



INSTITUTE FOR RESEARCH IN ECONOMIC AND FISCAL ISSUES

## IREF Working Paper Series

# From Libertarian Paternalism to AI-Powered Nudging: New Challenges for Freedom

Sergio Beraldo

IREF WORKING PAPER NO. 202505

MAY 2025

IN ENGLISH: [EN.IREFEUROPE.ORG](http://en.irefeurope.org)  
IN FRENCH: [FR.IREFEUROPE.ORG](http://fr.irefeurope.org)



INSTITUTE FOR RESEARCH IN ECONOMIC AND FISCAL ISSUES

# From Libertarian Paternalism to AI-Powered Nudging: New Challenges for Freedom

Sergio Beraldo

Department of Economics and Statistics, University of Napoli 'Federico II' and CSEF.

Address: Via Cinthia, Monte Sant'Angelo - 80126 - Naples, Italy.

e-mail address: [s.beraldo@unina.it](mailto:s.beraldo@unina.it)

May 12, 2025

Talk held at the IREF colloquium, organized at the Department of International Comparative Studies, Warsaw School of Economics, University of Warsaw, April 16, 2025. I would like to thank Agnieszka Slomka-Golebiowska for the kind invitation and hospitality, Pierre Garelo, Marek Lusztyń, and Enrico Colombatto for their thoughtful comments and discussion.

# 1 Introduction

The temptation to rely on findings from behavioral economics to condition human behavior has given rise to libertarian paternalism, an approach to public policy design that leverages psychological insights to steer decision-making. In this talk, I discuss whether this approach genuinely respects individual freedom. My discussion analyzes the philosophical underpinnings of libertarian paternalism, questions its compatibility with genuine autonomy, and proposes a reasons-based framework for evaluating its recommendations.

The fundamental tension explored here lies in whether freedom of choice is merely a matter of having the formal right to select from available options, or whether it requires deeper conditions, namely, the opportunity to make intentional, reason-based decisions.

My discussion will also address the consequences of the growing capacity of both public and private organizations to influence choices through increasingly refined variations in the choice context, made possible by the widespread use of artificial intelligence and big data analytics. These developments raise concerns about the erosion of individual autonomy in exercising choice, whether in market transactions or in other areas of life.

## 2 The Limits of Homo Economicus

Economic theory has long relied on the model of homo oeconomicus, which assumes individuals to have stable and internally consistent preferences, and to be capable of systematically evaluating all available options and selecting those that best serve their self-interest.

The homo oeconomicus model of human decision-making serves a dual purpose. First, it provides mathematical tractability, allowing economists to construct elegant models that yield testable predictions. Second, it offers a normative benchmark against which actual behavior can be assessed.

The homo oeconomicus model has been increasingly challenged since the late 1970s with the emergence of behavioural economics. By applying insights from cognitive psychology to economic behaviour, researchers like Daniel Kahneman and Amos Tversky demonstrated that human decision-making frequently and systematically deviates from the theoretical predictions considered standard in Economics (e.g. Kahneman, 2011). Their findings marked the rise of

a new conception of human decision-making, grounded in ideas from cognitive psychology. Among these are the notions that the human brain often operates as an energy-saving system - defaulting to mental shortcuts and heuristics rather than engaging in full-scale optimization - and that decisions can be highly susceptible to framing effects: contexts matter, and the way options are presented is often *as important as*, or even *more important than*, their objective attributes. These insights directly challenge the assumptions that preferences are stable and internally consistent, raising fundamental questions about the very foundations of choice.

### 3 Status quo biases and framing effects

Well-known departures from the predictions of the standard model are due to status quo bias and framing effects. Status quo bias refers to the human tendency to disproportionately prefer existing arrangements over change, even when switching would entail minimal costs or offer significant benefits. This *inertia* finds a possible explanation in loss aversion, the psychological tendency to weigh potential losses more heavily than equivalent gains (e.g. Kahneman et al., 1991). Samuelson and Zeckhauser (1988)'s classic investment choice experiments demonstrated, for example, that simply labeling an option as the default or status quo dramatically increased its selection frequency. Status quo bias manifests across numerous domains with significant policy implications. In retirement savings, employees overwhelmingly accept default contribution rates even when financially sub-optimal. In insurance markets, consumers rarely switch providers despite potential savings. In technology, users typically accept default privacy settings without review. This behavioural inertia creates opportunity for choice architects to influence outcomes by strategic default setting, a power that raises concerns about manipulation.

The framing effect, meanwhile, illustrates how presenting identical information in different ways can yield opposite choices (e.g. Tversky and Kahneman, 1981). McNeil et al. (1982)'s research on cancer treatment choices showed, for example, that describing outcomes in terms of survival rates versus mortality rates significantly altered patients' treatment preferences, despite conveying mathematically equivalent information. When presented with survival statistics, patients favoured surgery; when shown mortality figures, they preferred radiation therapy.

Beyond these examples, behavioural researchers have documented numerous other systematic departures from the predictions of standard rational choice theory. As reported by

Schmauder et al. (2023), at the time their paper was published the Wikipedia page on cognitive biases listed 251 of them. The study of these biases reveal that human choice is, among other things, influenced by seemingly irrelevant contextual factors. This poses a fundamental challenge: if choices vary based on presentation, which choice reflects a person’s true preference? This question constitutes the central philosophical dilemma for advocates of libertarian paternalism, who typically interpret biases as errors. This interpretation stems from evaluating behaviour against the standard economic model: if unbounded rationality is taken as the normative benchmark for assessing actual decision-making, then any deviation from it is seen as a failure; an error that, in turn, legitimises intervention. As emphasized by Rizzo and Whitman (2020, p.38), “economists often label people *irrational* or *boundedly rational* without sufficiently emphasizing that this simply means that the individuals are not behaving according to one narrow stipulation of rational behaviour”.

## 4 The Rise of Libertarian Paternalism

As Robert Sugden emphasized in his 2010 Keynes Lecture, until the beginning of the new century, behavioural economics was an almost wholly descriptive enterprise, identifying patterns in individual and small-group behaviour that deviate from traditional rational-choice theories but can be explained by psychology. Despite this fact, its findings posed serious challenges to conventional normative economics (i.e., the *should-be* or deontological perspective of economists on the world). Reflection on these challenges has led to the development of what is now called behavioural welfare economics: normative frameworks that seek to incorporate psychological realities while still offering guidance for policy-making Sugden (2018).

The most influential approach within the realm of behavioural welfare economics is certainly the one known as libertarian paternalism, developed by Cass Sunstein, Nobel price winner Richard Thaler, and others (e.g. Sunstein and Thaler, 2003; Thaler and Sunstein, 2008; Camerer et al., 2003). According to this approach, choice architects are entitled to design choice environments in ways that promote welfare-enhancing decisions, provided that freedom of choice is preserved in the process.

Sunstein and Thaler’s version of libertarian paternalism starts from the consideration that, if individuals possessed ‘complete information, unlimited cognitive abilities, and complete self-

control' - in other words, if they could 'think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi' - they would make different and better choices for themselves. The cognitive limitations from which individuals suffer thus open the door to the possibility of improving their welfare by influencing decision-making in ways that serve their own interests. This can be achieved by modifying the choice architecture - that is, the way options are presented to individuals - while preserving their freedom to choose.

This approach attempts to reconcile two seemingly contradictory commitments: respect for individual choice (the libertarian element) and promotion of individual welfare (the paternalistic element). The libertarian aspect is said to remain intact, as individuals are still free to ignore the intentional modifications of the choice architecture, i.e. no options are removed from the choice set. The paternalistic element, by contrast, lies in the choice architect's presumed ability to determine what individuals would truly want if they were free from cognitive limitations (see the discussion in Rizzo and Whitman, 2020). As Thaler and Sunstein put it, "So long as people are not choosing perfectly, some changes in the choice architecture could make their lives go better (as judged by their own preferences, not those of some bureaucrat)". From this perspective, the planner's task becomes that of reconstructing underlying or latent preferences by simulating what perfectly rational, informed, and temptation-resistant individuals would have chosen. This approach allows libertarian paternalists to retain the normative criterion of preference satisfaction traditionally used in economics, applying it, however, to what they regard as individuals' true underlying preferences (Infante et al., 2016).

## 5 Libertarian Paternalism in Action

The cafeteria example offered by Sunstein and Thaler illustrates how LP operates (Thaler and Sunstein, 2008). A cafeteria manager must decide how to arrange food options. Since some arrangement must be chosen, and any arrangement will influence what people select, the manager might as well organize options to encourage healthier choices. Placing fruits and vegetables at eye level while relegating desserts to less prominent positions nudges people toward healthier selections without removing any options. This example highlights several key features of libertarian paternalism. First, the choice architect (the cafeteria manager, in this case) inevitably influences decisions through environmental design, whether intentionally or not.

Second, the architect can use this influence to promote options that benefit choosers according to their own *true* judgment of personal welfare. Third, since no options are eliminated, the intervention supposedly respects freedom of choice despite steering behavior in a particular direction.

The cafeteria example is particularly interesting because it captures the basic structure of how nudges operate in many relevant contexts. Consider, for instance, organ donation policies. The rules governing organ donation can have a significant impact on donation rates. There is some evidence that countries with opt-out systems (where individuals are presumed to consent to organ donation unless they actively decline) generally achieve higher donation rates than those with opt-in systems (where individuals must actively register as donors). Johnson and Goldstein (2003) suggest that switching from opt-in to opt-out systems can increase potential donor pools by 40-60 percentage points, potentially saving thousands of lives. Similar default effects operate in retirement savings programs, where automatic enrollment substantially increases participation rates compared to opt-in systems.

The justification offered for changing these defaults is that they align choices with what people truly want but fail to achieve due to inertia or procrastination. However, if the power of the default stems instead from a lack of clear preferences or from individuals' inability to resolve a moral dilemma, its use becomes far more problematic, raising concerns in terms of the acceptability and justification of nudge policies. (Beraldo and Karpus, 2021).

## 6 Not All Nudges Are Created Equal

The use of nudges to influence individual behaviour without compromising freedom of choice raises significant questions about the very nature of that freedom (Hausman and Welch, 2010), although public discussion must acknowledge that not all nudges are alike, and it would be inappropriate to regard them as equally threatening to individual liberty. The word *nudge* captures strategies to condition behaviour based on various rationales; these include reminders, warnings, information disclosures, appeals to social norms, and default settings (Sunstein, 2016). While some nudges, like warnings and information disclosure, may bolster an individual's capacity for self-guidance and thoughtful consideration thereby mitigating external control risks, others - defaults, for example - potentially diminish this capacity.

To see the differences between nudges, consider the following three types. The first involves designing environments that highlight motivationally significant properties of options, as in an office building with an attractive, well-lit staircase that encourages physical activity. The second consists in simplifying complex information to aid comprehension, as with traffic-light food labeling systems that intuitively signal nutritional content. Both of these nudges can enhance autonomous choice by increasing the salience of certain option characteristics, thereby helping individuals make more informed decisions.

The third type concerns a software company that sets its privacy settings to share user data by default, knowing that most users will stick with this arrangement due to inertia rather than deliberate choice. The resulting decisions can hardly be considered expressions of genuine preference.

This distinction underscores a fundamental challenge to libertarian paternalism: merely having options available is not sufficient for meaningful autonomy. Genuine freedom of choice requires the capacity to make intentional, reasoned decisions, something that certain types of nudges may support, while others may subtly undermine.

## 7 A Reason-Based Account of Choice

Classical decision theory provides no reason-based account to explain individual choice. It considers preferences as mere matters of taste disciplined by consistency requirements, making the question “*why?*” irrelevant when observing behavior. This approach, however, fails to capture the intentional nature of human agency, or, at best, reduces it to a tautological exercise. As Anscombe (1957) argued, what distinguishes intentional actions from non-intentional ones is that they are “the actions to which a certain sense of the question *why?* is given application; the sense is of course that in which the answer, if positive, gives a reason for acting.”

In his 1993 paper on the internal consistency of choice, Sen (1993) presents an example involving the choice of a slice of cake to illustrate the complexity of decision processes and the limitations of standard decision theory. In this example, a person chooses between slices of cake offered to her. If the goal is simply to get the largest possible slice, then the choices can be straightforwardly evaluated for consistency. However, Sen introduces a twist: he supposes that the person tries to choose as large a slice as possible, but with the constraint of not picking the

very largest slice. This constraint might be due to the desire not to appear greedy, following a social convention or principle such as *never pick the largest slice*. In the case described by Sen, there are three slices of cake, whose size is such that  $z > x > y$ . An individual who tries to choose as large a slice as possible, subject to the constraint of not picking the very largest slice, chooses  $y$  when  $z$  is not in the choice set, and  $x$  otherwise. This pair of choices violates many of the standard conditions of internal consistency. It appears odd also to the common sense that a person who chooses  $y$  over  $x$  given the choice between  $x$  and  $y$ , chooses  $x$  over  $y$  when  $z$  is added to the menu. But the presumption of inconsistency may be easily disputed, in this as in other cases, as soon as we know a bit more about what the person is trying to do, i.e. a bit more about the reasons that motivate her choices. The example illustrates that, when an individual's reasons - possibly grounded in values, scruples, or adherence to social norms - are taken into account, what initially appears to be an inconsistent choice may turn out to be perfectly rational, as it aligns with the individual's broader goals or guiding principles.

Dietrich and List (2016) have developed a reason-based approach to explain individual choices. This approach is able to accommodate apparent inconsistencies in behaviour. The basic idea is that agents perceive the options not as monolithic entities, but in terms of their motivationally salient properties. In this regard, choices are to be understood not as the result of fixed preferences over options, but rather as arising from more fundamental preferences over bundles of motivationally salient properties, with each bundle representing an option. An apple, from this perspective, is more than just a fruit. It possesses multiple properties - such as its colour (red, yellow, green), its method of cultivation (organic or conventional), and others - that may be relevant for consumption decisions. However, not all properties of an option are relevant to the decision maker. Out of the full set of attributes an option may have, only a subset is relevant. Their relevance stems from the fact that these attributes justify choice by appealing to the decision maker's underlying reasons. This justificatory account lies outside the formal structure of the choice problem itself (Sen, 1993), yet it is a crucial element in making behaviour intelligible and rational.

What is key in Dietrich and List's perspective, is that motivationally salient properties are not exogenously given, but endogenously determined by the choice context. When consumers select among various products in a specific setting, such as orange juices in a grocery store,

they perceive each product not simply as an item, but as a combination of properties. While every product possesses numerous such properties, the consumer focuses only on a few of them, i.e. those motivationally relevant. In the grocery store, these might include whether the orange juice is pulp-free, fortified with extra vitamins, or made from organic oranges; but may overlook whether the juice comes in a bottle with an odd number of ounces or whether it's packaged in a recyclable container. Consumers then decide based on a fundamental preference for certain combinations of properties. They choose one orange juice over another in that context - for instance, an organic, vitamin-fortified juice over a regular, pulp-free juice - if and only if their fundamental preference ranks the set of motivationally relevant attributes of the first option, say organic, vitamin-enriched, higher than that of the second, say pulp-free, non-organic. The motivationally salient properties of any option may vary from context to context, and extend beyond the intrinsic properties of the options, including properties which are context-related. Examples are whether an option conforms to a context-specific social norm (e.g., is it polite having the largest slice of cake?), whether it is above average quality among the available options, or whether the choice menu offers luxury options. In Dietrich and List's language, an agent's choice behaviour is said to be reason-based explicable if there exists a motivational salience function which determines, for each choice context, the properties the agent cares about in that context. A motivational salience function coupled with a fundamental preference relation is then called a reasons structure. It is possible to think of each option's bundle of motivationally salient properties in a given context as the option in the subjective sense, as perceived by the agent. Which properties are motivationally relevant depends partly on context, including how options are presented.

In my view, any discussion about autonomy and the darker side of nudges inevitably involves the notion of intentional agency. Put differently, we want agents' autonomy to be preserved, and this is possible only if their actions are intentional.

In the perspective I adopt, intentional action is action motivated by reasons. In the context of choice among alternatives, such reasons must be grounded in the motivationally relevant properties of the options themselves. That is, in order to classify a choice as intentional, there must be a well-defined rationale that illuminates the decision-making process. This requires a clear articulation of the reasons that motivate the agent to prefer one option over the others.

In the reasons-based framework described above, the rationale is grounded in the set of motivationally relevant properties of the options. More specifically, a choice is intentional if there exists a property-based explanation, that clarifies *why* one is motivated to prefer a particular option over others in a given context.

When nudges alter behaviour without modifying the set of motivationally relevant properties of the options, the resulting choices lack intentionality, as they are not explicable in terms of the agent's reasons.

To stress the point, consider a case in which someone switches her choice from option *X* to option *Y* after being exposed to a nudge that does not alter the motivationally relevant properties of the options. In such a case, no reason-based explanation can account for the switch, simply because the properties that justify choice remain unchanged. By this standard, many implementations of libertarian paternalism fail to respect genuine autonomy.

In light of the reason-based approach to choice and intentional agency, the distinction between the three types of nudges discussed in Section 6 becomes particularly relevant. The first two types - those that highlight motivationally relevant properties or simplify complex information - operate by making certain features of the available options more salient to the agent. As such, they enhance the agents' capacity to act intentionally by helping them recognize and act upon reasons they already care about (what in Dietrich and List (2016)'s theory would be the *fundamental preference relation*). These nudges do not undermine autonomous choice. Indeed they preserve the link between the agents' decisions and the reasons that justify them. Any given choice can be rationalized by a property based account of why a given option is preferred.

The third type of nudge, by contrast, exerts influence not by enhancing the agent's awareness of motivationally salient properties, but by exploiting cognitive inertia or decisional passivity. In such cases, the choice does not reflect any underlying evaluative judgment or reasoned preference; it cannot be rationalized, in other words, on the basis of a property-based account of why the agent prefers one option over others. As a result, the decision cannot be considered intentional. Such nudges threaten individual autonomy.

## 8 The Challenge of Digital and Algorithmic Nudging

AI-powered nudges represent the new frontier in behavioural influence, combining the logic of choice architecture with the predictive and adaptive capabilities of artificial intelligence. Unlike traditional nudges, AI-powered interventions are often personalised, context-sensitive, and dynamically updated. On social media platforms such as Instagram, TikTok, or YouTube, algorithms recommend content based on users' past behaviour and preferences, subtly guiding not only entertainment consumption but also political opinions and adherence to social norms, potentially including discriminatory ones. In e-commerce environments, platforms like Amazon use AI-driven recommendation systems to structure product visibility—through highlights, rankings, and interface design, in ways that leverage cognitive biases such as anchoring or default effects. In the health and wellness domain, apps like Headspace or Fitbit send timely, data-informed notifications designed to nudge users toward healthier habits precisely when they are most likely to respond.

AI is also used to personalise default options in sectors such as banking, insurance, or digital privacy, where pre-selected configurations are tailored to user profiles but may align more closely with the provider's interests than the user's own goals. In the public sector, nudging techniques are deployed through urban informatics systems to encourage desired behaviours (e.g., choosing less congested routes or complying with health recommendations) based on real-time data. Finally, in educational technologies like Duolingo or Coursera, AI systems personalise learning pathways and deliver motivational prompts to sustain engagement and progress.

Across these domains, what distinguishes AI-powered nudges is their capacity to adaptively target individuals, often in ways that blur the line between assistance and manipulation, raising renewed concerns about autonomy, transparency, and informed consent.

The scale of digital nudging further distinguishes it from traditional interventions. While physical choice environments affect only those present in a particular location, digital nudges can reach billions of users simultaneously. This magnifies potential risks.

## 9 Toward Autonomy-Respecting Choice Architecture

If many nudges undermine rather than enhance autonomy, how should we approach behavioural interventions? One possible answer is to adopt what might be called an *autonomy-respecting choice architecture*, interventions specifically designed to engage, rather than bypass, individuals' reasoning faculties. Such an approach could be voluntarily embraced by both private and public organizations as part of a broader ethical commitment to preserving individual agency. This approach would favour interventions that support decision-making capacities - such as simplifying complex information or highlighting relevant considerations - over those that exploit inertia or the tendency to follow habitual patterns. And it would maintain genuinely neutral defaults where possible, or implement active choosing requirements for significant decisions. In the organ donation context, for example, this might mean adopting mandated choice systems rather than opt-out defaults. Individuals would be required to actively select their donation preference when renewing driver's licenses, ensuring that donation status reflects deliberate consideration rather than passive acceptance of defaults. While this approach might not maximize donation rates, it would better respect autonomy by ensuring that outcomes reflect intentional choices (in my paper with Jurgis Karpus we argue in favour of a system of voluntary active choice, presenting people with the choice of registering as an organ donor or not, but not requiring them to make a decision). In digital environments, autonomy-respecting design might include clearly labeling personalized recommendations as such, explaining the basis for algorithmic suggestions. These approaches would acknowledge the inevitability of some influence while ensuring it operates in ways compatible with user agency.

## 10 Conclusion

Behavioral economics has fundamentally challenged the traditional view of individuals as perfectly rational agents, opening the door to nudge-based interventions that leverage psychological insights to influence choice. While proponents of libertarian paternalism argue that such nudges preserve freedom while promoting welfare, a closer examination reveals that many implementations fail to respect genuine autonomy. The criterion for distinguishing between acceptable and unacceptable nudges lies in whether they influence choice by altering the set of motivationally

relevant properties in ways that engage decision-makers' reasoning faculties. By this standard, libertarian paternalism often fails to remain genuinely libertarian.

A reasons-based understanding of autonomy offers a framework for designing choice environments that truly respect individual agency.

The challenge ahead lies not in abandoning behavioral insights but in harnessing them in ways that enhance rather than diminish human agency. This requires a shift from viewing nudges primarily as welfare-enhancing tools to seeing them as potential supports for autonomous decision-making. By prioritizing transparency, reason-engagement, and genuine choice, it is possible to develop approaches that respect individuals as reasoning agents rather than treating them as objects to be steered toward predetermined outcomes.

## References

- Anscombe, G. (1957). *Intention*. Oxford: Blackwell.
- Beraldo, S. and J. Karpus (2021). Nudging to donate organs: do what you like or like what we do? *Medicine, Health Care and Philosophy* 24(3), 329–340.
- Camerer, C., S. Issacharoff, G. Loewenstein, T. O’Donoghue, and M. Rabin (2003). Regulation for conservatives: Behavioral economics and the case for ‘asymmetric paternalism’. *University of Pennsylvania Law Review* 151, 1211–1254.
- Dietrich, F. and C. List (2016). Reason-based choice and context-dependence: An explanatory framework. *Economics and Philosophy* 32(2), 175–229.
- Hausman, D. M. and B. Welch (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18(1), 123–136.
- Infante, G., G. Lecouteux, and R. Sugden (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology* 23(1), 1–25.
- Johnson, E. J. and D. Goldstein (2003). Do defaults save lives? *Science* 302(5649), 1338–1339.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kahneman, D., J. L. Knetsch, and R. H. Thaler (1991, March). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives* 5(1), 193–206.
- McNeil, B. J., S. G. Pauker, H. Sox, and A. Tversky (1982). On the elicitation of preferences for alternative therapies. *The New England journal of medicine* 306(21), 1259–1262.
- Rizzo, M. J. and G. Whitman (2020). *Escaping Paternalism*. Cambridge University Press.
- Samuelson, W. and R. Zeckhauser (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty* 1(2), 7–59.
- Schmauder, C., J. Karpus, M. Moll, B. Bahrami, and O. Deroy (2023). Algorithmic nudging: The need for an interdisciplinary oversight. *Topoi* 42(3), 799–807.

- Sen, A. (1993). Internal consistency of choice. *Econometrica* 61(3), 495–521.
- Sugden, R. (2018). *The community of advantage*. Oxford University Press.
- Sunstein, C. R. (2016). Autonomy by default. *The American Journal of Bioethics* 16(11), 1–2.  
PMID: 27749176.
- Sunstein, C. R. and R. H. Thaler (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review*, 1159–1202.
- Thaler, R. and C. Sunstein (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.
- Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science* 211(4481), 453–458.