AI & SOCIETY
https://doi.org/10.1007/s00146-023-01846-3

MAIN PAPER

# Socratic nudges, virtual moral assistants and the problem of autonomy

Francisco Lara[1] · Blanca Rodríguez-López[2]

Received: 16 October 2022 / Accepted: 14 December 2023

**Abstract**
Many of our daily activities are now made more convenient and efficient by virtual assistants, and the day when they can be designed to instruct us in certain skills, such as those needed to make moral judgements, is not far off. In this paper we ask to what extent it would be ethically acceptable for these so-called virtual assistants for moral enhancement to use subtle strategies, known as "nudges", to influence our decisions. To achieve our goal, we will first characterise nudges in their standard use and discuss the debate they have generated around their possible manipulative character, establishing three conditions of manipulation. Secondly, we ask whether nudges can occur in moral virtual assistants that are not manipulative. After critically analysing some proposed virtual assistants, we argue in favour of one of them, given that by pursuing an open and neutral moral enhancement, it promotes and respects the autonomy of the person as much as possible. Thirdly, we analyse how nudges could enhance the functioning of such an assistant, and evaluate them in terms of their degree of threat to the subject's autonomy and their level of transparency. Finally, we consider the possibility of using motivational nudges, which not only help us in the formation of moral judgements but also in our moral behaviour.

**Keywords** Nudges · virtual moral assistants · moral enhancement · autonomy · ethics of AI

## 1 Introduction

Nowadays, many of our daily activities are more convenient and efficient thanks to virtual assistants. By interacting with them we can locate our destinations, control the security of our home and the different devices in it, and access suggestions about our favourite music. But the day when assistants can be designed to instruct us in certain skills, such as those needed to make moral judgements, is not far off. Through conversation and interaction with these assistants, humans could become better informed and equipped to deliberate on what is right and wrong. This would be the realm of what is becoming known as "moral AIenhancement". In this article we propose considering to what extent it would be ethically acceptable for these so-called virtual assistants for moral enhancement to use subtle strategies, quite common in the fields of commerce and healthcare, known as "nudges", to influence our decisions. These involve making changes in the context of choice that are supposed to influence the behaviour of decision-makers so that (in non-commercial settings) they end up choosing what is in their best interest, but without introducing coercive mandates or increasing incentives.

To achieve our objective, we will first characterise nudges in their standard use and discuss the debate they have generated around their possible manipulative nature. In principle, if this manipulative character is true, this would already be a major obstacle to our project of using them for moral enhancement through AI. In order to see how this obstacle could be overcome, we will set out below the conditions we consider important to have in place. First, it will be necessary to determine the virtual assistant from which we should start. Not all types of moral AIenhancers may allow the use of ethical nudges. Once the ideal type of assistant that should accompany the nudges has been determined, we will establish the requirements that these should meet in order not to be considered manipulative. Finally, we will propose some appropriate nudges for the selected assistant and evaluate to

✉ Francisco Lara
flara@ugr.es

[1] Department of Philosophy I, University of Granada, Granada, Spain

[2] Department of Philosophy and Society, Complutense University of Madrid, Madrid, Spain

Published online: 27 January 2024

what extent they would meet those requirements and, in that regard, could be considered ethically acceptable.

## 2 Nudges and autonomy

Nudges, as strategies to subtly and indirectly influence our decisions, have long been of political interest and many countries already have specific departments in their governments to design and implement them in order to help citizens make better decisions for themselves and society as a whole. In academia however their study is much more recent, and has been developed in particular from behavioural science and economics.

But what are nudges really? They are strategies for influencing personal choices that take advantage of a human reality: the limitations of being rational[1] in deliberations and volitional acts. The complexity of the problems we face, lack of time, exhaustion or laziness prevent us from always being rational in our decisions; therefore, we are forced to resort to decisional shortcuts (unreflective and unconscious) with which to face these limitations, saving cognitive (concentration, acquisition and evaluation of information) and volitional efforts. The existence of these decisional shortcuts is corroborated by the discovery of two ways, present in all humans, of processing information and making decisions: one automatic and the other reflexive (Thaler and Sunstein 2008). Ultimately, these are the two ways of knowing and making decisions popularised by Kahneman (2012): "system 1" (S1) and "system 2" (S2), respectively. S1 or "automatic" is embodied in spontaneous, effortless, rapid, associative and unconscious thinking. To facilitate this rapid form of thinking, S1 uses cognitive barriers, heuristics and rules of thumb that allow instinctive or habitual responses to be made according to certain cues and without using all available information. S2, or "reflective", however, is more deliberate thinking: controlled, effortful, slow, deductive and conscious. We use S1 more often because it requires less effort and time, but at the cost of producing systematic biases, which can sometimes lead to irrational behaviour that is harmful to society or inconsistent with our long-term goals. In that sense, it can be said that these shortcuts, while unavoidable and pervasive, sometimes end up being negative and biased. Think, for example, of judging others according to stereotypes. The savings of mental resources that these shortcuts in judgement entail (and which in many cases may be accurate) may on occasions, when the individual is unrepresentative of the stereotype in question, be misguided and discriminatory.

The central assumption in nudge theory is that, rather than avoiding or combating the unreasonableness of S1, it should be accepted and used in a positive way. Standard nudges are then conceived as attempts, through small changes in the environment or "choice architecture", to activate those heuristics of S1, so that the behaviours of agents are influenced in a predictable way to lead to a more advantageous decision for them (in terms of health or financial improvement) or society (Thaler and Sunstein 2008).[2]

The basic notion in nudges is that of "choice architecture", referring to the conscious and deliberate attempt to shape the context in which people make decisions, rather than to alter or extend choice options. Nevertheless, it is important to note that not all forms of altering choice architecture are nudges. To be so, they must be easy to avoid, not eliminate options or constitute mandates (there is no coercion in nudges), and not introduce substantially new incentives. Examples of nudges are: healthy products being placed more visibly in self-service cafeterias to encourage good eating habits, or smaller plates being offered to "nudge" to unconsciously reduce calorie intake; children being offered duck-shaped carrots to encourage them to eat them; and default participation in a pension plan being added to employment contracts, or, in laws, the establishment of default consent to organ donation (Thaler and Sunstein 2008).

From an ethical point of view, the main problem with nudges revolves around the autonomy of the subject, insofar as the aim is to increase the probability that an option will be chosen by pushing the individual in one specific direction.

[1] Referred to in the literature as limited rationality, which contrasts with perfect rationality. In neoclassical economics, perfect rationality is what characterizes the human model represented by *Homo Economicus*, a hypothetical being who has perfect decision-making conditions, basically meaning complete information about the alternatives and perfect knowledge of the possible consequences of all his actions, which allow him to always act maximizing his utility. Real people are different (we are *humans* and not *econs*, in the terminology of behavioural economics). We have bounded rationality.

[2] We will not focus here on the use of nudges for strictly commercial purposes. Regarding non-commercial nudges, there are two versions: pro-self and pro-social. Originally, the proposal of nudges was linked to the theory of libertarian paternalism and therefore they are, in their aims, essentially paternalistic, having the goal of protecting individuals from their own mistakes and helping them make better decisions for their own well-being, but also libertarian, always leaving their freedom of choice intact (Sunstein 2015, pp. 7–8). However, Thaler and Sunstein (2008) already give many non-paternalistic examples in this sense (e.g., conservation of public spaces or payment of taxes, in ch. 2). There are even two chapters devoted to organ donation (chapter 11) and environmental care (chapter 12). In these cases, the aim of the nudges is to direct the agent's behaviour towards pro-social ends. Also, in chapter 16, there are several ideas of nudges that could be considered truly moral, such as those aimed at encouraging donations to charities. This distinction between pro-self and pro-social nudges has not gone unnoticed, and some authors characterise libertarian paternalism as the "advocacy of governmental use of pro-self nudges" (Barton and Grüne-Yanoff 2015, p. 344).

Some authors point out that this difficulty can be overcome if we consider that this "pushing" is carried out, especially in pro-self nudges, with the aim of directing behaviour within a libertarian paternalism view that, while intending the best for the individual, always preserves their freedom to oppose (Thaler and Sunstein 2008, pp. 4–6). According to Thaler and Sunstein (2008, p. 5), when third parties are not at risk and the well-being of decision-makers is the only relevant matter, the goal of the nudge is "to influence choices in a way that will make choosers better off, as judged by themselves".

However, it could be argued that the freedom of choice preserved by nudges is insufficient because of the way in which they attempt to influence individuals to improve their decisions and behaviour. Arguably, although individual freedom of action is not reduced in the absence of coercion (they still have the option to do the opposite of what is intended by the nudge), their autonomy would be put at risk. But in what sense would nudges threaten the agent's autonomy? We believe that two versions of the objection tend to converge, which should, however, be differentiated.

One of them argues that nudges threaten autonomy because they prevent reflection. Instead of trying to influence with reasons, they take advantage, with imperceptible strategies, of unconscious and unreflective psychological mechanisms that do not require any effort or deliberation, thus reducing the agent's involvement in the decision-making process (Bovens 2009; Grüne-Yanoff and Hertwig 2016). As Bovens (2009) says, "what is driving my action does not constitute a reason for my action". This intended adherence of the individual would correspond more to immediate, superficial and blind acceptance than to a reflective personal identification therewith.

In short, nudges would be ethically questionable according to this first version of the critique, because the tactic that characterises them in itself entails an attempt to circumvent the deliberative capacities of the individual, thus significantly limiting autonomy (Ashcroft 2013; Bovens 2009; Conly 2012; Glod 2015; Hausman 2018; MacKay and Robinson 2016; Saghai 2013; White 2013; Wilkinson 2013; Yeung 2012).

However, a nudge, as we have characterised it, does not threaten autonomy just by dodging reflection, for while it may not increase this, it can hardly be said to decrease its degree in the agent. If I tend to pick the closest thing in sight in a cafeteria, I am still just as autonomous whether I choose the fruit as it is placed before me by chance or as part of a nudge strategy.

Moreover, it should be added that, despite the standard characterisation we have put forward, some nudges can be considered as reflexive. As opposed to standard nudges, which by changing choice architecture aim to use heuristics to guide people's behaviour in a specific direction without involving any reflective (conscious and deliberate) thought, there would also be those that would use the mechanisms of S1 to induce reflection. Barton and Grüne-Yanoff (2015) distinguish these two types of nudges as "heuristics-triggering" and "heuristics-blocking", respectively, which Hansen and Jespersen (2013, pp. 14–15) prefer to refer to as "Type 1" (T1) and "Type 2" (T2). T1 nudges are more effective in situations where there is the pressure of decision making with a high cognitive load, which is difficult to manage given our memory limitations, while T2 nudges are more useful when the important thing is to achieve persistent, long-term behavioural changes (Hansen and Jespersen 2013; Weijers et al. 2020).

T2 nudges or heuristics-blocking thus aim to use the automatic knowledge system so that the agent precisely ends up blocking the shortcuts that such a system uses. They attempt to turn that which is initially unconscious into a deliberate choice. The best known example of this type of nudge is the use of a sticker in the urinal with the image of a fly. In this way, S1 mechanisms are used to attract the user's attention in order to trigger an attentive action to the act of urination and even a reflection on the meaning of the sticker, which ultimately leads to cleaner urinals. Another example of T2 nudges would be to shorten the side lines on a road in order to get drivers to feel a sense of increased velocity and automatically reduce their speed as a consequence. In this case, conscious attention or possible reflection would come after the reduction, when the driver perceives that his or her quick response had been caused by an optical illusion (Hansen and Jespersen 2013, p. 15). In these cases, heuristics (quick attention to unfamiliar images or quick reaction to sensations of danger) are being used to paradoxically block heuristic behaviour and, with the activation of a certain reflective and active attitude, "push slightly" towards more rational decisions and behaviour (keeping urinals clean and moderating speed). These would therefore be nudges that, contrary to T1, aim to preserve or increase the freedom of the individual and can be considered libertarian "understood in its thicker sense as autonomy-respecting" (Yeung 2012, p. 137).

But we think that this typology is incomplete if we do not add nudges that subtly attempt to provoke a predictable behaviour as a result of an intended act of reflection, only now without seeking an activation of particular S1 mechanisms. These would be those that, for example, promote decision-making resulting from previous cool off periods or access to data or reminders. They would be those characterised by some authors as educational or informing nudges (Barton and Grüne-Yanoff 2015) and which we will also refer to here as T3.

In short, T1 nudges by themselves do not threaten autonomy by preventing reflection on the part of the agent, and T2 and T3 nudges can be considered autonomy-promoting instruments in the sense that their modus operandi is based on influencing a more deliberate decision by the agent.

A second version of the objection that nudges pose a threat to autonomy would be based on their alleged manipulative nature. A certain strategy can be said to seek to influence agents when it meets certain conditions. The first is that it seeks to influence the agent with criteria that are alien to him or her, be these business interests, conceptions of the common good or moral principles. This would, in principle, exclude from the charge of manipulation nudges that seek to influence agents to make decisions more in line with their own interests (when these have not been sufficiently considered by them), as liberal paternalism argues.

A second essential condition of manipulation, the presence of which may be an objection to all types of nudges, would be their lack of transparency, that is that they attempt to influence the subject's decisions in a way that is not obvious to the subject, entailing hidden or disguised deceptions (Thaler and Sunstein 2008, p. 244; Blumenthal-Barby and Burroughs 2012; Grüne-Yanoff 2012; Hausman and Welch 2010), and evidencing a certain disregard or disrespect for individuals as beings capable of making rational decisions in their own affairs (Hausman and Welch 2010; White 2010; Yeung 2012).

To avoid this reprehensible lack of transparency of nudges, Thaler and Sunstein (2008, p. 244) argue that, when used by states, they should be governed by J. Rawls' principle of publicity, which bans governments from selecting policies they would not be able or willing to defend publicly to their own citizens. This principle involves the idea of respect. “The government should respect the people whom it governs, and if it adopts policies that it could not defend in public, it fails to manifest that respect. Instead, it treats its citizens as tools for its own manipulation. In this sense, the publicity principle is concerned with the prohibition on lying. Someone who lies treats people as means, not as ends” (244–5).

However, the principle of publicity might not be sufficient. We can illustrate this point with subliminal advertising, an example of a decision-influencing strategy that is clearly inadmissible due to its manipulative character, and which could meet the requirement of Rawls' publicity principle. Imagine a government announces it is going to use subliminal advertising to promote something beneficial to all, such as toothbrushing. It could defend this publicly without any problems, but citizens would have to trust the government that its use only be for that purpose, as they have no technical means to check at any given moment whether such technology is being employed and for what reason. This lack of transparency, by not allowing its use to be monitored, would make it highly vulnerable to abuse and therefore illegitimate, which is why Grüne-Yanoff (2012) considers any nudge that is not transparent to be manipulative.

But the principle of publicity and the monitoring requirement presented as an alternative to it are not the only options for taking a position on the transparency or non-transparency of nudges.

To some extent, nudges, at least the T1 variety, cannot be entirely transparent. If they were they would cease to be functional, as their very operation relies on heuristics that are often unconscious (Marchiori et al. 2017). As Bovens (2009) argues, "they are most effective in the dark". However, this "obscurity" admits degrees and types. With this in mind, perhaps we can establish a transparency requirement for nudges in a weak sense. Thus, it could be a requirement that even if nudges were " dark " in their immediate presentation, they should be evident at a later point in time and provided certain conditions are met. In line with the intention of Thaler and Sunstein, Bovens (2009) suggests that for nudges to be truly transparent a watchful agent should be able to identify the intention of the underlying decisional architecture change, and choose not to be influenced by the nudge. Similarly, Hansen and Jespersen (2013) define a transparent nudge as "a nudge provided in such a way that the intention behind it, as well as the means by which behavioural change is pursued, could reasonably be expected to be transparent to the agent being nudged as a result of the intervention" (Hansen and Jespersen 2013, p. 17). This weak transparency requirement would be fulfilled by all T2 and T3 nudges, such as the fly on the urinal or colour labelling of unhealthy products. T1 nudges, however, would have more difficulty in meeting this requirement. This would be the case with subtly elaborated formulations of certain questions that lead to predictable answers, exposing people to pictures of faces that make them more cooperative or changing the seating arrangement in a classroom to reduce bullying (Weijers et al. 2020).

The third condition that would define a strategy as manipulative is that it be designed to diminish the degree of resistance the subject may have towards following the intended orientation of his or her behaviour. A nudge is usually characterised as a strategy that, while aiming to influence the agent, must maintain freedom of decision. However, this is usually interpreted in an undemanding sense: it is only a requisite that there be no coercion. But sometimes the influence exerted by the nudge can be so strong that freedom of choice is not effective. Therefore, a requirement that would guarantee autonomy, even if it circumvented certain rational capacities, would be that the influence exerted by the nudge be easily resistible. This would require that the nudge should not undermine in the nudgee either the ability to become conscious or attentive to pressure, or the ability to inhibit the propensity that the nudge exploits. When this happens, it could be said that the uses of cognitive shortcuts preserve autonomy because nudgees are in control of their choices, that they have a real opportunity to dissent from what the nudge intends (Saghai 2013).

## 3 Nudges for virtual assistants

Having outlined the two possible interpretations of the objection that nudges pose a threat to the autonomy of human beings, we have given reasons to grant more plausibility to that which poses the threat in terms of manipulation (rather than avoidance of reflection). The question now is whether there can be nudges in moral virtual assistants that are not exposed to this interpretation of the objection, i.e., that are not tainted with manipulation.

To this end, it is important to clarify two issues. First, to what extent the use by these assistants of a new technology, such as AI, could substantially modify their ability to nudge users and, above all, whether this in itself would not make them potentially more manipulative entities. In the literature on the subject, this use of AI is often captured in what is generically known as "digital nudging" and in a particular type of nudging that has been termed "hypernudging".[3] Both terms refer to the possibilities of influencing behaviour in a similar way to standard nudging, but in these cases within a 'digital environment'. Thus, digital nudging aims to direct user behaviour in general by virtue of different user interface design elements, for example, when mobile payment apps include a default tip in the payment order (Weinmann et al. 2016). However, with hypernudging, various data processing techniques are used so that the influence on the user is governed by a constant personalisation and updating of choice architectures (Yeung 2017; Morozovaite 2021), especially present in recommender systems (Lanzing 2019; Jesse and Jannach 2021).

Digital nudges, including hypernudges, have developed as we make more and more decisions on screens, such as websites or mobile apps, ranging from choosing travel, insurance, all kinds of products and even a partner. In the digital world, even worse decisions can be made than in the real world, because due to the large amount of information on the internet, users may fail to notice the relevant details to reach an optimal decision. Rather, decisions on screens are often made in a hasty and automated way (Benartzi and Lehrer 2015). In this context, nudging can be a great tool for enhancement of users' decision making. Compared to offline scenarios, the implementation of digital nudges is easier, faster and cheaper, and thanks to the user tracking enabled by the internet, nudges can be personalised, as we have seen with hypernudges, and thus achieve their goals in a much more effective way (Mirsch et al. 2017). Thus, smartphones, wearables and internet of things technologies allow the monitoring of our activity, and from it we can derive personalised advice that aims to positively influence our behaviour, making it, for example, healthier.

But it is also true that the influence of new technologies in this area may not always be beneficial. One example is the construction of digital environments known as "dark patterns". These are strategies that :knowingly confuse users, make it difficult for users to express their actual preferences, or manipulate users into taking certain actions" (Luguri and Strahilevitz 2019), or where "designers use their knowledge of human behaviour and the desires of end users to implement deceptive functionality that is not in the user's best interest" (Gray et al. 2018). An example of the latter would be websites that nudge shoppers to make quicker and therefore less thoughtful decisions by displaying countdown timers or misleading stock reports, like "Only 1 left!" (Susser and Grimaldi 2021: 246).

However, the potential manipulative risks of these new nudges have been pointed out most often in reference to hypernudging. Yeung (2017: 119) has indicated that "by configuring and thereby personalising the user's informational choice context, typically through algorithmic analysis of data streams from multiple sources claiming to offer predictive insights concerning the habits, preferences and interests of targeted individuals, these nudges channel user choices in directions preferred by the choice architect through processes that are subtle, unobtrusive, yet extraordinarily powerful". In addition to this real time, personalised feedback dynamic and their predictive capacity by virtue of algorithms that "learn" from collected data and allow for a constant reconfiguration of the choice architecture of individuals, for some authors hypernudges would be particularly problematic because of their hiddenness. They argue that, compared to standard nudges, which although not immediately detectable must be visible in some way in the physical world, hypernudges would be hidden in a more sophisticated way as they are embedded into the design of complex, machine learning algorithms, which are highly opaque. This could undermine the transparency requirement discussed above and render users unable to determine whether hypernudges respond more to illegitimate intentions of political institutions or companies than to user welfare (Yeung 2017; Lanzing 2019: 555; Mills: 6–7).

All these digital nudging and hypernudging strategies are specific to contexts and interfaces very different from the one we are going to examine here, occurring above all in internet activities and from the not at all explicit intentions of certain organisations to influence us in order to satisfy their own interests, which can even give rise to such clearly manipulative practices as dark patterns. The context that would be most relevant to our interest in moral virtual assistants would be that of devices that recommend healthy practices by virtue of monitoring our habits. In these cases, as in the case of moral virtual assistants, the efficiency of

[3] Mills (2022) disagrees, however, with this conception of "hypernudging" as a type of "digital nudging", even if, as he himself acknowledges, his interpretation is not the most widespread.

updating decision architectures according to personal profiles does not necessarily entail a lack of transparency. Both the target behaviour (e.g. to reduce carbohydrate consumption) and the means employed (e.g. emoticons with sad faces when this reduction is not being achieved) can be perfectly transparent.Ignorance of the algorithmic formulas that lead to these architectures and their changes should not be an impediment for the user to perceive their meaning. To hold otherwise would be tantamount to maintaining, in the case of standard nudges, that their effectiveness is hidden from users by virtue of their ignorance of the psychological or even physical laws that explain the changes in choice architectures. If this is true, and if it can be established that what is important is not the transparency of the algorithms that explain the changes, but the meaning of these changes, it could be inferred that, in principle, digital nudges could meet the same weak transparency requirements that we defended in the previous section for standard nudges. In Sect. 3 we will examine how these requirements can be met in the case of moral virtual assistants.

The second issue to be taken into account prior to arriving at the objective of this section, which is to determine to what extent the use of nudges by virtual assistants may entail manipulation, is the essentially educational status of the type of assistant we are interested in. Thus, the basic aim would be to use AI to increase moral capacities and predispositions by means of more or fewer robotic systems. Therefore, the appropriate nudges for these systems would not be strategies to promote, through political management, decisions by citizens more in line with their interests or those of the community, without this necessarily implying better skills or predispositions. Rather, they would aim to modify the "choice environment" in order to subtly influence behaviour, but as a consequence of the acquisition of stable decision-making patterns through them. We will therefore attempt to discover whether these types of assistant-accompanying nudges could truly be enhancement strategies that do not entail an unacceptable reduction in autonomy.

The not very extensive literature on the subject is dominated by the idea that these nudges could form part of robots designed to promote the necessary attitudes and skills for humans to behave following some ethical standards, such as Rawls' principles of justice (Borenstein and Arkin 2016a, 2016b) or stoic practice (Klincewicz 2019). Although these proposals would specifically target robots to take advantage of certain benefits of humanoid chassis, such as emotive influence (Asada et al. 2009) or the inspiration of more authority (Aroyo et al. 2018), they could also be implemented in simple computer programmes that assist users by guiding their behaviour (Klincewicz 2019, pp. 426–427).

The main problem with the use of nudges in these social robots is that personal autonomy is put at risk since they reproduce in the field of moral enhancement the same scheme that underlies standard nudges, used in political management, of intervention according to a behavioural guideline previously established by a third party. This already entails a certain manipulative character because such nudges are intended to impose criteria that are alien to the subject.

In the case of Borenstein and Arkin (2016a, b), nudges would be oriented towards the enhancement of humans according to substantive ethical principles or attitudes selected by a designer. Such values do not necessarily coincide with those of the individual who is going to be nudged, which would imply some kind of value imposition. Thus, Borenstein and Arkin argue (2016a) that social companion robots could nudge individuals, especially children, to dislike inequality, for example by smiling or displaying other social cues that encourage the sharing of toys between playmates, or by mimicking expressions of disappointment if a child refuses to share. They could also influence children to distance themselves from possible parochialism, for example, by nudging them to interact with each other. They are convinced that encouraging such attitudes would morally enhance individuals by making them more concerned with social justice according to Rawls, and in particular with the second principle of his theory, according to which social and economic inequalities are only fair if they result in compensatory benefits for all and in particular for the least advantaged (Rawls 2009, p. 13). In another publication, Borenstein and Arkin (2016b) show a preference for nudges, rather than directly modifying specific attitudes, to affect the capabilities that make them possible. They specifically argue that robots, through affective computing, should foster empathy that is responsible for the performance of charitable acts, and promote the good of society. Whether the emphasis is on the attitudes themselves or on the mechanisms or capacities that cause them, the problem is that these are substantive proposals in the sense that they are debatable in themselves, without leaving open the possibility that it might be questioned whether these principles and empathy are the foundation of morality (Klincewicz 2019, pp. 430–432). Moreover, these proposals would be debatable from the perspective of the relevance of autonomy because their principles and motivations need not coincide with the subject's chosen or eligible moral perspective.

On the other hand, the robotic nudgers proposed by Klincewicz should be designed to promote in users practices or skills with clear stoic profiles: to make them differentiate between what does and does not depend on them (by inducing them to reflect on the valuation of possible control or by the robot itself doing the reflection and confronting it with the subject's valuation); by asking users to commit to imagining the loss of things that are valuable to them; asking and guiding them to review the day and stay in the present (reminding them of their place in the world and preparing them for a mindful confrontation with the world); or encouraging them to reflect on the causes of their emotional disturbances (Klincewicz 2019, pp. 436–438).

Klincewicz justifies these moral nudges by virtue of their supposed neutrality and usefulness for moral enhancement. Of them he says that in order to differentiate his proposal from that of Borenstein and Arkin, "instead of being designed to promote Rawlsian principles or empathy they would need to be designed with strategies for improvement of other psychological capacities relevant to moral behaviour and moral decision-ranking" (Klincewicz 2019, p. 435). But this proposal is neither neutral—since it opts for a particular, stoic, way of understanding the good life and right behaviour -, nor does it guarantee moral enhancement—since the subject in question, by adopting these stoic practices, could be more equanimous and willing to be content with what he or she has and can do, but this would not necessarily lead to better moral decisions.

Therefore, we believe that digital nudges would only respect people's autonomy if we start in principle from the conception of a virtual assistant (or social robot) whose ultimate purpose is not the direct influence on actions or attitudes according to previous principles alien to the agent, but rather the achievement, by means of instruction, of certain stable capacities and predispositions in the user, which allow them to be critical of other people's and their own moral approaches. An example of such an assistant is that proposed by Lara and Deckers (2019). It would seek moral enhancement through constant interaction with the user, in the Socratic style, and from strictly procedural criteria, such as empirical, conceptual, logical or ethical-discursive rigour. This type of assistant, which we will call SocrAI, is justified precisely by virtue of its commitment to an "open" and neutral moral enhancement that promotes and respects the autonomy of the individual to the maximum (Lara 2021, pp. 12–17).[4]

The question now arises as to what the nudges that could be incorporated into SocrAI would be like if we want to free them from the accusation of being manipulative and therefore infringing on personal autonomy. By being incorporated into an assistant that aims for neutrality and therefore avoids the influence of criteria external to the agent, we can already say that it fulfils the first requirement for not being considered manipulative. It remains for us to check whether, once we know their possible design, they would also meet the requirements of transparency and easy resistibility.

## 4 Socratic nudges

In what ways could nudges enhance the functioning of an assistant such as SocrAI? In principle, they could be useful for overcoming obstacles to moral learning. Let us first see what such obstacles would be and what nudges could be incorporated into the design of the assistant to successfully overcome them. Then, we will evaluate them in terms of their degree of threat to the subject's autonomy.

Some nudges could be devised to neutralise common biases in our decisions. As we have seen, heuristics are tools that in principle facilitate learning, reducing the costs of constant rational and conscious deliberation. But these heuristics are not always reliable and can sometimes give rise to biases that make decision-making irrational.

In the case of our interaction with SocrAI, there could be the following tendencies contrary to the rational formation of judgements and, thus, to the learning of moral skills and attitudes. On the one hand, there are anchoring biases, meaning humans tend to "anchor" their decisions in irrational judgements of probabilities based on data and solutions to problems in which, because of their greater knowledge or occurrence, they end up placing excessive trust. These biases would lead to a second type of bias, based on the backfire effect. This is a heuristic that, based on anchoring, leads users to adopt more "defensive" attitudes of not only disregarding other people's testimonies that are obvious, but of reinforcing their own beliefs even more when presented with strong evidence to the contrary.

A third type of bias worth noting here is representativeness bias, which leads us to give a determining role in our decisions to the beliefs, values and behaviours of the majority.

To overcome all these biases, nudges would be useful for reinforcing the relevance of the data and the recognised authority of the sources, so as to facilitate the emergence in the user of a self-critical attitude with its anchors and strong external influences. They would basically consist of informative indications which, by highlighting the credibility of the sources, would predispose the user to be more open towards considering opposing positions and arguments. But they could also consist of explicit indications as regards the influence of the bias in question on the subject's deliberations, as well as, in keeping with the assistant's Socratic method, frequent rebuttals that would undermine persistent adoption of such unconscious and automatic heuristics on the part of users. On the other hand, such nudges could also take the form of warnings about the inconsistencies of these biased predispositions with other approaches previously held by them. Thanks to the machine's ability to quickly process past data about the behaviour of users in previous interactions and taking advantage of the habitual tendency of humans

[4] To avoid misunderstanding, we are talking about ‘axiological neutrally’. We do not mean that AI systems in general are “neutral” technologies that are not imbued with values. Rather, Lara and Deckers (2019) use this expression to refer to one of the classic requirements of procedural ethics, namely, the absence of substantive position-taking at the normative level.

to present themselves and others as coherent agents, SocrAI could make users think about the irrationality of their persistence in certain anchors and backfire effects.

Nevertheless, obstacles to learning an open morality may not only come from these habitual tendencies to take cognitive shortcuts in order to avoid certain mental costs; they could also be due to negative physiological, psychological and environmental conditions for moral deliberation, such as lack of sleep (Olsen et al. 2010), hunger (Danziger et al. 2011), alterations in neurotransmitters and hormones such as serotonin (Tse and Bond 2002) or oxytocin (de Dreu et al. 2010; Lara 2017), noise (Berkowitz 1993) or heat (Anderson et al. 1995). The assistant, through its monitoring of the user's physiology, mental states and environment, could indicate the degree of appropriateness of the deliberation conditions (Lara and Deckers 2019; Savulescu and Maslen 2015), and could do so, for example, by means of a colour traffic light.

Finally, it should be noted that the human limitation on noticing the framing effect, i.e., how different framings of the same information or question can produce different predictable reactions, would also be an obstacle to moral learning with assistants. Unawareness on the part of subjects of the use of these framings can be a serious impediment to their deliberative independence. Given this, SocrAI could take advantage of this cognitive limitation precisely in order to broaden their critical capacity. To this end, it could be accompanied by nudges that frame the questions and information of the assistant in its interaction with users in such a way as to subtly push them in the direction of an inclination towards doubt and neutral-procedural rigour.

All these nudges, designed to avoid the aforementioned biases, the negative internal and external conditions for deliberation or the inability to notice framings, could in principle constitute important advances for the autonomy of human beings since their ultimate aim is to overcome impediments for the subject to reach, from the fulfilment of formal requirements of deliberative rigour, his or her own moral perspective. But whether they really represent an advance will ultimately depend on the type of strategy on which the nudge is based. Three features will be decisive in order to examine them more particularly in terms of autonomy, as we saw above when referring to manipulation (since the requirement of non-influence with extraneous criteria would be guaranteed by the aims of SocrAI): whether they are T1, T2 or T3, on the one hand; transparent, on the other; and, ultimately, whether they are easily resistible.

As for the type of nudge, most of those suggested so far would be informative (T3) and therefore not undermine autonomy. By providing relevant information about the particular issue or the suitability of the situation for deliberation, they directly aim (without activating heuristics) to gain the agent's attention and reflection, and to make him or her wonder about certain aspects he or she had not previously considered relevant.

One might think these informational nudges are T1, in that they aim to influence the agent's attitude and decision by selecting certain information and excluding other information, thus preventing the user from engaging in complicated reflective processes (Weijers et al. 2020, pp. 8–9), or by using that selected information to affect our emotions or anchors so that a certain situation is viewed so strongly it changes our behaviour (Bovens 2009). However, it is questionable whether this justifies categorising them as T1, as no cognitive shortcut is properly used to direct the decision. Even so, it could not be argued that, irrespective of their typology, autonomy be ultimately infringed. It is true that the nudge in question seeks reflection only in certain situations (in which the relevant information is given) but not in others, since it is designed to ensure that nudgees both ask questions about certain aspects they had not previously considered relevant and that they do not do so when they are not. But rather than threatening autonomy it increases it, for we should not forget that there is a strong relationship between autonomy and time management. Autonomy does not always require choices, as this would prevent the individual from concentrating on what is important. People should be allowed to devote their attention and focus on the issues that, from the point of view of their real interests, are important (Mills 2013; Sunstein 2013). The key, then, is how such reflection is sought, which, to be appropriate, must be limited.

Other proposed nudges, which would take advantage of heuristics in their strategies, but in this case, not to unconsciously direct subjects in a certain direction, but invite them to reflect, deserve a different consideration These would therefore belong to T2, and among them would be those which, using automatic reactions to certain indications, would increase the likelihood that subjects would be aware of, and reflect on, their tendency to deny arguments and obvious data. This could be done on the one hand by the human claim to want to perceive oneself and show oneself to others, as a coherent subject. But also, the predictably positive reaction we tend to have to data from sources to which we attach a high degree of credibility in order, on certain occasions, to cast doubt on the subject's stubborn stance with indications about the authority of the sources and the subject's inconsistent past positions. In such cases, autonomy would not be undermined, because the targeted use of heuristics would ultimately be aimed at the user's reflection.

But in order to determine the ethical acceptability of nudges, by virtue of whether or not they respect autonomy, the essential thing is not to determine what kind of nudges they are, but rather to what extent they are transparent. We said that in a weak sense of transparency, nudges should be designed in such a way that the nudgee could, without much effort, be aware of the meaning of the strategy used. This requirement would be easily fulfilled in nudges that limit themselves to selecting certain information that is considered relevant for the user to make a certain decision (T3), but also when the information given is about the information sources, supposedly reliable for the user or about some previous statements of his or hers that are contradictory (T2) with the current ones. In the case of the latter nudges, the subject will easily interpret the indications as seeking his or her reaction, by virtue of certain personal references, so as not to be biased or incoherent.

Thanks to the (weak) transparency of these T2 and T3 nudges, it could be argued that there is no psychological manipulation, that users can realise with little effort the attempt to influence them to be more reflective. Hansen and Jespersen (2013, p. 24) even go so far as to consider T2 nudges that are also transparent as examples of "empowerment" because they facilitate "freedom of choice" in complex environments. They allow nudgees "to change their actions and behaviour in a predictable way, *while simultaneously leaving them free to choose otherwise—not just as a matter of principle, but also in practice*" (italics are theirs).

A very different case is that of those nudges which, taking advantage of the human limitation on noticing subtle framings in the way of giving information and asking questions, directed users towards less firm positions and more in line with the methodical doubt characteristic of SocrAI. While consistent with the purpose of SocrAI, these nudges would be T1 and, in most cases, would not give nudgees epistemic access to the intentions and means by which they are being influenced. To a certain extent, it is a form of deception, as the intention is to direct the subject's decision with framings that, because they are very elaborate and hidden, would require a high degree of attention and knowledge of the influence of framing in order to be noticed, and are therefore not within the reach of the most common consciousness. We would therefore be dealing with clear cases of decision manipulation.

Such manipulation would become more accentuated because these nudges, in a certain sense, would be irresistible because, although it would be possible to appreciate the strategy and understand its meaning, this occurs once the desired influence has taken place.

A similar paradigmatic case could be the nudge of decreasing the size of painted road signs to get the driver to activate instinctive and automatic braking responses. Conscious perception of and reflection on such a strategy may occur on the part of the nudgee, but only after the fact. The manipulative nature of such nudges is that a behavioural reaction is sought that leaves no real freedom for individuals to choose to do something else if they so decide, because by the time they realise the influence exerted on them, it is simply too late to do anything else. That is why in these cases one can speak of a complete manipulation of automated behaviours and their consequences, not of the decision itself (Hansen and Jespersen 2013, p. 25). Freedom of choice is only in theory, as in practice the nudge effect is unavoidable, leading to instinctive and automatic behaviour. The fact that they are transparent in a weak sense does not make these nudges permissible, for although the nudgee may come to understand the meaning of the nudge, in our case, to realise the framing, this will be a posteriori and therefore does not make it possible to avoid the decision making that is foreseen with the nudge, even if that decision is only part of a mere dialogical interrelation, since for this to take effect it also requires adequate commitments in the time required by each argumentative phase. Anachronistic perceptions that the cause of a certain response was due to a certain premeditated framing would invalidate the achievements of the argumentation, which could have gone in a different direction if the subject had been made aware of the framing at the very moment of its occurrence.

In addition to this lack of "argumentative productivity", the limitation that this type of nudge entails for the nudgee as regards being aware of it in time also poses a serious impediment to his or her ability to effectively dissent from the meaning of the strategy itself. Even if the programming of the framing nudge responds to the supposed formative neutrality to which SocrAI aspires, directing the subject towards doubt and rigour in deliberation, the fact that transparency is not immediate means that the agent loses the option of being able to argue in time that the nudge really responds to this ultimate aspiration of neutrality. In this way, we would ultimately be diminishing the autonomy of the subject, preventing the possibility of dissent, as well as contradicting the very purpose of SocrAI and Socratic nudges to train users in deliberative skills.

## 5 Nudges for motivation

Given its essentially educational character, SocrAI will be a good ethical assistant if, in addition to enhancing users' deliberative skills morally, enabling them to make good choices, it also helps them to translate their choices into action. There are many constraints on even a morally educated person behaving in accordance with his or her qualified values, including those that could be achieved in interaction with an assistant such as SocrAI. The following are some obstacles to moral motivation that, like cognitive ones, could be targeted by nudges designed to overcome them and thus facilitate comprehensive moral enhancement.

First, there would be the difficulty for moral agents to do what they consider right by virtue of certain apathetic tendencies, such as procrastination through inertia or weakness of will.

To neutralise this type of volitional tendency, nudges could be devised to indicate the user's past affirmations or commitments registered by the assistant, and which would clearly show the agent's willingness to act in accordance with those resolutions that now do not motivate him or her. The aim would be to activate the tendency to show ourselves and others as a coherent agent, which we saw before in the level of showing ourselves without inconsistencies between our statements and values, and which also manifests itself, in this case, between our judgements and our actions. These nudges could increase willpower by means of remembering and confirming commitments made in the past by the user himself. Once users have set their own goals with such commitments, these will become salient reference points that motivate them in order to avoid the psychological costs of not achieving them (Clark et al. 2017; Koch and Nafziger 2011), and SocrAI could use the information gathered from its past interactions with users to make them reflect on their inconsistencies in terms of willingness to act.

Motivational nudges could also be set up as a result of giving SocrAI a trustworthy appearance and language. In this case, trustworthiness would not go hand in hand with the claim of affective links between assistant and user. The neutrality intended in the formative attitude that should prevail in SocrAI should exclude nudges that seek to motivate the subject by means of emotions, which could ultimately pervert their open value development. To this end, they should be designed without any discernible human or animal form. Recent studies show that companion robots, manufactured with the appearance of pets or human beings, elicit in users consolidated emotions of attachment to the robots which even lead to attributing some type of mental state or social status to them (Friedman et al. 2003; Melson et al. 2009). Therefore, if a non-provocative design is used, the user would be emotionally distanced from the assistant, facilitating reflective independence.

With that same intention of optimally reducing emotional influences, we should expressly forgo the “affective computing” techniques with which automated systems aim to imitate user emotions and attitudes. Based on the psychological tendency for people of a similar nature to be attracted to each other, companion robots emotionally identical to users are designed with these techniques to gain their trust and thus fulfil their emotional deficits or make them change their unhealthy habits. In our case, interaction based on this emotional affinity could lead to either an excessive dependence of users on the assistant or easier manipulation of them by a malicious designer. In both cases, the results are counterproductive to a virtual assistant that only seeks the development of intellectual abilities, with maximum autonomy.

The reinforcement of trustworthiness in the assistant, therefore in its motivating force, should rather take the form of nudges that would increase the positive evaluation that the user would make of an effortful deliberative process in which the final decision responds entirely to a protagonism that is not overshadowed by the merely "procedural" and neutral, but very effective in data processing, help of the assistant. These nudges would be aimed, therefore, at making individuals perceive the decisions resulting from the dialogue with the virtual assistant as their own, which would make their deliberations much more motivating (Lara 2021, p. 19).

Thirdly, there would be the volitional limitations that have to do not strictly speaking with apathy, but rather with the user's lack of self-confidence to advance in the instructional process involved in their use of SocrAI. Let us remember that this type of assistant is not an oracle or ethical advisor. As mentioned above, it assists the user with constant questioning and feedback, but the choice of substantive ethical criteria is entirely up to the subject. This inquisitive process, together with the responsibility to conclude it personally, contrary to the motivating force this may have, may also overwhelm some individuals and ultimately lead to a lack of self-confidence to achieve the ultimate goal of moral enhancement.

For these constraints, nudges could be designed to detract from the user's negative self-beliefs about his or her capacity for enhancement. They would be based on the automatic reactions of personal satisfaction that could be derived from frequent indications of his or her deliberative achievements. These indications would be the result of processing the assistant's own recordings of the user's evolution in decision-making skills.

On the other hand, social nudges that appeal to comparison with others would also be relevant. These nudges aim to inform about what the majority does and thinks, given that we are gregarious, as well as competitive and our behaviours can therefore be oriented by attracting attention to what others do. This would be the “motivational” version of cognitive representativeness bias. A well-known example of such a nudge is the successful letter that the UK tax office sent to suspected tax avoiders, informing them that most of their fellow citizens had paid their taxes on time. In our case, nudges could be designed that, taking advantage of this tendency to compare ourselves with others, seek a motivational reaction based on the satisfaction derived from the knowledge of the superiority of personal achievements over those of others. One of them could be based on the processing by the assistant of comparative data obtained through digital social networks or access to shared information on the results of similar assistants and, when appropriate (positive), communicating them to the subject. All these informative cues would automatically trigger satisfying feelings that would reinforce self-confidence and, with it, the willingness to follow the instruction despite difficulties.

All these motivational nudges are basically informative, although they could be interpreted as similar to T1, since they seek a predictable, more or less conscious reaction; however, instead of seeking it by taking advantage of common cognitive heuristics they would resort to certain volitional tendencies, such as being coherent or feeling stimulated by personal or comparative achievements. In any case, the motivational nudges would be weakly transparent, in the sense that any "astute" user could realise with little effort that such indications respond to a purpose of the assistant to motivate him or her. This is a perception that could also be immediate.

Despite this, while immediacy makes it possible for the subject to resist the nudge, it is more difficult to establish to what extent these nudges might be easily resistible. This will depend on the psychology of the user, as there will be subjects in whom these tendencies towards consistency or the satisfaction of enhancement or surpassing others are stronger than in others. The ethical or unethical nature of the use of this type of nudge would therefore require a prior study of the user's psychology, which could be modulated according to these characteristics of the subjects.

A separate consideration would require SocrAI's use of VR as an integral part of motivational nudges. As some applications have shown, this technology is capable of strongly influencing subjects, accentuating the feeling of compassion or empathy in virtual scenarios that can then translate into related attitudes of emotional empathy for real behaviours and situations, and has therefore come to be labelled “the empathy machine” (Bollmer 2017, p. 63; Herrera et al. 2018, p. 2). Insofar as empathy could reinforce a motivation towards impartial or altruistic behaviour, one could think of nudges that reinforce moral motivation by sometimes using VR.

It would be difficult not to consider such nudges as manipulative as they would be T1 nudges which, given their direct and strong impact on emotions, would considerably reduce the user's ability to resist their influence.

However, there is also room for a different interpretation of empathy and, consequently, a different possible use of VR as a nudge. The aim would be to develop skills useful for moral deliberation through virtual embodiment in avatars. The use of VR, given certain conditions, should allow users to put themselves in the place of others not so much to feel like them but to “know” their perspective (Rueda and Lara 2020). This would facilitate the adoption of the impartial point of view that characterises morality and with it a greater consideration of the interests of others, and thus strengthen the subject's motivation to behave morally. In this way, the nudge would not seek immediate and irrepressible reactions as a result of the activation of feelings resulting from full identification with the victim of the simulation (seriously limiting freedom of choice). If the aim is for one to just "put oneself in the other's shoes" (without pretending to be the other) and feel as one would if one were in the other's position, the subject's reaction would leave room for reflection derived from adopting a different point of view but without ceasing to be oneself (Lara and Rueda 2021). With this use of VR as a tool for "cognitive" empathy, nudges would be much more resistible, and the subject would still be free to do something other than what the nudge intends.

## 6 Conclusions

AI is with us and here to stay. Among the many contributions it can make is that of helping us improve morally in our judgements and behaviour. Our moral lives are far from being satisfactory, or from always being satisfactory, even for the best of us. We often make decisions in haste, when we are in an inauspicious mood, or we allow ourselves to be swayed by thoughtlessly formulated judgements that lead to biases. Acting in accordance with these biases can have detrimental consequences for us and, in the case of morality, for others.

In this paper we have asked whether it would be acceptable for moral assistants using AI to employ nudges that avoid such biased judgments in the realm of morality. To this end, we distinguished three types of nudges and considered one particular assistant, SocrAI, which seems to us in principle to be more respectful of the moral autonomy of individuals. We have analysed the objections raised to the use of nudges in terms of autonomy and the possibility of manipulation. Our conclusion is that well-designed nudges could be useful and effective tools for our moral enhancement in a way that respects our autonomy and is free from manipulation. With the use of such non-manipulative nudges, a neutral assistant such as SocrAI can help us to formulate better moral judgements and also to overcome some obstacles to behaving according to our best judgements.

**Funding** Funding for open access publishing: Universidad de Granada/CBUA. This study was funded by MCIN/AEI/10.13039/501100011033—FEDER (Grants numbers PID2019-104943RB-I00 and PID2022-137953OB-I00) and FEDER-Junta de Andalucía (Grant number B-HUM-64-UGR20). Funding for open access charge: Universidad de Granada / CBUA.

## Declarations

**Conflict of interest** The authors have no competing interests to declare that are relevant to the content of this article.

## References

Anderson CA, Deuser WE, DeNeve KM (1995) Hot temperatures, hostile affect, hostile cognition, and arousal: tests of a general model of affective aggression. Pers Soc Psychol Bull 21(5):434–448. https://doi.org/10.1177/0146167295215002

Aroyo AM, Kyohei T, Koyama T, Takahashi H, Rea F, Sciutti A, Yoshikawa Y, Ishiguro H, Sandini G (2018) Will people morally crack under the authority of a famous wicked robot? In: RO-MAN 2018—27th IEEE International Symposium on Robot and Human Interactive Communication. https://doi.org/10.1109/ROMAN.2018.8525744

Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C (2009) Cognitive developmental robotics: a survey. IEEE Trans Auton Ment Dev 1(1):12–34. https://doi.org/10.1109/TAMD.2009.2021702

Ashcroft RE (2013) Doing good by stealth: comments on "Salvaging the concept of nudge." J Med Ethics 39(8):494. https://doi.org/10.1136/medethics-2012-101109

Barton A, Grüne-Yanoff T (2015) From libertarian paternalism to nudging—and beyond. Rev Philos Psychol. https://doi.org/10.1007/s13164-015-0268-x

Benartzi S, Lehrer J (2015) The smarter screen: surprising ways to influence and improve online behavior. Portfolio/Penguin

Berkowitz L (1993) Pain and aggression: some findings and implications. Motiv Emot 17(3):277–293. https://doi.org/10.1007/BF00992223

Blumenthal-Barby JS, Burroughs H (2012) Seeking better health care outcomes: the ethics of using the "nudge." Am J Bioeth 12(2):1–10. https://doi.org/10.1080/15265161.2011.634481

Bollmer G (2017) Empathy machines. Media Int Aust 165(1):63–76. https://doi.org/10.1177/1329878X17726794

Borenstein J, Arkin R (2016a) Robotic nudges: the ethics of engineering a more socially just human being. Sci Eng Ethics 22(1):31–46. https://doi.org/10.1007/s11948-015-9636-2

Borenstein J, Arkin RC (2016b) Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. AI Soc 32(4):499–507. https://doi.org/10.1007/S00146-016-0684-1

Bovens L (2009) The ethics of nudge. In: Grüne-Yanoff T, Hanson SO (eds) Preference change. Springer, Netherlands, pp 207–209. https://doi.org/10.1007/978-90-481-2593-7_10

Bruns H, Kantorowicz-Reznichenko E, Klement K, Luistro Jonsson M, Rahali B (2018) Can nudges be transparent and yet effective? J Econ Psychol 65:41–59. https://doi.org/10.1016/j.joep.2018.02.002

Clark D, Gill D, Prowse V, Rush M (2017) Using goals to motivate college students: theory and evidence from field experiments. Natl Bureau Econ Res. https://doi.org/10.3386/W23638

Conly S (2012) Against autonomy: justifying coercive paternalism. Cambridge University Press, Cambridge

Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. Proc Natl Acad Sci USA 108(17):6889–6892. https://doi.org/10.1073/PNAS.1018033108/SUPPL_FILE/PNAS.201018033SI.PDF

de Dreu CKW, Greer LL, Handgraaf MJJ, Shalvi S, van Kleef GA, Baas M, ten Velden FS, van Dijk E, Feith SWW (2010) The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328(5984):1408–1411. https://doi.org/10.1126/SCIENCE.1189047/SUPPL_FILE/DE_DREU_SOM.PDF

Friedman B, Kahn PH, Hagman J (2003) Hardware companions? What online AIBO discussion forums reveal about the human-robotic relationship. CHI 2003

Glod W (2015) How nudges often fail to treat people according to their own preferences. Soc Theory Pract 41(4):599–617. https://doi.org/10.5840/SOCTHEORPRACT201541433

Gray CM, Kubo Y, Battles B, Hoggat J, Toombs AL (2018) The dark (patterns) side of UX design. In: Proceedings of the 2018 CHI conference on human factors in computing systems—CHI´18. ACM Press, pp 1–14

Grüne-Yanoff T (2012) Old wine in new casks: Libertarian paternalism still violates liberal principles. Soc Choice Welfare 38(4):635–645. https://doi.org/10.1007/s00355-011-0636-0

Grüne-Yanoff T, Hertwig R (2016) Nudge versus boost: how coherent are policy and theory? Mind Mach 26(1–2):149–183. https://doi.org/10.1007/s11023-015-9367-9

Hansen PG, Jespersen A (2013) Nudge and the manipulation of choice. a framework for the responsible use of nudge approach to behaviour change in public policy. Eur J Risk Regul 4(1):3–28. https://papers.ssrn.com/abstract=2555337

Hausman DM (2018) Nudging and other ways of steering choices. Intereconomics 53(1):17–20. https://doi.org/10.1007/s10272-018-0713-z

Hausman DM, Welch B (2010) Debate: to nudge or not to nudge. J Polit Philos 18(1):123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

Herrera F, Bailenson J, Weisz E, Ogle E, Zak J (2018) Building long-term empathy: a large-scale comparison of traditional and virtual reality perspective-taking. PLoS ONE 13(10):e0204494. https://doi.org/10.1371/JOURNAL.PONE.0204494

Jesse M, Jannach D (2021) Digital nuding with recommender systems: survey and future directions. Comp Hum Behav Rep 3:100052. https://doi.org/10.1016/j.chbr.2020.100052

Kahneman D (2012) Thinking, fast and slow. Penguin Random House

Klincewicz M (2019) Robotic nudges for moral improvement through stoic practice. Techné Res Philos Technol 23(3):425–455. https://doi.org/10.5840/techne2019122109

Koch AK, Nafziger J (2011) Self-regulation through Goal Setting. Scand J Econ 113(1):212–227. https://doi.org/10.1111/J.1467-9442.2010.01641.X

Lanzing M (2019) "Strongly recommended" revisiting decisional privacy to judge hypernudging in self-tracking technologies. Philos Technol 32:549–568

Lara F (2017) Oxytocin, empathy and human enhancement. Theoria 32(3):367–384. https://doi.org/10.1387/THEORIA.17890

Lara F (2021) Why a virtual assistant for moral enhancement when we could have a socrates? Sci Eng Ethics 27:1–42. https://doi.org/10.1007/s11948-021-00318-5

Lara F, Deckers J (2019) Artificial intelligence as a socratic assistant for moral enhancement. Neuroethics 13:275–287. https://doi.org/10.1007/s12152-019-09401-y

Lara F, Rueda J (2021) Virtual reality not for "being someone" but for "being in someone else's shoes": avoiding misconceptions in empathy enhancement. Front Psychol. https://doi.org/10.3389/FPSYG.2021.741516

Luguri J, Strahilevitz LJ (2019) Shining a light on dark patters. SSRN

MacKay D, Robinson A (2016) The ethics of organ donor registration policies: nudges and respect for autonomy. Am J Bioeth 16(11):3–12. https://doi.org/10.1080/15265161.2016.1222007

Marchiori DR, Adriaanse MA, de Ridder DTD (2017) Unresolved questions in nudging research: putting the psychology back in nudging. Soc Personal Psychol Compass. https://doi.org/10.1111/SPC3.12297

Melson GF, Kahn PH, Beck AM, Friedman B (2009) Robotic pets in human lives: Implications for the human—animal bond and for human relationships with personified technologies. J Soc Issues 65(3):545–567. https://doi.org/10.1111/J.1540-4560.2009.01613.X

Mills C (2013) Why nudges matter: a reply to goodwin. Politics 33(1):28–36. https://doi.org/10.1111/j.1467-9256.2012.01450.x

Mills S (2022) Finding the `nudge´ in hypernudge. Technol Soc 71:1–9

Mirsch T, Lehrer C, Jung R (2017) Digital nudging: altering user behavior in digital environments. In: Leimeister JM, Brenner W (Hrsg.) Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, pp 634–684

Morozovaite V (2021) Two sides of the digital advertising coin: Putting hypernuding into perspective. Market Compet Law Rev 5(2):105–145

Olsen OK, Pallesen S, Eid J (2010) The impact of partial sleep deprivation on moral reasoning in military officers. Sleep 33(8):1086–1090. https://doi.org/10.1093/sleep/33.8.1086

Rawls J (2009) A theory of justice. Harvard University Press

Rueda J, Lara F (2020) Virtual reality and empathy enhancement: ethical aspects. Front Robot AI. https://doi.org/10.3389/frobt.2020.506984

Saghai Y (2013) Salvaging the concept of nudge. J Med Ethics 39(8):487–493. https://doi.org/10.1136/medethics-2012-100727

Savulescu J, Maslen H (2015) Moral enhancement and artificial intelligence: moral AI? In: Romportl J et al (eds) Beyond artificial intelligence. Springer, Berlin, pp. 79–95. https://doi.org/10.1007/978-3-319-09668-1_6

Steffel M, Williams EF, Pogacar R (2016) Ethically deployed defaults: transparency and consumer protection through disclosure and preference articulation. J Mark Res 53(5):865–880. https://doi.org/10.1509/JMR.14.0421

Sunstein C (2013) The storrs lectures: behavioral economics and paternalism. Yale Law J. https://digitalcommons.law.yale.edu/ylj/vol122/iss7/3

Sunstein CR (2015) Why nudge: the politics of libertarian paternalism. Why nudge: the politics of libertarian paternalism. Yale University Press. https://doi.org/10.5860/choice.186901

Susser D, Grimaldi V (2021) Measuring automated influence: between empirical evidence and ethical values. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES´21), May 19–21, 2021, Virtual Event, USA. ACM, New York

Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press. https://doi.org/10.1016/s1477-3880(15)30073-6

Tse WS, Bond AJ (2002) Serotonergic intervention affects both social dominance and affiliative behaviour. Psychopharmacology 161(3):324–330. https://doi.org/10.1007/S00213-002-1049-7

Weijers RJ, de Koning BB, Paas F (2020) Nudging in education: from theory towards guidelines for successful implementation. Eur J Psychol Educ 36:883–902. https://doi.org/10.1007/s10212-020-00495-0

Weinmann M, Schneider C, vom Brocke J (2016) Digital nudging. Bus Inf Syst Eng 58:433–436

White MD (2010) Behavioral law and economics: the assault on consent, will, and dignity. In: Gaus G, Favor C, Lamont J (eds) Essays on philosophy, politics and economics: integration and common research projects. Stanford University Press, pp 203–224

White MD (2013) The manipulation of choice: ethics and libertarian paternalism. Palgrave Macmillan. https://doi.org/10.1057/9781137313577

Wilkinson TM (2013) Nudging and manipulation. Polit Stud 61(2):341–355. https://doi.org/10.1111/j.1467-9248.2012.00974.x

Yeung K (2012) Nudge as fudge. Modern Law Rev 75(1):122–148. https://doi.org/10.1111/j.1468-2230.2012.00893.x

Yeung K (2017) ´Hypernudge´: big data as a mode of regulation by design. Inf Commun Soc 20(1):118–136