways we are "not consciously aware of" or which we are "not able to control or resist". Such characterizations apply to a

number of techniques that are not covered by the Narrow Definition discussed above.

In contrast to the operative provisions, such as article 5(1)(a), recitals in EU law are not legally binding. However, they state the reasoning behind the operative provisions and judges can use them to disambiguate provisions. Since the notion of 'subliminal techniques' is not a well-established concept in the law, the above passage may well play an important role in determining what kinds of techniques will be covered. That said, there is legal uncertainty as to how the term will be interpreted in practice.

### *B. Scientific and Methodological Limitations*

Some might be tempted to defend the narrow definition of subliminal techniques on the grounds that this is the *scientific* meaning of the term and therefore (presumably) a more objective definition or one that can be more reliably applied.

However, despite the long tradition of research on the potential to influence behavior through subliminal perception, much disagreement remains about the phenomenon's scope and practical relevance. Researchers agree that the effects of subliminal stimuli on behavior, if they exist, are much weaker than popular discussions assume [11]. They are extremely short-lived, tending to leave no traces of lasting effects beyond 100ms ([15], [16]), and seem capable of triggering actions only if the individual already intends to perform them [17].

Methodologically, determining whether a given presentation of information is subliminal or not involves a number of problems [12]. Different people have different perceptual sensitivities, so one and the same stimulus can be subliminal to some and supraliminal to others. Moreover, the same person can be more or less perceptively acute depending on, e.g., fatigue levels or how divided their attention is. Thus, it is methodologically challenging to determine that a certain stimulus is *subliminal in general*, as opposed to *subliminal for a certain person in a specific context*. Finally, researchers often rely on self-report to assess whether a token stimulus is subliminal, but participant reports can differ with respect to the level of confidence they require to report having perceived a stimulus [13].

### *C. Ethical Limitations*

The narrow definition's most important drawback is that it leaves out many of the potentially problematic cases of AI-based influence. These cases have been discussed in the existing literature on dark patterns (technology design elements that benefit service providers at the cost of users [18], [19]) and the ethics of digital nudging [20]. Here are a few examples of existing AI-driven influence techniques that have raised concerns about autonomy-undermining influence in recent years.

- *Psychological targeting.* Tailoring communication to the psychological characteristics of the recipient may enhance its persuasiveness. The availability of large amounts of social media data and other digital footprints has enabled the development of machine learning systems capable of predicting the psychological profile of large numbers of people accurately enough to target persuasive messages to their individual profile. Matz and colleagues [21] have tested the effectiveness of this technique as a persuasion strategy, through experiments in which different versions of an advert were shown to Facebook users, based on their predicted personality. For instance, a make-up advert targeting individuals with high predicted extroversion would have the tag-line "Dance like no one's watching (but they totally are)", while adverts for the same retailer targeting individuals with low predicted extroversion read "Beauty doesn't have to shout". In this example, adverts congruent with their predicted personality resulted in 50% more purchases than non-targeted ads. While there are methodological worries about this particular study ([22], [23]), studies of this type fuel ethical concerns, namely the potential to influence people's choice beyond their awareness.
- *Digital nudging.* The concept of nudging refers to persuasion techniques that influence behavior without removing the options available to an individual or significantly altering their incentives [24]. Drawing on insights from behavioral science, nudging relies on changes to "choice architecture", such as the order in which options are presented or how they are framed. Nudging techniques are increasingly used in the design of digital platforms. For instance, Uber has been known to deploy a variety of data-driven nudges to discourage drivers from finishing their shifts, even when it is in the drivers' best interest [25]. One of these nudges involved informing the driver how close they were to reaching some arbitrary money target, e.g., "You are $6 away from making $40 in net earnings", thereby framing the decision to log off as a loss, which is known to increase motivation more than gain framings [26]. Moreover, whereas traditional nudges are static and uniform (e.g., all users have to actively opt out of tracking cookies), AI-powered techniques raise additional challenges by allowing digital nudges to be deployed in more dynamic and individually-tailored ways ([20], [27], [28]).
- *Search engine manipulation effects*. Search engines are essential to navigating the vast amounts of information on the internet. However, research suggests that the way they present results can subtly influence our attitudes and decisions. A 2014 study showed that the order in which the same 30 search results were ranked could change political voting preferences by more than 20%, while leaving a large proportion of the experimental participants unaware of the intervention [29]. While this study did not involve AI, AI systems could be used to boost this effect, e.g., to identify and target a particular population (e.g., undecided voters) or to identify the way to bias results most effectively and most opaquely.
- *Recommender systems and goal misalignment*. Recommender systems present tailored options predicted to appeal to the individual user. Using machine learning to analyze large datasets, recommender systems are designed to present those options that optimally elicit certain kinds of behavior, such as buying more products or staying engaged with a social media platform. While these

behaviors are often portrayed as proxies for the satisfaction of user preferences, there is concern in among media and the public that they can become misaligned in ways that instead exploit and reinforce negative emotions, such as outrage or negative social comparison [30]. Existing studies suggest that automated recommendations can change user preferences instead of learning how to satisfy them ([11], [31], [32]).

As this non-exhaustive list illustrates, ethically problematic forms of AI-driven influence may utilize a host of different techniques. Importantly for our purposes, none of these examples involve the subliminal presentation of stimuli, so they would not be covered by the narrow definition. Nonetheless, something covert is going on: many users may not be aware of how AI-driven techniques are used to influence their evaluations and decisions. Moreover, even in cases where users show some awareness of the intervention, it can still have an effect on their behavior. Becoming aware that one is being nudged need not make nudges ineffective ([33], [34]); likewise, being aware of automated influence need not neutralize its effect on behavior. This raises the question of whether users, even if aware, could in fact control or resist the effects of the intervention.

While a closer analysis of these examples goes beyond the scope of this paper, we take them to be part of what the AIA's protection against manipulative uses of AI is meant to cover, given the text in Recital 16. At the very least, Recital 16 opens the door to consider these as belonging to the set of practices targeted by Article 5(1)(a)'s prohibition.

## IV. Toward a Better Definition of Subliminal Techniques

In sum, the Narrow Definition of 'subliminal techniques' does not fit the legal context that inspired the AIA's prohibition, lies on shaky scientific and methodological grounds, and is too narrow to tackle many (or most) of the ethical misuses of AI influence. This section proposes a way forward.

### A. Desiderata

Based on the limitations outlined in the preceding section, we stipulate four desiderata for an ethically useful definition of subliminal techniques. The definition should

(a) be consistent with the legal text;

(b) cover all or most cases of ethical concern that motivate the legal text;

(c) avoid placing excessive administrative and regulatory burdens; and

(d) provide guidance for ethical innovation and management of AI systems.

The first desideratum should be self-explanatory, but it is worth briefly commenting on (b)–(d). The definition should be sufficiently but not excessively broad. Focusing exclusively on (b) may lead to a maximally broad definition. While this would ensure that all problematic AI-based influence techniques count as 'subliminal', all non-problematic ones would as well. From an industry perspective, this would create significant disincentives to the development and use of AI systems that interface with user decision-making. It would mean that companies and other organizations would be forced to do extensive checks on the impacts of any such AI system, to ensure that they cannot accidentally cause harm. Combined with the possibility of a fine of up to 6% of annual worldwide revenue for companies that are determined to have violated the prohibition, this would create strong incentives against innovation in this field. Thus, (b) should be balanced with (c): if it is to be useful in practice, the definition should help identify the influence practices and techniques that involve the greater ethical risks, so that companies can set up due diligence mechanisms to mitigate these risks, without unduly disincentivizing the development and use of new technologies. In connection with this point, (d) states that the definition should be useful as guidance for AI producers seeking to develop technology in ethically responsible ways. It should provide them with a tool to distinguish between different levels of required oversight depending on the AI system's ethical relevance.

### B. How to Influence Others without Threatening Autonomy: A Model of Ethical Influence

Subliminal techniques are ethically relevant because, as we will discuss in this section, not having awareness of the ways in which we are being influenced may undermine our ability to self-govern, and thus our autonomy. This worry is present in the AIA. As mentioned above, Recital 16 understands 'subliminal techniques' broadly to include those techniques "that subvert or impair a person's autonomy, decision-making or free choices in ways that people are not consciously aware of, or even if aware not able to control or resist". Three key ideas are used to describe the scope of this broader category: (1) influence techniques that threaten personal autonomy; (2) conscious awareness; and (3) the ability to control or resist. In this section we build a model of ethical influence that incorporates these ideas in a way that seeks to be theoretically well-grounded, i.e., largely philosophically uncontroversial and coherent with our current scientific understanding of human action.

We understand influence as an intervention aimed at changing a person's beliefs, values, goals, or behaviors [35]. Some influence is clearly non-problematic, like the truthful and non-biased offering of reasons for a certain view; other forms are clearly problematic, like coercing someone to do something under threat of violence. However, there are also many forms of influence, including many putative cases of manipulation, that are less clear. Where is the line between ethical and unethical forms of influence? The discussion is complex ([9], [36], [37]), but there is at least one uncontroversial point: forms of influence that do not threaten the autonomy of their target audience are ethically less problematic than those that do. Discussing autonomy will thus help clarify the ethics of influence.

Personal autonomy can be articulated in various ways, but generally involves the ability "to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces" [38], or similarly, the ability "to act, reflect, and choose on the basis of factors that are somehow [the agent's] own (authentic in some sense)" [38]. In other words, to be autonomous involves having values and beliefs that are *authentic*, and being able to

*guide* one's actions in accordance with those values and beliefs [2].

While a person's beliefs and values can change in ways that increase their authenticity [39], we focus here on how external influences can distort belief- and value-formation processes in ways that reduce the person's authenticity. We also discuss how external influences can undermine an agent's ability to guide action in accordance with their own values and beliefs. External influences count as distortive if they reduce the agent's ability to form or maintain authentic values and beliefs, or if they reduce the agent's ability for action guidance. Indoctrination and 'brainwashing' are forms of authenticity-undermining influence. Violent coercion is a form of influence that undermines the agent's guidance abilities. However, other, more subtle forms of influence can also threaten an agent's autonomy.

We constantly form beliefs and acquire interests unintentionally and unconsciously [40]. This entails that our ability to consciously and reflectively nurture authentic values and beliefs is limited. Such limitations highlight an ethical requirement on influence: to ensure an influence attempt does not further reduce the agent's already limited capacity to cultivate authenticity, the influencer should provide the agent with opportunities to consciously understand the way they are being influenced and how the influence methods used impact their beliefs and values. External influences that operate outside the agent's awareness are harder (perhaps impossible) to revise and endorse, and thus can more easily compromise the authenticity of the agent's beliefs and values [41].

Action guidance relies on a set of abilities and processes that allow us to control our actions so that they reflect our beliefs and values. We sometimes guide our actions by explicitly deliberating and making decisions or plans, but we often also guide actions through habits that we have cultivated, or through emotional and intuitive cues and reactions. Part of being able to guide action is thus being able to check whether these processes that shape our actions are in line with our values and beliefs, and correcting course when they do not.

Cognitive science provides ample evidence that our abilities for action guidance and control often break down, displaying failures of self-control, habitual action slips, and planning errors [42]. These limitations impose an ethical requirement on responsible influence: we should avoid influencing others in ways that predictably undermine their already limited capacities to guide action in accordance with their authentic values and beliefs. We should at least provide them with opportunities to understand the ways in which the influence methods used may impact their guidance abilities.

To summarize, when an action is autonomous it is consistent with the agent's authentic values, and the agent is able to guide its performance, i.e., to notice if the action deviates from their authentic values and intervene to correct it when required. Awareness is crucial for autonomy because it enables us to notice such deviations and attempt to correct them. Granted, awareness is not sufficient for successful correction (you may be unable to correct an action despite knowing it is misaligned with your values), but it is necessary for attempting correction (if you do not notice the misalignment, you would not even be able to try to correct it).

As mentioned above, evidence from cognitive science reveals that the capacities underlying personal autonomy (for authenticity and guidance) are limited and fragile. These vulnerabilities enable exploitation by others. Thus, ethical attempts at influence need to take care that they do not further undermine such capacities. In order to do this, it is crucial that the influenced agent is given the opportunity to become consciously aware of and understand how their capacities for authenticity and guidance may be impacted by the influence attempt.

More specifically, influence techniques ought to be deployed in ways that allow the influenced agent to become aware of (1) the fact that they are being influenced, (2) how the influence techniques work, and (3) how the techniques affect their beliefs, values, and decisions. Doing this is required to make sure that the influence attempt does not threaten the autonomy of agents.

### C. *The Risks of AI-Based Influence for Autonomy*

We summarize these points as a model of autonomous action (Figure 2). Within this model, we can distinguish three ways that external influences—such as AI-based influence techniques discussed above—can threaten autonomy. They can undermine action guidance directly (through e.g. computer-brain interfaces and other emerging technologies) by compelling us to take misaligned actions [41] (Figure 2, [c]). But more commonly, they can distort our belief- and value-formation processes (Figure 2, [a]) or distort our decision-making process, leading us to make choices and to act in ways that are not aligned with our beliefs and values (Figure 2, [b]).

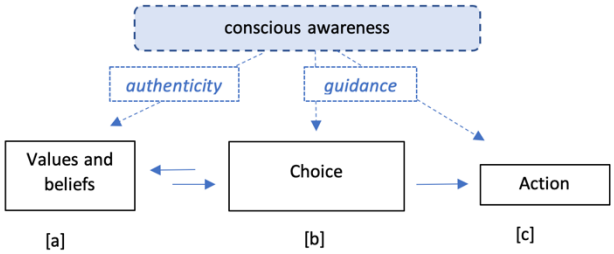


Figure 2: A model of autonomous action. Autonomy can be undermined by external influences that change [a] the agent's values and beliefs, [b] decision-making processes, or [c] actions. Consciousness awareness helps agents to control and resist those influences.

This can explain why it makes sense for Recital 16 to target AI-based influence techniques that circumvent conscious awareness. By keeping some aspects of the influence process outside of consciousness, these techniques make it harder for the agent to exercise higher-order guidance when the action deviates from their values. The examples of problematic AI influence mentioned above illustrate this:

- Search engine manipulation effects and the cases of goal misalignment via recommender systems illustrate how AI-based influence can undermine authenticity by distorting belief- and value-formation processes. If the agent remains unaware of how these influence methods operate, or of the effect they have on them, their ability to attempt resisting said effect is diminished.
- Through psychological targeting and other forms of digital nudging, AI-based influence can undermine guidance over decision-making, since the agent

remains unaware of how these influence methods work and of the consequences they have on their deliberation, thereby being less able to resist their influence.

Thus, influence techniques can be problematic for autonomy not only when the agent is unaware of the influencing stimulus itself, but also when they lack awareness of the stimulus' effects on their behavior, or of how the influence method being used on them operates.

In short, the model of autonomy presented in this section motivates a broader definition of the term 'subliminal techniques' that goes beyond mere unconscious stimulus perception.

## V. Subliminal Techniques: A Broader Definition

On a broader interpretation, an influence technique can count as *subliminal* even if the stimulus itself is consciously perceived, as long as its methods and its effects on behavior are not conscious. We thus propose the following:

- *Broad Definition:* Subliminal techniques aim at influencing a person's behavior in ways in which the person is likely to remain unaware of (1) the influence attempt, (2) how the influence works, or (3) the influence attempt's effects on decision-making or value- and belief-formation processes.

This definition includes all the techniques included in the Narrow Definition, i.e., techniques that aim at producing influence via a subliminal stimulus. But it also includes the ethically concerning techniques that we have discussed above that fall outside of the Narrow Definition's scope: To take few examples: Many *digital nudging* techniques alter the presentation of options in ways that predictably affect evaluation and decision-making without the user's awareness of the fact that they are being influenced, of the way they are being influenced, and of the effects this has on their behavior. And when exposed to ads using *psychological targeting*, people can be aware of the fact that the ad is an attempt at influence and even of the effect that the ad has on their decision-making; but they may remain unaware of the way in which they are being influenced (i.e. through personality-based ad targeting). The Broad Definition thus makes sense of the passage in the AIA's Recital according to which subliminal methods can undermine autonomy when the agent, even if aware, is not able to control or resist. This can be understood if we distinguish forms of lack of awareness. Even if the agent is aware (1) of the influence attempt, their control over the influence can still be reduced if they remain unaware of (2) how the influence works, or (3) what effects it has on them. The Broad Definition thus explains how the agent's autonomy can be reduced due to lack of awareness *even* in cases where the agent is aware of being influenced.

The Broad Definition satisfies our desiderata (a) and (b). We discuss (c) and (d) in the rest of this section.

### A. Is the Broad Definition overly broad?

Our definition may, at first glance, seem very broad. Of course, this is to some extent intended. As argued in Section IV, since the motivating ethical concern is to protect people's autonomy, any influence technique that predictably circumvents the awareness of the person being influenced should be included. Nonetheless, it might be thought that the definition is *overly* broad, thereby running afoul of our desideratum (c). A few things can be said in response to this.

First, the question of whether an AI system uses subliminal techniques applies only to some AI systems. Systems used for energy consumption optimization, industrial automation, image and video compression, and many others do not interface with persons' decision-making processes in the ways that make the question of subliminal techniques relevant.

Second, the definition refers to *techniques*, that is, methodological procedures that predictably produce certain results. System features that in a few cases produce the relevant kinds of unawareness, but not in a systematic way, and that thus do not predictably produce the same effect elsewhere, would not count as subliminal *techniques*. Thus, many systems that do interact with human decision-making would likely not count. For instance, while systems involved in speech and face recognition, language translation, or even financial credit approval raise ethical questions (e.g., about bias and fairness) and can be said to influence people's decisions, users are not predictably likely to be unaware of how this influence works. That said, some systems of this kind would count as using subliminal influence techniques. For instance, systems that use a certain digital nudge, or that develop their influence messages through multiple rounds of A/B testing and personalization techniques to identify the message that maximizes influence. In such cases, the relevant kinds of techniques would be present and the issue of subliminal influence would become relevant (since users would tend to be unaware of whether there are influence techniques at work behind the message received, or of how they work).

And third, it should be noticed that the AIA does not prohibit the use of subliminal techniques *per se,* but only those that *materially distort a person's behavior* in ways *that cause or are reasonably likely to cause physical or psychological harm.* AI system producers are free to use any and as much subliminal techniques as they want, *as long as* they make sure to not distort people's behavior in potentially harmful ways.

The term 'material distortion' is significant. It signals that more is at stake here than a mere *change* in behavior. For instance, Article 1 of the European Union's Unfair Commercial Practices Directive (UCPD) defines 'materially distorting consumer behavior' as "using a commercial practice to appreciably impair the consumer's ability to make an informed decision, thereby causing the consumer to take a transactional decision that he would not have taken otherwise" [43]. Here, to distort someone's behavior means more than simply changing it from what it would otherwise have been. Rather, it involves changes to behavior that result from reducing their capacity for informed decision-making.

Now, the UCPD governs the realm of commercial transactions, so it focuses on ensuring informed decision-making. Since the AIA's realm of application is broader, and article 5(1)(a) is motivated by concerns around autonomy-undermining influence techniques, we propose expanding the scope here from informed decisions to autonomous decisions more generally. Going back to our model of autonomous action, we thus propose that a subliminal technique *materially distorts* a person's behavior when it significantly impairs the abilities required for authenticity and guidance.

In short, while our definition of subliminal techniques is broad, it does not entail that organizations would have to evaluate every single AI system for potential harms in order to comply with article 5(1)(a). It of course still includes more than the Narrow Definition. However, as argued in Section III, some broadening of this definition is likely necessary, for both legal-compliance and ethical reasons. We suggest that the Broad Definition provides an ethically robust approach, which captures most if not all problematic types of influencing, without over-including irrelevant cases.

### *B. Can the Broad Definition provide practical guidance?*

To address desideratum (d), we sketch a decision procedure that designers and organizations deploying AI systems can use to help identify whether an AI system includes subliminal techniques, and what to do if it does. We summarize this procure as a decision tree in Figure 3. While the decision tree depicts this as a straightforward process, we want to emphasize that each step will still require careful reflection on the part of technology designers.

The first step is to determine whether the very question of subliminal techniques makes sense for ethically assessing the AI system in question. As discussed above, the question is relevant only if the system influences human decision-making in predictable ways. This is because only for this kind of influence does the lack of awareness introduce risks of autonomy reduction.

The next step in this procedure is to check whether subliminal techniques are present according to the Broad Definition. To do this, separate checks would be required for each of the Definition's three clauses: lack of awareness of (1) the influence attempt, (2) its mode of operation, or (3) its effects on decision and judgment. If the system tends to produce one or more of these lacks of awareness, then the next step is to determine whether these techniques are used in a distorting way, that is, in a way that significantly reduces the person's capacity for authenticity or guidance. If they do, then an ethical risk assessment is required to assess whether the system's use of subliminal techniques implies an increase in the risk of harm for any persons whose behavior is in this way distorted, to anyone that may be affected by those behaviors.

Finally, even if the subliminal techniques used are accurately deemed not to significantly impair authenticity and guidance, this assessment may change in the course of users' interaction with the system. We therefore recommend that the system continues to be monitored through its life cycle, so that any new impairments of autonomy can be identified.

What these risk assessment and monitoring processes entail is beyond the scope of this paper and will be discussed in future work. However, we take this flowchart (Figure 3) to illustrate that the Broad Definition can in principle be used to specify in which cases systems require further monitoring or risk-assessment procedures. It in other words, the Definition provides AI-producing companies and organizations with a set of criteria to check an influence technique's level of ethical risk. Companies that want to develop AI systems in an ethically responsible way can incorporate into their design process awareness checks for each of the steps in the flowchart. The documentation of such a process can be incorporated into a risk-management system, the records of which can be kept as evidence of due diligence.

## VI. Conclusion

While the AIA seeks to tackle AI-based influence systems that can have harmful effects on personal autonomy, the choice of 'subliminal techniques' as a target of prohibition may lead to either interpretations of the prohibition that are too narrow to tackle the most serious and pervasive forms of problematic manipulation, or to a reading that is so broad as to impose prohibitive burdens on any organization wishing to develop or use AI systems.

To avoid both of these extremes, we propose a definition which identifies subliminal techniques as the influence techniques where the influenced agent remains predictably unaware of the influencing stimulus, the way the techniques operate, or their effects on their values, beliefs, and decisions. This definition is grounded on widely shared views of personal autonomy, offers a coherent interpretation of the legal text, and provides ethical guidance for technology development practice.

There is still much uncertainty about how the notion of 'subliminal techniques' will be interpreted in legal practice. Adopting practices based on the Broad Definition can also allow companies to make a case for responsibly assessing the ethical risks associated with automated influence, which may be significant in reducing the risks involved in legal uncertainty.

We thus recommend that the Broad Definition be adopted by all developers seeking to design AI systems ethically and responsibly.

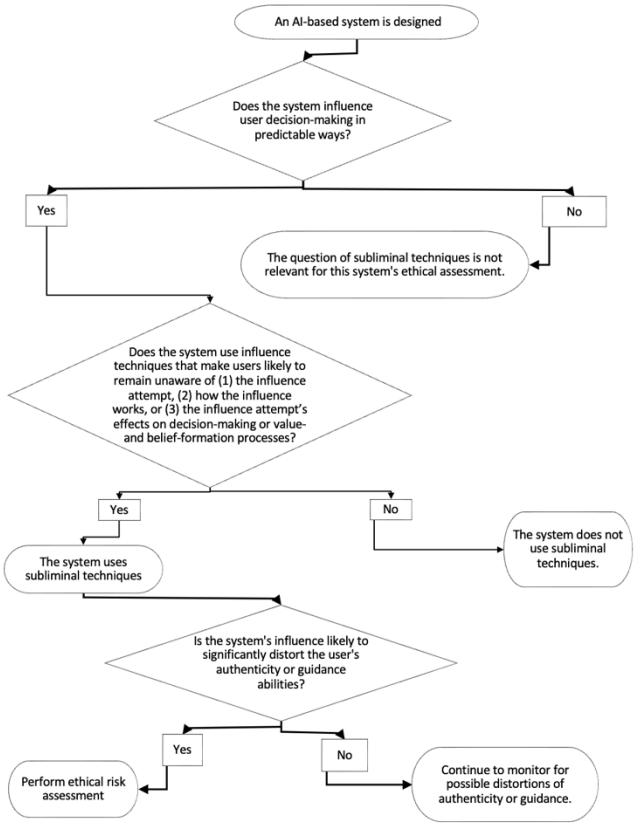


Figure 3: Does my system use subliminal techniques? And if so, what should I do? A decision tree according to the Broad Definition.

## VII. Future Directions

In line with the AIA's Article 5(1)(a), and with the discussion above, we highlight two directions for future research crucial to ensuring the responsible use of automated influence.

(1) *Subliminality measures* are required to make it possible to identify when influence attempts count as subliminal in relation to each of the broad definition's clauses. Such measures will be required for putting the Article into practice and eventually settling debates in relation to the subliminal nature of particular influence techniques used by specific technologies.

(2) *Studies of algorithmic mental harm*. The AIA prohibits the use of subliminal techniques that distort people's behavior if that is reasonably likely to cause mental or physical harm. The notion of physical harm is well studied , but the concept of mental harm remains under-developed both in legal theory, psychology, and even philosophy. Work on the nature and kinds of mental harm will be crucial to specifying the kinds of effects that subliminal techniques should not have. To make progress in this direction, we can take inspiration from existing work on taxonomies of algorithmic or sociotechnical harms ([44]–[47]), and in particular work on harms caused by dark patterns ([48], [49]).

## Author Contribution Statement

JPB and RN led the conceptualization, investigation, drafting and editing of the paper. RC and SD contributed to the conceptualization, editing, project administration and supervision. FY, LM and CM supported conceptualization and review of the final manuscript.

## References


[1] A. Jobin, M. Ienca, and E. Vayena, 'The global landscape of AI ethics guidelines', *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[2] C. Prunkl, 'Human autonomy in the age of artificial intelligence', *Nat. Mach. Intell.*, vol. 4, no. 2, pp. 99–101, Feb. 2022, doi: 10.1038/s42256-022-00449-9.

[3] R. A. Calvo, D. Peters, K. Vold, and R. M. Ryan, 'Supporting human autonomy in AI systems: a framework for ethical enquiry', in *Ethics of Digital Well-Being*, C. Burr and L. Floridi, Eds., in Philosophical Studies Series. Cham: Springer, 2020, pp. 31–54. doi: 10.1007/978-3-030-50585-1_2.

[4] R. A. Calvo, D. Peters, D. Johnson, and Y. Rogers, 'Autonomy in technology design', in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, Toronto Ontario Canada: ACM, Apr. 2014, pp. 37–40. doi: 10.1145/2559206.2560468.

[5] Council of the European Union, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach', Council of the European Union, 2021/0106(COD), Nov. 2022. [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

[6] R. Uuk, 'Manipulation and the AI Act', The Future of Life Institute, 2022. Accessed: Sep. 05, 2022. [Online]. Available: https://futureoflife.org/wp-content/uploads/2022/01/FLI-Manipulation_AI_Act.pdf

[7] M. Veale and F. Z. Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act', *Comput. Law Rev. Int.*, vol. 22, no. 4, pp. 97–112, 2021.

[8] N. A. Smuha *et al.*, 'How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act', *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.3899991.

[9] J. C. Bublitz and T. Douglas, 'Manipulative Influence via AI Systems and the EU Proposal for Regulation of Artificial Intelligence', Feedback on the European Commission's Draft Artificial Intelligence Act, Aug. 2021. Accessed: Oct. 21, 2022. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665640_en

[10] C. Boine, 'AI-enabled manipulation and EU law', *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.4042321.

[11] M. Franklin, H. Ashton, R. Gorman, and S. Armstrong, 'Missing mechanisms of manipulation in the EU AI Act', *Int. FLAIRS Conf. Proc.*, vol. 35, May 2022, doi: 10.32473/flairs.v35i.130723.

[12] M. Elgendi, P. Kumar, S. Barbic, N. Howard, D. Abbott, and A. Cichocki, 'Subliminal priming—state of the art and future perspectives', *Behav. Sci.*, vol. 8, no. 6, p. 54, May 2018, doi: 10.3390/bs8060054.

[13] M. L. Still and J. D. Still, 'Subliminal techniques: considerations and recommendations for analyzing feasibility', *Int. J. Human–Computer Interact.*, vol. 34, no. 5, pp. 457–466, May 2018, doi: 10.1080/10447318.2017.1358973.

[14] A. Riener, 'Subliminal persuasion and its potential for driver behavior adaptation', *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 71–80, Mar. 2012, doi: 10.1109/TITS.2011.2178838.

[15] A. G. Greenwald, S. C. Draine, and R. L. Abrams, 'Three cognitive markers of unconscious semantic activation', *Science*, vol. 273, no. 5282, pp. 1699–1702, Sep. 1996, doi: 10.1126/science.273.5282.1699.

[16] D. R. Shanks *et al.*, 'Priming intelligent behavior: an elusive phenomenon', *PLoS ONE*, vol. 8, no. 4, p. e56515, Apr. 2013, doi: 10.1371/journal.pone.0056515.

[17] S. J. Brooks, V. Savov, E. Allzén, C. Benedict, R. Fredriksson, and H. B. Schiöth, '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*, vol. 59, no. 3, pp. 2962–2973, Feb. 2012, doi: 10.1016/j.neuroimage.2011.09.077.

[18] A. Narayanan, A. Mathur, M. Chetty, and M. Kshirsagar, 'Dark patterns: past, present, and future: the evolution of tricky user interfaces', *Queue*, vol. 18, no. 2, pp. 67–92, 2020.

[19] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, 'The dark (patterns) side of UX design', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–14. doi: 10.1145/3173574.3174108.

[20] K. Yeung, '"Hypernudge": Big Data as a mode of regulation by design', *Inf. Commun. Soc.*, vol. 20, no. 1, pp. 118–136, Jan. 2017, doi: 10.1080/1369118X.2016.1186713.

[21] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, 'Psychological targeting as an effective approach to digital mass persuasion', *Proc. Natl. Acad. Sci.*, vol. 114, no. 48, pp. 12714–12719, Nov. 2017, doi: 10.1073/pnas.1710966114.

[22] B. Sharp, N. Danenberg, and S. Bellman, 'Psychological targeting', *Proc. Natl. Acad. Sci.*, vol. 115, no. 34, Aug. 2018, doi: 10.1073/pnas.1810436115.

[23] D. Eckles, B. R. Gordon, and G. A. Johnson, 'Field studies of psychologically targeted ads face threats to internal validity', *Proc. Natl. Acad. Sci.*, vol. 115, no. 23, Jun. 2018, doi: 10.1073/pnas.1805363115.

[24] R. H. Thaler and C. R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

[25] N. Scheiber, 'How Uber uses psychological tricks to push its drivers' buttons', *The New York Times*, Apr. 02, 2017. Accessed: Sep. 29, 2022. [Online]. Available: https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html

[26] A. Tversky and D. Kahneman, 'The framing of decisions and the psychology of choice', *SCIENCE*, vol. 211, p. 30, 1981.

[27] A. Stemler, J. E. Perry, and T. Haugh, 'The code of the platform', *Ga. Law Rev.*, vol. 54, pp. 605–662, 2020.

[28] S. Mills, 'Finding the "nudge" in hypernudge', *Technol. Soc.*, vol. 71, p. 102117, Nov. 2022, doi: 10.1016/j.techsoc.2022.102117.

[29] R. Epstein and R. E. Robertson, 'The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections', *Proc. Natl. Acad. Sci. - PNAS*, vol. 112, no. 33, pp. E4512–E4521, 2015, doi: 10.1073/pnas.1419828112.

[30] G. Wells, J. Horwitz, and D. Seetharaman, 'Facebook knows Instagram is toxic for teen girls, company documents show', *Wall Street Journal*, Sep. 14, 2021. Accessed: Jan. 12, 2022. [Online].

Available: https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

[31] M. Alfano, A. E. Fard, J. A. Carter, P. Clutton, and C. Klein, 'Technologically scaffolded atypical cognition: the case of YouTube's recommender system', *Synthese*, Jun. 2020, doi: 10.1007/s11229-020-02724-x.

[32] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang, 'Do recommender systems manipulate consumer preferences? A study of anchoring effects', *Inf. Syst. Res.*, vol. 24, no. 4, pp. 956–975, Dec. 2013, doi: 10.1287/isre.2013.0497.

[33] H. Bruns, E. Kantorowicz-Reznichenko, K. Klement, M. Luistro Jonsson, and B. Rahali, 'Can nudges be transparent and yet effective?', *J. Econ. Psychol.*, vol. 65, pp. 41–59, Apr. 2018, doi: 10.1016/j.joep.2018.02.002.

[34] G. Loewenstein, C. Bryce, D. Hagmann, and S. Rajpal, 'Warning: You are about to be nudged', *Behav. Sci. Policy*, vol. 1, no. 1, pp. 35–42, 2015.

[35] D. Susser and V. Grimaldi, 'Measuring automated influence: between empirical evidence and ethical values', in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, Jul. 2021, pp. 242–253. doi: 10.1145/3461702.3462532.

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

[37] C. Benn and S. Lazar, 'What's wrong with automated influence', *Can. J. Philos.*, p. 50, 2021, doi: 10.1017/can.2021.23.

[38] J. Christman, 'Autonomy in moral and political philosophy', in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Fall 2020.Metaphysics Research Lab, Stanford University, 2020. Accessed: Oct. 05, 2022. [Online]. Available: https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

[39] S. Varga and C. Guignon, 'Authenticity', *Stanford Encyclopedia of Philosophy*. 2020. [Online]. Available: https://plato.stanford.edu/archives/spr2020/entries/authenticity/

[40] J. St. B. T. Evans, *Thinking twice: Two minds in one brain*. Oxford University Press, 2010.

[41] J. C. Bublitz and R. Merkel, 'Crimes against minds: on mental manipulations, harms and a human right to mental self-determination', *Crim. Law Philos.*, vol. 8, no. 1, pp. 51–77, Jan. 2014, doi: 10.1007/s11572-012-9172-y.

[42] S. Amaya, 'Agency and mistakes', in *The Routledge Handbook of Philosophy of Agency*, L. Ferrero, Ed., Routledge, 2021, pp. 149–158.

[43] European Commission, *Unfair Commercial Practices Directive [UCPD]*. 2005.

[44] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami, 'Toward user-driven algorithm auditing: investigating users' strategies for uncovering harmful algorithmic behavior', in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–19. doi: 10.1145/3491102.3517441.

[45] M. Banko, B. MacKeen, and L. Ray, 'A unified taxonomy of harmful content', in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online: Association for Computational Linguistics, 2020, pp. 125–137. doi: 10.18653/v1/2020.alw-1.16.

[46] J. Redden and J. Brand, 'Data harm record', 2017. [Online]. Available: https://orca.cardiff.ac.uk/id/eprint/107924/1/data-harm-record-djl2.pdf

[47] R. Shelby *et al.*, 'Sociotechnical harms: scoping a taxonomy for harm reduction'. arXiv, Oct. 11, 2022. Accessed: Feb. 07, 2023. [Online]. Available: http://arxiv.org/abs/2210.05791

[48] J. Gunawan, C. Santos, and I. Kamara, 'Redress for dark patterns privacy harms? A case study on consent interactions', in *Proceedings of the 2022 Symposium on Computer Science and Law*, Washington DC USA: ACM, Nov. 2022, pp. 181–194. doi: 10.1145/3511265.3550448.

[49] C. M. Gray, C. Santos, N. Bielova, M. Toth, and D. Clifford, 'Dark patterns and the legal requirements of consent banners: an interaction criticism perspective', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–18. doi: 10.1145/3411764.3445779.