ARTICLE

The ethics of self-aware behavioural public policies: any different to standard nudges?

Peter John

School of Politics and Economics, King’s College London, London, UK
Email: peter.john@kcl.ac.uk

(Received 31 January 2023; revised 31 January 2023; accepted 31 January 2023)

Abstract
Nudges – light-touch interventions aimed to help people achieve socially desirable outcomes – can take place without individuals being aware of them. It would seem to be ethically superior to tell individuals that they are being nudged, encouraging them to be aware of the reasons for the official interest in their behaviours. Aided by internal reflection, individuals may make informed choices whether to go along with officially-preferred options or not. In general, this paper adopts this line of argument, justifying self-awareness from the liberal belief in autonomy of the person. However, awareness and/or reflection are not always necessarily ethically superior to passivity, as in cases where manipulation is also present with information provision, when there is framing of deliberative exercises, and where there is harm done to others due to reflectively-driven actions. Most of the time self-awareness is to be preferred, but not always.

Keywords: nudge; ethics; reflection; transparency

Nudges – low-cost prompts that help people get to socially desirable outcomes – have become very common in recent years (Halpern, 2016; Benartzi et al., 2017; Hallsworth & Kirkman, 2020). Many nudges take place without people being aware of them as they appeal to unconscious biases and preferences, which might not be noticed in the designs of letters and e-mails, the default options presented on websites, or in the wording of posters conveying that others are doing an action. But nudges could be made more transparent by telling people about them (Loewenstein et al., 2015; Bruns et al., 2018) and encouraging individuals to think (Banerjee & John, 2021).

So, is being aware of nudges, in particular reflecting more consciously upon them, ethically superior to reacting unconsciously through automatic and psychologically-induced responses? The obvious answer to this question might – at first blush – be thought to be in the affirmative. Human beings, who have independent thoughts, are valued as conscious actors and are not necessarily subject to naturally-given or manipulated reactions: they have agency and a degree of autonomy, given the right conditions. When considering a choice that might affect their and society’s welfare,
such as whether to take out a new pension or to change to a healthier diet, it would seem to be worthwhile that individuals first reflect on the decision in terms of their objectives and values, then change (or not) their behaviours consciously, rather than being guided by other human beings to the same outcome without being self-aware. This is what Sen calls process freedom: ‘processes that allow freedom of actions and decisions’ (Sen, 2014: 527). Fully conscious humans can make decisions based on thinking about their lives, considering the welfare of others and taking into account the interests of the wider society. Nudges that are received wholly automatically cannot achieve these aims. When individuals who are nudged realise what has happened to them, and agree or maybe even disagree with the policy, this process comports with human-centred values of autonomy, self-ownership and agency.

But there are some powerful arguments against this position, which rest on the objective of behavioural public policy (BPP) to improve individual and collective welfare. If individuals choose options that might harm themselves and/or impose harm on others, would it not be ethically better to help them choose a course of action (or inaction) that leads them to achieve more desirable outcomes? This is the classic liberal position. According to this argument, welfare considerations can beat process freedom in certain circumstances. Moreover, in terms of respecting choice, it may be possible to inform the person after the intervention (Thaler & Sunstein, 2009: 244) rather than before or during it. In the original libertarian paternalist framework, both welfare and autonomy are promoted by nudges in a way that could not be done if people are more self-aware. People might thank the paternalist for keeping them on track in ways that could have never been achieved if left to their own devices. Partly because nudges are thought to work ‘better in the dark’ (Bovens, 2009), there are good reasons to stick to the notion that socially-beneficial outcomes are best achieved through automatic processes. Self-awareness is not a necessary condition for successful and legitimate nudges, and can be introduced before or after their delivery, part of the democratic processes that authorise BPPs just like any other public policy.

Being self-aware may give the illusion of autonomy when people are still being constrained by behavioural interventions. Being conscious is not a fool-proof guard against manipulation, and the opportunity for reflection might lull someone into thinking they have autonomy when in fact they are being manipulated, an example of what the Marxist social theorist Herbert Marcuse called ‘repressive tolerance’ (Wolff et al., 1969). People are reassured they have choice when actually they have accepted the framing of the intervention, even when encouraged to think. Just because people reason does not mean they are not influenced by emotions and their biases, which are still at work in deliberative settings (Ryfe, 2005). In addition, doing things automatically is not inconsistent with the idea that people might feel free when making these choices and be content for their automatic side to be in play for matters that are not controversial. As nudges are usually not contentious, it may not be worthwhile for individuals to expend their cognitive energy to think about these kinds of issues very much. In any case, even when using their automatic processes, people are usually ready to zoom in on a problem when it becomes more acute or there is some danger in what they are being asked to do. Agency can be at work in a nudge environment, as people might reason ecologically, in a more incremental and ‘frugal’ way (Schmidt, 2019), rather than as rational actors as in traditional
economic theory. Also, what if, as Peter Singer argues (Singer, 2000: 228), consciousness were associated or causally linked to a morally bad decision, how would that compare on ethical grounds with individuals who are using their habits and instincts to make beneficial choices? For the same situation, something that is planned consciously can be worse than something not, if it encourages people to put their self-interest first. For example, if someone falls into the sea and someone else jumps in spontaneously to help them, the outcome might be better than when the person hesitates on the shoreline thinking of all the pros and cons while the helpless victim slips further under the waves. Overall, there might be an ethical advantage for nudge processes to be automatic: less time for overthinking and strategic cunning; more space for genuine and authentic responses.

To adjudicate these arguments, this paper proceeds in several steps: first, it sets out the ethics of standard nudges and why they are deemed superior to other kinds of public policies. Then the second section explains how more self-aware BPPs – nudge pluses – work and why they might be better in ethical terms than standard nudges. This section reviews an additional argument in favour of self-aware policies: they can take place without detriment to the effectiveness of nudges. In the final section, the similarities between nudge and nudge+ in both implementation and ethical senses is stressed, while not giving up on the claim that self-aware nudges are indeed ethically superior in most circumstances.

The ethics of nudge

The core idea of nudge is that policy-makers or choice architects use their discretion over the light-touch and non-coercive tools of government to encourage individuals to make choices – courses of action or inaction – that maximise their and societies’ welfare. Examples of nudges are letters that inform taxpayers that nine of ten people had already paid (Hallsworth, 2014), prompted choices to donate organs on a driver’s licence renewal website (Johnson & Goldstein, 2003), or defaults for new employees to join a pension scheme (Thaler & Benartzi, 2004). Nudges take advantage of automatic and learnt habits or biases that help guide individuals to welfare-enhancing outcomes without them being directly aware of the interventions. Central to nudge is that these interventions are not directive and that individuals can, if they become aware of the choice between one kind of action or another, change their minds or just not take the preferred choice. They are not compelled by the nudge itself to act. They usually can opt out of it without prejudice or disadvantage, though they might sometimes feel bad for doing so, for example when they recieve injunctive norms (messages invoking social disapproval), which can sound pejorative. The classic nudge is the default that can only be reversed if they choose to do so. Defaults take advantage of status quo bias so someone might get to this choice without much conscious thought. It takes a bit of effort to reject the nudge. The good policy-maker explains what is going on, but usually there is not much time to do this when the nudge is being delivered. So, even in this welfare-orientated approach, there is a concession that transparency and by implication some human agency is an important aspect in explaining how nudges work, even if in practice it is hard to achieve in their delivery. This argument points to the superiority of more self-directed human choices.

For this reason, the paternalism of nudges is thought to be consistent with respecting human freedom, conveyed by the term ‘libertarian paternalism’ (Sunstein & Thaler, 2003; Thaler & Sunstein, 2003), and is part of the ethical project of nudge. This term has come in for a lot of criticism on the grounds that if the influence of nudges on human behaviour is known about in advance and the objective is to increase human welfare, then this is as paternalistic as any law or other form of government control (Sugden, 2009; Anderson, 2010; Ryan, 2017). The libertarian element is not really taking place. Despite the protest of its inventors (Sunstein & Thaler, 2003), the concept could be thought to be an oxymoron.

The key argument that libertarian paternalists use to defend the term is that human choice is anyway constrained by many prior government decisions. What governments face day-to-day is how to enhance choices within this framework, as there is no original state of nature of freedom to depart from. Moreover, when applying more careful definition of liberty-enhancing outcomes, Dowding and Oprea (2022) argue that there is no systematic difference between nudges and many other policies on freedom grounds. Nudges and other policies can be freedom limiting or enhancing depending on their design. This fits with the idea that most tools of government have an information aspect in terms of communication and persuasion, increasingly so with the use of behavioural techniques (John, 2013). This gives all tools of government a communication dimension, so that individuals are invited to act as well as being compelled. Most laws and regulations cannot be universally enforced, especially if most people do not readily comply. Such an idea is core to the responsive regulation research programme (Ayres & Braithwaite, 1992; Braithwaite, 2002; Baldwin & Black, 2008), whereby it is recognised that top-down laws are likely to fail through non-compliance but that a communicative and participatory approach to policy implementation is more likely to work. This soft approach to regulation shares a few features of nudge in seeking softer approaches to behaviour changes, but without nudge being communicative or participatory. The comparison leads naturally to more deliberative nudges and thinks discussed in the next section.

In spite of the libertarian side to nudge, automatic processes that do not appear to be coercive could in fact be manipulating citizens rather than being straight with them with clear policy choices (see review by Schmidt & Engelen (2020)). This is particularly important because policy-makers are making assumptions about what is in people’s interest as a way to justify the decision of overriding their immediate preferences and for using their instincts to adhere to the goals that the policy-maker has chosen for the individual. But it is not clear exactly what is in someone’s interest as interest is complex and may vary from situation to situation (White, 2016). White gives the example of chocolate cake: it might be bad for health, but in the context of friendship and families the giving of cake may be in someone’s interest. Its consumption could satisfy peer expectations and altruistic desires (which could have a health benefit). Now of course on average the interest of the individual is likely to be in having less cake, but in a particular situation policy-makers may be using their skills at the presentation of information to act against the interests of the individual. That is manipulation.
The use of awareness and reflection could limit manipulation. In the case of the chocolate cake, it might be useful to convey the idea that it is bad for you in the long term if you ate it too much, and for the individual to be aware of what the policy-maker is doing; but also allow the individual to say, ‘I understand you, but in this case, I will eat the cake with my friend, but I will consider cutting back in future when is it not a special occasion’. It might be seen to be manipulative, a more subtle way of doing of getting government policy across, particularly as there is often an unclear boundary line between manipulation and non-manipulation (Dowding & Oprea, 2021), but it less likely. Policy-makers and citizens need to have open conversations (if that is possible) about these matters. If citizens get a voice, governments are more likely to act correctly and change their policies if they get things wrong. This is responsive government, a crucial feature of democracy (Powell, 2000; Besley & Burgess, 2002).

By being open to the charge of manipulation, nudge can be criticised by libertarians and others (for example, Sugden, 2018). Should policy-makers prioritise interventions that allow people to make mistakes and learn rather than subtly guiding them, which implies a degree of autonomy and self-awareness, with the state at some remove from the citizen to encourage that flourishing. The question is whether there is a half-way house between welfare and autonomy-protecting policies, another version of paternalism, but this time saved by self-awareness. This new attempt to reconcile freedom and welfare is core to the nudge+ alternative discussed in the next section.

The ethics of nudge+

The nudge+ idea comes from earlier work comparing nudges and thinks (John et al., 2009; John, 2011), whereby standard nudges were compared with interventions that have citizen participation embedded in them, such as citizen forms, juries and other mini-publics. The conclusion of this research programme was that think as a policy tool has several limitations, often having very small effects on a self-selected samples and causing occasional backfires. Thinks have advantages in ethical and democratic terms, but are hard to scale up, partly because it is challenging to select representative samples into them, except perhaps deliberative polls (Fishkin, 2011). The question that emerges is whether it is possible to take advantage of the think dimension in a way that is closely linked to the delivery of nudge. Hence nudge+ (John & Stoker, 2019; Banerjee & John, 2021, 2022). This is not a deliberative nudge as in Sunstein’s formulation (Sunstein, 2016b), whereby nudges are delivered incorporating deliberation, but when a deliberation is added on to an existing nudge, usually taking place at the same time as the nudge is being delivered or very soon after or beforehand. The nudge and think are happening more or less at the same time, but still remain distinct interventions. This dual approach takes advantage of what are called system 1 (automatic, fast) and system 2 (reflective, slow) cognitive processes (Kahneman, 2011), with both in play (and relate to each other) in the delivery of the nudge and the plus. People react to the nudge with system 1, but reflect on it with system 2.

The element of reflection gives agents some autonomy when they receive the nudge. They can then choose whether to accept it or not. The plus is there to promote autonomy over the nudge which is consistent with the libertarian paternalist
framework but gives the agent more choice over what is happening to them than classic nudges. As nudge+ promotes agency, then it is ethically superior to nudge on its own. From a paternalist perspective, it is commonly thought there are no negative impacts on making the agent aware of these choices (Loewenstein et al., 2015; Bruns et al., 2018). They make the same welfare-enhancing choices whether they are being told they are being nudged or not; moreover studies of transparent nudges show that there is a positive impact on agents’ sense of autonomy (Wachner et al., 2020). Overall, the public is more supportive of deliberative nudges than standard ones (Sunstein, 2016b), though some studies find no difference (John et al., 2022). One recent study, however, does show a lack of beneficial effects of being told about the nudge (Michaelsen et al., 2021), which appears to show the limitation of reflection in some situations. It is an ongoing research agenda, but one that in general supports transparency and self-awareness.

**Maybe not so different ethically after all?**

It is tempting to see nudge and think as operating very differently, based on their different conceptions of the roles of the citizen and taking unique ethical stances as a result (John et al., 2009). The move to nudge+ brings these two worlds closer together, less as competitors, more as complements, each needing the other. If they are closer in how they are used, as they work together, then they are closer together ethically. This line becomes more plausible when thinking about the cognitive science behind nudge and nudge+ (Banerjee & John, 2021) as there are close links between system 1 and system 2 processes and the two parts of the brain may be regarded as part of the same cognitive system. People are often partially aware of the nudges against them, and likewise charges of manipulation and framing can be present even in a deliberative setting. With nudges it might be more obvious what the government is doing once it is worked out, whereas the goals behind a think or a plus might be rather diffuse and hidden because the decision has been handed over to the individual. It is a bit like the patient listening to their doctor who says there are three choices about what to do about their health condition and it is totally up to the individual. It is a bit like the patient listening to their doctor who says there are three choices about what to do about their health condition and it is totally up to the patient to decide which one to choose. But it is usually clear the doctor is leading the patient to choose one option from the way these choices are framed. Nudge+ should state what the objectives of the policy-maker are and then ask the person to reflect freely upon them, but it takes an extra effort to ensure this happens and there is always the risk that individuals defer to policy-makers’ preferences, even if they are hidden. People can be less aware of manipulation if they are given a more open choice than when they are being nudged what to do.

Then there is the question as to whether reflection is always necessary. People often keep a look out about what is being done to them and do not necessarily want to stop and think about every cue, a bit like drivers being aware of road signs and other cars, but ready to take action when something looks wrong, using their intuitions to get back to their fully conscious selves. Moreover, nudges vary a lot in their level of transparency and respect for autonomy, and can be designed with ethical principles weaved into them (Lades & Delaney, 2019), and of course openness and transparency are part of this new turn in BPP. But even with the similarities between nudge and nudge+, it is still preferable ethically in most situations to give people...
some opportunity for reflection and self-awareness in the delivery of nudges in an active deliberate way. This is what nudge+ aims to do.

Acknowledgements. I would like to thank the participants of the second ‘Behavioural Limits to the State’ seminar to whom I presented my paper on 6 May 2022. I also thank my colleague Paul Lewis for his comments on an earlier draft.

References


https://doi.org/10.1017/bpp.2023.10 Published online by Cambridge University Press