Friedrich-Alexander-Universität
Erlangen-Nürnberg

Ethical Implications of AI-Driven Nudging

by

Dennis Mustafić

Matriculation Number	23332080
E-Mail	dennis.mustafic@fau.de
Degree	M. Sc. Data Science
Course	Ethics and Philosophy of AI
Submission Date	23.01.2025
Term	Winter Term 24/25

Abstract

This paper explores the implications of AI-driven nudging on user autonomy, focusing on the ethical challenges posed by personalized, data-driven interventions. Drawing on the theory of Libertarian Paternalism by Thaler and Sunstein, the discussion highlights how traditional nudges aimed at improving decision-making fundamentally differ from AI-powered nudges in their personalization and complexity. The key issues of opt-out availability and transparency are examined, demonstrating that the lack of straightforward opt-out mechanisms and the opaque nature of AI algorithms significantly constrain user autonomy. The role of Explainable AI (XAI) is analyzed as a potential solution to the transparency problem, emphasizing the need for interpretable models to ensure users can make informed choices. In conclusion, this essay supports the implementation of explainable algorithms - even with certain concerns - to promote an ethical AI landscape where user autonomy and transparent decision-making are prioritized.

1 Introduction

Nudging – a concept originally rooted in behavioral economics introduced by Thaler and Sunstein in "Nudge. Improving Decisions about Wealth, Health and Happiness" – has been a part of our daily lives for many years, subtly influencing our decisions and guiding us towards particular choices without coercion. Traditionally, nudges have been designed to influence behavior by altering the "choice architecture" (Thaler and Sunstein, 2008, p. 6), such as the placement of healthy food in grocery stores in more easily accessible areas. However, with the rise of artificial intelligence (AI), the sophistication of nudging has increased significantly. AI-powered nudges now influence our day-to-day interactions through platforms like Google Maps, which recommends eco-friendly routes, Netflix suggesting movies based on recent user activity, and social media algorithms deciding what content we see on platforms like TikTok or YouTube. These data-driven nudges are becoming an inevitable part of our decision-making process, influencing us in almost every aspect of our lives.

At the heart of the phenomenon of nudging is the theory of *Libertarian Paternalism*, as outlined by Thaler and Sunstein, which advocates for guiding individuals toward better decisions while preserving their freedom of choice. Nudging, in this framework, seeks to improve well-being without restricting options or economic incentives, relying on the subtle manipulation of choice environments to encourage preferred behaviors (Thaler and Sunstein, 2008, p. 6). Nevertheless, the use of AI has introduced new challenges and ethical questions regarding the extent to which these nudges impact autonomy and freedom. Unlike classical nudges, which are often static and apply universally, AI-driven nudges are tailored to individuals based on personal data—such as screen time or browsing habits—allowing for highly personalized nudges. This customization, coupled with the use of reinforcement learning, enhances the effectiveness of AI nudges (Nallur et al., 2024, p.2-3), but also raises concerns about the extent to which individuals retain control over their decisions.

Autonomy, in the context of this paper, is understood as the ability for individuals to make decisions in a self-determined manner, free from manipulation or excessive external influence. Drawing from Immanuel Kant's definitions of autonomy as "the property of the will to be a law unto itself" (Kant, 1870, p. 75, as quoted in Waldegge, 2017, p. 1), it involves not just the freedom to act according to one's desires, but also the ability to reason and choose without coercion or subtle pressures, including those exerted by AI-based systems. For this discussion, the term autonomy covers preservation of individual freedom to make informed decisions without being subtly nudged in ways that compromise one's self-determination, even if these nudges are designed to promote individual well-being or societal goals.

Based on this definition, two key questions arise: "At what point do we lose our autonomy when exposed to AI-nudging?" and, following from this, "Is it possible to provide AI-nudging without subtly manipulating individuals?". Answering the first question requires a closer examination of the principles of Libertarian Paternalism as outlined by Thaler and Sunstein. From their work, two main dimensions in the context of preserving autonomy emerge: *availability of opt-out options* (subsection 1.1) and *transparency of the nudging process* (subsection 1.2).

1.1 Availability of Opt-Out Options

Thaler and Sunstein claim that the absence of transparency and the lack of an low cost opt-out option in nudging mechanisms lead to a significant loss of autonomy (Thaler and Sunstein, 2008, p. 237). An opt-out option is critical, as it allows users to disengage from these influences when they choose to, thus protecting their ability to make independent decisions. Without such an option, autonomy is severely restricted.

1.2 Transparency of the Nudging Process

Thaler and Sunstein emphasize the importance of transparency in nudging. Their interpretation of the *publicity principle* calls for nudging mechanisms to be openly transparent, enabling individuals to understand the motives and methods behind these interventions. This transparency ensures that users can make informed decisions, preserving their autonomy in the face of these subtle influences (Thaler and Sunstein, 2008, p. 244-246). Additionally, a recently published review by Michaelsen demonstrates that several experiments indicate transparency of the nudge does not diminish the effectiveness (Michaelsen, 2023, p. 815). This means that nudges can be both effective and transparent, thereby enhancing user autonomy. This underscores the significance of transparency in preserving autonomy.

If either of these factors—opt-out options or transparency—is neglected, the autonomy of individuals is endangered. Considering these two guidelines provided by Thaler and Sunstein, this paper argues that unlike traditional nudges, users often have no real means of opting out of these influences, making manipulation irresistible. Additionally, it is argued that even if transparency is guaranteed, the user cannot understand how the decision by the AI-system was made, resulting in a limitation of autonomy.

2 The Problem of Opting-Out

Thaler and Sunstein define “opting out” as the ability to avoid a specific action, rule, or nudge (Thaler and Sunstein, 2008, p. 5). A key property of this concept is that opting out should be easy and inexpensive (Thaler and Sunstein, 2008, p. 236) . For instance, when signing up for a newsletter, the checkbox for subscribing is often pre-checked, and all the user needs to do is uncheck it to avoid receiving the newsletter. This simplicity ensures that individuals can make decisions freely, even when nudges influence their behavior.

In AI-driven systems, however, opting out is often far from straightforward. There are typically two ways to opt out of AI nudges: disabling the nudge within the system (subsection 2.1), or avoiding the technology or platform entirely (subsection 2.2). Both options present significant barriers to autonomy.

2.1 Disabling Nudges

Many AI-driven platforms do not offer users a straightforward option to disable nudges. For example, while it is possible to hide YouTube’s recommender system using third-party browser extensions on a PC, this option is not available on smartphones. More importantly, the platforms themselves often do not provide any built-in mechanisms for users to turn off these nudges. This lack of functionality means that even if users wish to opt out, they may be unable to do so without some technical intervention or expertise. This does not meet the principle that the opt-out should be simple and inexpensive (Thaler and Sunstein, 2008, p. 247-248), which leads to a considerable restriction of user autonomy.

2.2 Missing Alternatives and Market Dominance

When disabling nudges within a system is not feasible, the only alternative is to switch platforms. However, dominant digital platforms, like Google, Facebook, and YouTube, often make this choice impractical. Opting out may mean disengaging from widely-used digital spaces, effectively isolating users from social aspects of modern life. Users are faced with a choice: either they accept being part of the system, engage with social media, and remain exposed to AI-driven nudges, or they opt out and risk becoming disconnected from the social and digital landscape.

The dominance of a few digital platforms worsens the issue of user autonomy because it leaves users with limited options to escape AI-driven nudges. As these monopolies grow, the lack of alternatives not only restricts choice but also increases the influence of these platforms, making it harder for individuals to maintain their autonomy.

3 Why Transparency is Not Enough

While transparency is important for preserving autonomy, it proves insufficient in the context of AI-driven nudging due to the inherent complexity of the algorithms. Unlike traditional nudges, which are often straightforward and easy to interpret, AI-driven nudges are powered by complex algorithms that even experts may find challenging to fully understand. This can lead to information overload, which will be discussed in subsection 3.1.

Furthermore, due to the opaque ”black box”-like nature of machine learning algorithms, it is often impossible to understand the decisions made by the model. This leads to a further loss of autonomy, since the publicity principle gets violated. Subsection 3.2 looks further into this issue.

3.1 Information Overload

Transparency can lead to information overload, as many users lack the technical background to process complex machine learning details. When faced with too much information, users may feel overwhelmed, experience ”filter failure”, and suffer from information anxiety, reducing their ability to make informed choices (Bawden and Robinson, 2020, p. 1-2, 20, 24) and hence diminishing their autonomy.

Choice overload is another issue, where too much information about nudging mechanisms can cause cognitive fatigue and decision paralysis. Instead of empowering users, this surplus of information can make it harder for them to navigate choices effectively, leading to anxiety and poor decision-making (Bawden and Robinson, 2020, p. 21).

The combination of information and choice overload can subtly manipulate users, compromising their autonomy by making it difficult to understand the consequences of their actions or how to opt out.

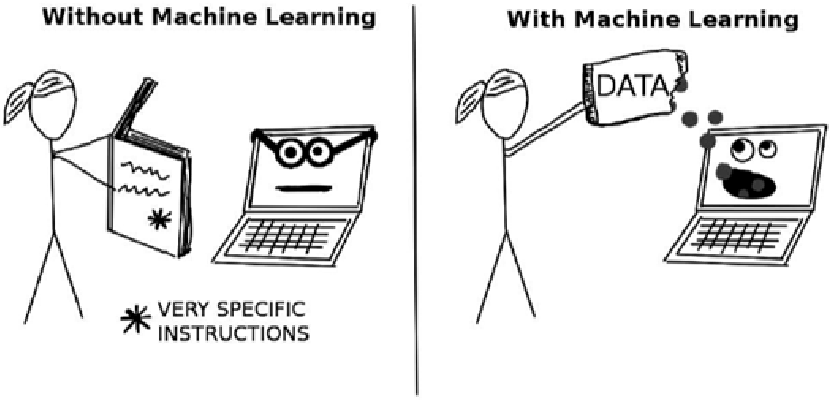

Figure 1: Comparison of classical algorithms (left side) to machine learning algorithms (right side), highlighting their methodological differences (Molnar, 2020, p. 12).

3.2 The Lack of Interpretability

The most severe problem of AI-driven nudges is the lack of interpretability in their decision-making processes. Unlike traditional nudges, where transparency and opt-out options are relatively straightforward, AI-driven nudges face significant challenges in this area. The algorithms that power these nudges are often highly complex, and even experts mostly struggle to understand how they make decisions. This lack of clarity is commonly referred to as the "black box problem."

To clarify this further, it is important to differentiate between traditional algorithms and machine learning algorithms. An algorithm, in general, is a set of rules that a machine follows to achieve a particular goal. For example, an algorithm can be thought of as a cooking recipe: the ingredients are the inputs, the finished dish is the output, and the steps taken to prepare the dish are the rules or instructions to follow. Machine learning, on the other hand, represents a shift in how machines are programmed. Instead of explicitly providing all instructions, as in traditional programming, machine learning allows computers to learn from data in order to make and improve predictions. This shift from "normal programming" to "indirect programming" means that the machine learns how to make decisions from the data provided, usually without human intervention or clear instruction at each step (Molnar, 2020, p. 12).

An illustrative example of how the understanding of decision-making is lost in neural networks can help clarify this difference. Consider a neural network trained to classify images of animals, initially using labeled images of cats and dogs. At first, the relationship between the input (image) and output (label) seems straightforward: if the image resembles a cat, the network outputs "cat"; if it resembles a dog, the network outputs "dog." However, as training progresses, the network begins to learn abstract features—such as edge patterns or texture gradients—that are imperceptible to humans. This means that the network might eventually misclassify a dog image as a cat if, for example, the background of the dog image contains elements more commonly seen in the training data for cats, like grass or sky [1]. This decision-making process is obscure to humans, as the neural network's decisions rely on high-dimensional, abstract features that cannot be easily interpreted.

[1] Adapted example of Maier et al., 2023, p. 17

In contrast to classical algorithms, which are inherently interpretable due to their explicit, rule-based decision-making processes, machine learning algorithms—particularly neural networks—rely on learned representations that are often too complex for human understanding. As a result, these algorithms are often treated as ”black boxes,” meaning their decision-making is difficult, if not impossible, to fully comprehend. This lack of interpretability prevents the achievement of the publicity principle, which advocates for transparency and understanding in AI systems, as the complexity of machine learning systems makes it challenging to fully realize this ideal.

4 Discussion

The field of XAI (or Interpretable Machine Learning) tries to address the issues outlined in subsection 3.2 by developing methods that make the behavior and predictions of machine learning systems understandable to humans (Molnar, 2020, p. 13). It aims to provide insights into why a model makes certain decisions, not just what those decisions are (Molnar, 2020, p. 15). As XAI research progresses, achieving true transparency in AI systems could become a realistic goal. Such advancements would significantly reduce the manipulative potential of AI-driven nudges, thereby enhancing user autonomy. If users can understand how AI nudges function, they are better equipped to make informed decisions, aligning with the principles of Libertarian Paternalism while preserving their self-determination.

While the promise of XAI is providing explanations regarding model decisions, it is still uncertain whether these methods will ultimately solve the transparency problem. As models become more complex, the challenge is not just providing insights into the model's decisions but ensuring that such insights are accessible and useful to end users. If interpretability methods overwhelm users with excessive technical details (as discussed in subsection 3.1), they risk worsening the issues they aim to resolve. Thus, while XAI is a step in the right direction in providing transparency regarding decisions, further advancements in this field need to ensure that the explanations remain understandable, making it feasible to provide transparency in AI-driven nudges.

5 Conclusion

AI-driven systems, despite progress in XAI, still face challenges related to opacity and information overload, which hinder true user autonomy. While XAI offers potential solutions, it must balance interpretability with accessibility to avoid overwhelming users. Prioritizing more interpretable algorithms, such as decision trees, could improve transparency and align with ethical standards that respect user autonomy. However, even with opt-out options, practical barriers and the lack of technical knowledge may prevent meaningful user control, highlighting the need for more ethical and user-friendly AI systems.

References

Bawden, D. and L. Robinson (2020). Information overload: An overview.

Kant, I. (1870). *Grundlegung zur metaphysik der sitten*, Volume 28. L. Heimann.

Maier, A., V. Christlein, K. Breininger, Z. Yang, L. Rist, M. Nau, S. Jaganathan, C. Liu, N. Maul, L. Folle, K. Packhäuser, and M. Zinnen (2023, April). Deep learning lecture slides: Visualization and attention mechanisms.

Michaelsen, P. (2023). Transparency and nudging: an overview and methodological critique of empirical investigations. *Behavioural Public Policy*, 1–11.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Nallur, V., K. Renaud, and A. Gudkov (2024). Nudging using autonomous agents: risks and ethical considerations. *arXiv preprint arXiv:2407.16362*.

Thaler, R. H. and C. R. Sunstein (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New York: Penguin Books.

Waldegge, F. H. v. (2017). Was ist autonomie? *Hegel-Jahrbuch 2017*(1), 197–201.