Nudging, Transparency, and Watchfulness

Ivanković, Viktor; Engelen, Bart

Nudging, Transparency, and Watchfulness

in *Social Theory and Practice* 45 (1)

https://doi.org/10.5840/soctheorpract20191751

Viktor Ivanković, Central European University, Doctoral School of Political Science, Public Policy and International Relations, Nador u. 9, 1051 Budapest, Hungary, ivankovic_viktor@phd.ceu.edu

Bart Engelen, Tilburg School of Humanities, Department of Philosophy, Warandelaan 2, 5037 AB Tilburg, The Netherlands, b.engelen@uvt.nl

1. Introduction

In recent years, behavioral techniques have been shown to benefit individual welfare and alleviate collective action problems in a range of areas. Nudges have been implemented in situations where people make systematic errors in reasoning and experience difficulties in realizing whatever they (or policy makers) find valuable. As aspects of the choice architecture that predictably, yet non-coercively, alter people's behavior, nudges can take on a multitude of forms, some of which "typically work better in the dark."[1] Famous nudges like the cafeteria food arrangement may be found less effective if telegraphed to their intended targets. Opponents of nudges point out the concern that because nudges rely on psychological quirks, they "will be more effective if they are not transparent to the individuals subjected to them."[2]

[1] Luc Bovens, "The Ethics of Nudge," in *Preference Change: Approaches from Philosophy, Economics and Psychology*, ed. Till Grüne-Yanoff and Sven Ove Hansson (Berlin & New York: Springer, 2009), pp. 207–219, 209.
[2] Till Grüne-Yanoff, "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles," *Social Choice and Welfare* 38 (2012): 635–45, 637.

The worry here does not solely concern nudge effectiveness. This would be easily answered by findings which suggest that nudges can work well even if people are fully aware of them. Transparency is primarily a concern about the purported (il)legitimacy of government policies as it is intertwined with fundamental notions of accountability, respect, deliberation and consent.[3] Governments that nudge people 'behind their backs' without the latter being (able to become) aware of the influence are considered to be overstepping serious boundaries. Using such policy tools does not allow for the kind of scrutiny and contestation that should be possible in liberal democracies.

In this paper, we develop an account of nudge transparency, which enables us to analyze the force of nudge critics' concerns and find ways of meeting them. The aim of the paper is both conceptual (to define and analyze what transparency in nudging entails) and normative (to assess how the transparency of a nudge affects its permissibility). In section 2, we illustrate the worries critics have about the non-transparency of nudges and show why the standard responses fail. In section 3, we address the empirical question concerning how transparency relates to the effectiveness of nudges. In section 4, we build on insights from Luc Bovens and Andreas Schmidt to provide a detailed account of what exactly transparency means and make some conceptual distinctions that enable us to locate the normative concerns more clearly. In section 5, we reduce the scope of our moral inquiry into the legitimacy and permissibility of different kinds of nudges by setting uncontroversial cases aside. In section 6, we analyze the pivotal notion of *watchfulness* and develop a conception that is both feasible and acceptable in democracies like

[3] Pelle G. Hansen and Andreas M. Jespersen, "Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy," *European Journal of Risk Regulation* 4 (2013): 3–28, 15.

our own. In section 7, we investigate its normative implications before rebutting some objections in section 8. Section 9 concludes.

2. The worry, the responses, and why they do not convince

As mentioned, one of the main worries with governments using nudge techniques is that they play into less-than-rational psychological mechanisms and thus influence people's behavior 'behind their backs'.[4] By using them, governments are covertly steering citizen behavior without this being obvious to the citizens themselves, which would make nudges impermissible at least prima facie. Some critics argue that nudges are therefore illegitimate policy tools because they violate people's autonomy. It is the quality of being covert that is said to make nudges more problematic than their coercive but more forthright counterparts, such as mandates and sanctions.[5] Riccardo Rebonato argues that nudges have a greater accountability deficit, "exactly because the means employed … are not transparent."[6]

In response to these worries, Richard Thaler and Cass Sunstein, the main proponents of nudge policies, develop two counterarguments. First, they stress that resisting nudges should be "easy and cheap."[7] If the nudgee perceives a nudge as not in line with his long-standing goals and values, he can always go against its influence. Why criticize nudges if they are liberty-preserving and can thus be easily resisted? Opponents, however, are not convinced. In their view,

[4] Daniel M. Hausman and Brynn Welch, "Debate: To Nudge or Not to Nudge," *Journal of Political Philosophy* 18 (2010): 123–36, 135; Bovens, "The Ethics of Nudge," 216; Riccardo Rebonato, "A Critical Assessment of Libertarian Paternalism," *Journal of Consumer Policy* 37 (2014): 357–96; Jeremy Waldron, "It's All for Your Own Good," *The New York Review of Books*: October 9, 2014.
[5] For example, see Richard E. Ashcroft, "Personal Financial Incentives in Health Promotion: Where Do They Fit in an Ethic of Autonomy?," *Health Expectations* 14 (2011): 191–200.
[6] Rebonato, "A Critical Assessment of Libertarian Paternalism," 360.
[7] Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New Haven, CT: Yale University Press, 2008), 6.

resistibility depends at least in part on transparency. If non-transparent nudges steer choices "while flying below the radar screen of rational deliberations,"[8] people may not realize they are being influenced and may make decisions they do not want to make. A lack of transparency makes it harder to sidestep the nudge, thereby inhibiting people's capacities to autonomously pursue their conceptions of the good life. Some nudges are controversial because they constrain reflective decision-making and thereby "prevent the reflective act of will required for a decision-maker to avoid the nudge consciously and willfully."[9]

Second, Thaler and Sunstein claim that governments wanting to employ nudges should do so publicly and transparently, with officials being "happy to reveal both their methods and their motives."[10] According to Sunstein, nudges "should be visible, scrutinized and monitored"[11] in order to reduce the likelihood of having illicit ends, reducing welfare or violating people's autonomy or dignity.[12] According to Chris Mills, nudges that are in line with people's ends, avoidable, made public and transparent, need not violate autonomy.[13] Some critics agree that nudges can be legitimate as long as they are implemented transparently: "the more clearly visible, the better."[14] However, other critics remain unconvinced and argue that "the espousal of transparency and publicity constraints comes across as an artificial and ad hoc declaration of values that belies a lack of real interest in the importance of ensuring that those subjected to

[8] Rebonato, "A Critical Assessment of Libertarian Paternalism," 366.
[9] Chris Mills, "The Choice Architect's Trilemma," *Res Publica* 24 (2018): 395–414, 407.
[10] Thaler and Sunstein, *Nudge*, 245.
[11] Cass R. Sunstein, *Why Nudge? The Politics of Libertarian Paternalism* (New Haven, CT: Yale University Press, 2015b), 148.
[12] Cass R. Sunstein, *The Ethics of Influence: Government in the Age of Behavioral Science* (Cambridge: Cambridge University Press, 2016), 42.
[13] Chris Mills, "The Heteronomy of Choice Architecture," *Review of Philosophy and Psychology* 6 (2015): 495–509, 502.
[14] Rebonato, "A Critical Assessment of Libertarian Paternalism," 392.

these subtle forms of state power understand the underlying rationale."[15] Nudge techniques exploit cognitive and motivational heuristics–instead of explicitly mandating and sanctioning individuals–and thus seem extremely convenient to governments not wanting to face full scrutiny. Disclosing information about nudges, these critics argue, does not address the problem that nudges fail, by their very nature, to respect people as autonomous and rational persons.[16] In other words, "we cannot be confident that publicity and transparency in combination would take away a government's motive and opportunity to manipulate."[17] And even if we could ensure that governments nudge citizens with only the best of intentions, covert nudge techniques still seem objectionable for not living up to democratic standards of accountability, deliberation, and contestation.

In our view, there are three kinds of reasons why critics tend to remain unconvinced by Thaler and Sunstein's standard response that nudges are permissible when disclosed. First, as we will see in the next section, critics like Till Grüne-Yanoff believe that, when disclosed, nudges will no longer be effective; while disclosure does not render them impermissible, it takes away their very purpose.[18] Second, if nudges are manipulative, critics argue, disclosing them does nothing to right this wrong, because transparent manipulation can be wrong as well.[19] Third, Thaler and Sunstein's response is underdeveloped; they never offer a proper account of the kind of transparency required, what exactly it consists in and which nudges can be called transparent and which cannot.[20] In this paper, we address all three of these issues. We show that most nudges are

[15] Joel Anderson, "Review of Nudge: Improving Decisions about Health, Wealth, and Happiness," *Economics and Philosophy* 26 (2010): 369–76, 374.
[16] Waldron, "It's All for Your Own Good."
[17] T. M. Wilkinson, "Nudging and Manipulation," *Political Studies* 61 (2013): 341–55, 344.
[18] Grüne-Yanoff, "Old Wine in New Casks."
[19] Andrés Moles, "Nudging for Liberals," *Social Theory and Practice* 41 (2015): 644–67, 655; Wilkinson, "Nudging and Manipulation."
[20] Bovens, "The Ethics of Nudge."

permissible as long as they are transparent and can be resisted as a result. For this, we will need to provide a proper account of transparency, thus addressing the third issue and rising above the shortcomings of Thaler and Sunstein's account.

Before turning to these tasks, we address the first worry in the following section and show that, under the right conditions, transparency does not render at least certain nudges ineffective.

3. Transparency and the effectiveness of nudges

What exactly do we know about the effects of making nudges transparent? The most common expectation is that this impact will be negative. Bovens, for example, hypothesizes that the more specific nudgers are when disclosing information about nudges, "the less effective these techniques are."[21] Grüne-Yanoff argues that this will likely be the case, either because people will come to realize and resent that they are being steered and look to resist this, or will come to see through the bag of tricks that the behavioral techniques are using, thus neutralizing their influence.[22] If I come to understand how the scary pictures on cigarette packages work, "I will no longer find the drastic slogans and images shocking. Thus the effectiveness of the policies requires their being not fully transparent."[23] If this turns out to be true, nudge enthusiasts face a conundrum: either they should go for nudges that are transparent but ineffective, or effective but non-transparent.

Scarce empirical evidence, however, suggests that transparency does not necessarily decrease effectiveness. A study by Loewenstein et al. shows that informing people about the use of defaults in steering decisions about advanced directives does not significantly weaken their

[21] Ibid., 217.
[22] Grüne-Yanoff, "Old Wine in New Casks," 637–638.
[23] Ibid., 638.

impact, suggesting that "such defaults can be transparently implemented, addressing the concerns of many ethicists without losing defaults' effectiveness."[24] This finding can be welcomed by nudge enthusiasts as evidence that nudges can be both transparent and effective. On the other hand, it seems to contradict the idea that transparency increases resistibility against covert influences. The latter conclusion, however, is not as straightforward as it seems. Sunstein rightly points out that individuals might react to the presence of defaults differently if the disclosure helped them to fully appreciate the impact of defaults on behavior.[25] Furthermore, it is yet to be determined whether transparency changes the effectiveness of other kinds of nudges. That this will likely be the case at least for certain nudges is suggested by an experiment in which On Amir and Dan Ariely show that attempts at exploiting reflexive reasoning can be overcome even when individuals are merely instructed to carefully reflect on their decisions.[26]

The empirical evidence still allows for multiple possible scenarios owing to the impact of transparency. It might: reduce the effectiveness of nudges (because it induces reflectiveness or because people do not endorse the underlying goals), make nudges counterproductive (because people show reactance when they do not like being nudged), make nudges even more effective (because people would understand and support the underlying goals) or have no real impact on effectiveness at all.[27]

One venue for further empirical research is whether the effectiveness depends on the trust of the nudgee in the nudger. When the nudgee trusts the nudger, one can intuitively expect him to agree

[24] George Loewenstein, Cindy Bryce, David Hagmann, and Sachin Rajpal, "Warning: You Are About to be Nudged," *Behavioral Science and Policy* 1 (2015): 35–42, 40.
[25] Cass R. Sunstein, "Do People Like Nudges?," *Administrative Law Review* 68 (2016): 177–232. Sunstein extensively discusses the findings of Loewenstein et al. and stresses that quite a few things could be going on here, such as people not focusing on the information or not caring about it. See also Sunstein, *The Ethics of Influence*, 154–57.
[26] On Amir and Dan Ariely, "Decisions by Rules: The Case of Unwillingness to Pay for Beneficial Delays," *Journal of Marketing Research* 44 (2007): 142–52, 146–48.
[27] Sunstein, *The Ethics of Influence*, 154.

with the direction of the nudge and see the nudge as supportive of his own goals. While we know that these aspects are crucial in generating public support,[28] it makes intuitive sense to claim that transparency would increase the perceived legitimacy and thus also the effectiveness of the nudge. If you want to eat healthily and you are invited to a party at your friend's place, then you will welcome him transparently influencing you by placing the potato chips in a distant corner.[29] If these conditions are met, it seems people might not oppose the use of nudges. Also, highlighting the effectiveness of a nudge can increase its acceptability, but there is no evidence that this is the case when one is transparent about the underlying processes (nudges tapping into less-than-conscious heuristics).[30]

More problematic, however, are situations in which making nudges transparent helps nudgees realize that they are being manipulated by someone whom they do not trust, leading them in a direction they do not endorse. Since people perceive nudges in these cases as illegitimate, transparency about them can lead to "reactance."[31] According to Robert Baldwin, disclosing nudges is "as likely to provoke protest as to reassure potentially targeted citizens."[32] One study provides evidence for such reactance amongst people who oppose the use of (default) nudges by government.[33] This reveals a connection between the empirical issue about the impact of

[28] Cass R. Sunstein, *Human Agency and Behavioral Economics: Nudging Fast and Slow* (Cham: Palgrave MacMillan, 2017), 38.
[29] Lucia Reisch and Cass Sunstein show that people generally welcome nudges that are targeted at policy goals they support and that are implemented by governments they trust. This evidence in itself, however, does not say anything about the impact of transparency on effectiveness. See Lucia A. Reisch, Cass R. Sunstein, "Do Europeans Like Nudges?,", *Judgment and Decision Making*, 11 (2016): 310–25.
[30] Dragos C. Petrescu, Gareth J. Hollands, Dominique-Laurent Couturier, Yin-Lam Ng, and Theresa M. Marteau, "Public Acceptability in the UK and USA of Nudging to Reduce Obesity: The Example of Reducing Sugar-Sweetened Beverages Consumption," *PLoS ONE* 11 (2016). Available at: https://doi.org/10.1371/journal.pone.0155995].
[31] Sunstein, *The Ethics of Influence*, 119.
[32] Robert Baldwin, "From Regulation to Behaviour Change: Giving Nudge the Third Degree," *The Modern Law Review* 77 (2014): 831–57, 854.
[33] Ayala Arad and Ariel Rubinstein, "The People's Perspective on Libertarian-Paternalistic Policies," accessed online on September 18 2018, at https://www.tau.ac.il/~aradayal/LP.pdf.

transparency on nudge effectiveness and the normative issue regarding the degree of nudge transparency required.

Transparency could, however, render many nudges ineffective in one sense. If people agree with or take up as reasonable the behavior that a transparent nudge helps them pursue, this means that the intervention often no longer operates 'qua nudge' in the narrow heuristics-triggering sense. Its effectiveness no longer comes from the behavioral technique it employs (which taps into the heuristics of System 1) but from the information it conveys (which speaks to the reflective capacities of System 2).[34] Take a person whose considerations of fairness result in his decision only to pay his taxes if a vast majority of citizens do so as well. When he gets a letter from the tax authorities stating that 90% of citizens have already filed their taxes,[35] his System 2 processes the information and updates his beliefs, unlike many others whose System 1 heuristic is triggered to establish conformity with others. In both scenarios, the modified choice architecture is effective in steering people's behavior, but transparency may have different effects. Realizing how effective a nudge can be and what it aims for may lead to reactance amongst people who come to see conformism as an irrelevant factor (I *shouldn't* pay my taxes merely because others

[34] Understanding the processes and capacities of the human mind in terms of System 1, with less-reflective processes like heuristics and biases, and System 2, with more reflective processes, comes from dual-process theory (Daniel Kahneman, *Thinking, Fast and Slow* [London: Penguin, 2011])**.** This paper uses the vocabulary of dual-process theory (System 1 and 2, shallow and reflective processing) when analyzing how nudges work. While this mainstream view is contested in behavioral science, our largely conceptual analysis does not take a stance on its descriptive accuracy. Our claim here assumes that only heuristics-triggering interventions that steer behavior via System 1 work 'qua nudges.' This narrow view clashes with the broader view of nudges that Sunstein uses: "some nudges enlist or exploit System 1 whereas other nudges appeal to System 2" **(***The Ethics of Influence*, 34**)**. However, it is in line with the general idea that it is their reliance on less-than-rational processes that distinguishes nudges from (the rational processes involved in) more traditional strategies like informing and persuading people. For more information on the merits and drawbacks of dual-process theory, see *Dual-Process Theories in Moral Psychology: Interdisciplinary Approaches to Theoretical, Empirical and Practical Considerations*, ed. Cordula Brand (Weisberg: Springer, 2016).

[35] Michael Hallsworth, John A. List, Robert D. Metcalfe, and Ivo Vlaev, "The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance," *Journal of Public Economics* 148 (2017): 14–31.

do) while other people's conviction might be strengthened (it is only fair that I pay my taxes if others do as well).

As we will argue in this paper, making nudges transparent meets potential concerns about autonomy and agency. On the one hand, when disclosure renders nudges less effective, it probably means that nudgees come to see the nudge as contrary to their reasons and goals, in which case their decreased effectiveness is, ceteris paribus, a good thing. On the other hand, when disclosure does not make nudges less effective, there are two possible explanations. Either the nudgees perceive the nudge(r)s as pursuing legitimate goals (in which case we fail to see the problem), or the particular nudge may not be easily resistible even when made transparent (which does raise worries).

In this paper, our main purpose is to analyze the conditions under which less-than-fully-transparent or non-transparent nudges that only work qua heuristic triggers can be justified, keeping in mind the possible transparency effects just stated. We argue for a technically elaborated account of transparency that does not require full disclosure of such nudges in the most standard sense. As we will show, there are reasons for nudges to remain undisclosed and work as heuristic triggers.

4. Transparency: a typology

Arguing that governments should be transparent about (the reasons behind) their nudging techniques leaves open what kind of transparency is desirable.[36] Bovens's seminal paper on "the

[36] Sunstein, *The Ethics of Influence*, 73.

ethics of nudge"[37] sets up the conceptual landscape for discussing issues surrounding nudge transparency. Bovens's crucial distinction is between *type* and *token interference transparency*. *Type interference transparency* stands for governments informing citizens that certain techniques will be used to increase their individual welfare or solve collective action problems. These governments are thus transparent about the types of interventions they are going to implement. According to Bovens, however, this "is not enough."[38] Subliminal messaging, he argues, becomes no more permissible if it is openly admitted, and there seems to be little or no difference between type interference transparency of non-transparent nudges and disclosing the use of subliminal messages. Therefore, type interference transparency is not demanding enough. This conclusion is shared by Baldwin: "general strategies of disclosure will be seen as doing little to change the semi-covert nature" of nudges that target people's less-than-conscious cognitive processes.[39]

Conversely, *token interference transparency* requires transparency about every particular nudge intervention. A number of problems arise here. First, the effects of nudges that owe their effectiveness to their covertness could be dulled, making token interference transparency a non-starter.[40] Second, some nudges may be like subliminal messages in that they resist token interference transparency: "transparency about the particular instances of such policy applications … is often not achievable."[41] Third, even if feasible, it seems absurd to require each and every nudge to be accompanied by a notification that a nudge is in place. Given the fact that choice architecture is often inevitable, token interference transparency seems overly demanding:

[37] Bovens, "The Ethics of Nudge."
[38] Ibid., 216.
[39] Baldwin, "From Regulation to Behaviour Change," 854.
[40] Bovens, "The Ethics of Nudge," 217.
[41] Grüne-Yanoff, "Old Wine in New Casks," 638.

"beyond a certain level of effort, it will become absurd."[42] Lastly, we do not require such transparency for other kinds of policies, such as laws or financial incentives. At any given moment, there is a plethora of regulations and stipulations in the background which are not being disclosed every single time they may become relevant.[43] Token transparency, therefore, is too demanding.

What would suffice, claims Bovens, is an *in principle token interference transparency*, which entails that "a *watchful* person would be able to identify the intention of the choice architecture and she could blow the whistle if she judges that the government is overstepping its mandate."[44] Bovens stresses people's (not necessarily fully realized) ability to notice a particular nudge when subjected to it. Reformulating this, one can say that an *in principle* transparent (or 'detectable') nudge is not always *de facto* transparent (or 'detected').[45] Perhaps it works best when not detected, and thus with people who are not watchful, but it is not impossible to become aware of it and defy its influence. This is what distinguishes an in principle transparent nudge from subliminal messages, which are undetectable even for the watchful.

A recent contribution to the transparency typology, following Bovens, is made by Schmidt, who offers the principle of *reasonable token inference transparency*, according to which "a watchful person is someone who, with not unreasonable effort and understanding, would be able to detect a token of a nudge and would comprehend the intention behind it."[46] The important desideratum of this principle, which lies somewhere in between token and type transparency is that it does not

[42] Shlomo Cohen, "A Philosophical Misunderstanding at the Basis of Opposition to Nudging," *The American Journal of Bioethics* 15 (2015): 39–41, 40.
[43] See also Andreas T. Schmidt, "The Power to Nudge," *American Political Science Review* 111 (2017): 404–17, 410.
[44] Bovens, "The Ethics of Nudge," 217; *emphasis added*.
[45] Hereinafter, we refer to Bovens's three categories as *type*, *token*, and *in principle* transparency.
[46] Schmidt, "The Power to Nudge," 410.

pose unreasonable burdens on people. In Schmidt's view, most nudges can be suitably transparent in this sense, especially compared with other kinds of policies, which might be much more opaque and complex.

While Bovens and Schmidt provide a useful framework, they never fully explain what exactly their notions of 'watchfulness' entail and what the implications are for nudge transparency. In what follows, we aim to analyze in much more detail what kinds of capacities are required for people to watchfully navigate the behavioral influences of nudges, following or rejecting them as they see fit. Although our preferred conception of watchfulness requires people to become more nudge-savvy than they currently are, we keep our expectations within constraints of feasibility. If the conditions for watchfulness are satisfied, we argue, in principle transparency is an attainable ideal that allows for the use of at least some nudges that work in the dark. We also argue that in principle transparency resembles and, in some cases, boils down to type transparency whenever watchfulness is in place. When watchful, it is enough for agents to be informed about the use of nudges at a more general level.

Before going into the core of our arguments, it is useful to distinguish between three 'axes' of transparency. First, Bovens's distinctions can be said to refer to different *degrees* of transparency along a continuum that runs from only disclosing general information (type transparency) to disclosing specific information about specific interventions (token transparency). To show what lies in between these extremes, consider the other aspects of a nudge that can be disclosed: the timing and place of its implementation, the goal or intention, the behavioral technique and its established impact, or the presumed causal mechanism that runs from the interference to the

outcome.[47] The more details disclosed, the closer we get to the token side of the continuum. Take the implementation of nudge-enhanced food arrangements in public cafeterias. The government can be specific about one of the aspects (e.g., goal) while remaining vague about another (e.g., timing). Or think of a cafeteria that puts a placard next to its entrance after having rearranged the food items. While this would count as token transparency on Bovens's account, the amount of information about the different aspects of this intervention can differ substantially.

Second, transparency can depend on the nudge's *design*. Some nudges are fully *transparent by design* and work exactly because they are visible to the nudgee, while other nudges do not and owe their effectiveness to their covertness. But even nudges that are transparent by design differ in terms of *when* the nudgee notices them. In cases of *ex ante transparent* nudges, the nudgee sees the nudge beforehand and can sidestep it if he so desires. Think of the urinal fly or traffic light labels (green, yellow, red) for healthy, less healthy and unhealthy food products. Conversely, a nudge is *ex post transparent* if the target notices the influence only when it has already taken effect. Think of fake potholes painted on roads to slow down drivers, or the use of defaults in contracts (such as health insurance). Only after you come to experience the consequences of the influence do you realize that you were nudged and that your status quo bias had an impact on your situation. Still other nudges are *non-transparent by design*. They are invisible to most nudgees, and only a well-trained eye can spot them. Think of framing effects, which are notoriously hard to detect, that tap into, for example, loss aversion.

[47] We should exercise caution about the last point. Grüne-Yanoff recently argued that the behavioral sciences are not adequately able to explain the causal mechanisms underlying behavioral change. Numerous behavioral studies successfully establish that alterations in the choice environment predictably change behavior, but fail to provide evidence that would account for the process that drives it. Here, we do not go against Grüne-Yanoff. All we claim is that the findings in behavioral sciences should convince people that there are *some* mechanisms at play, many of which are in the less-than-rational domain of cognitive processing. Grüne-Yanoff rightly cautions scientists against offering hasty explanations about which exact mechanisms are at play in any given case. We thank an anonymous reviewer for raising this concern. See Till Grüne-Yanoff, "Why Behavioural Policy Needs Mechanistic Evidence," *Economics and Philosophy* 32 (2016): 463–83.

Third, transparency can vary for different *people*. Here we determine whether nudges are detected or merely detectable depending on the capacities of the nudgee. A detectable nudge may actually be noticed only by some people. Think of governments trying to raise tax revenues by sending out messages that most people have already filed their taxes. Such a nudge is hiding in plain sight. Even though some may not notice the influence at all, others find its influence much easier to detect.[48]

Of course, these distinctions can be combined. A government can, for example, be type transparent about the intentions behind nudges (e.g., health promotion), token transparent about timing, placement and the techniques involved (e.g., the way food items are arranged in this or that cafeteria), but not disclose information about the way the techniques are presumed to work. Or one can have a look at one technique (e.g., lines on the road) and notice that its timing and placement are *ex ante* transparent but its intention is *ex post* transparent, except for people who have heard about this nudge and its underlying rationale before.

5. Reducing the scope of inquiry

The above typology helps us to clarify when the transparency issue is most relevant and which cases we can safely bracket from discussion as morally unproblematic. This section aims to reduce the scope of inquiry by putting aside those nudges where transparency is not an issue for the permissibility of implementation. This allows us to focus on the problematic cases in the next section.

[48] Not all nudges are like this. Even when disclosed, some nudges are notoriously hard to detect, and our inability to cope with their covertness gives good reason to avoid their use. More on this in section 8.

First, let us bracket nudges that are transparent by design.[49] For a nudge to be transparent by design, Pelle Hansen and Andreas Jespersen argue, both the intended shift in behavior and the means of achieving it must be overt and visible to the nudgee.[50] The famous urinal fly, for instance, has all the features of a transparent nudge. The visual stimulus provides sufficient content for the nudgee to consciously work out what behavior is being pursued (improving the aim) and by what means. Hansen and Jespersen mention other examples, including the 'look right/left/both-ways' signs on streets in the UK, car alarms for seat belts, GPS navigation, or colored footprints that lead individuals towards recycle dumpsters.[51] Since people are able to grasp these interventions before they take effect on behavior, they are what we have called *ex ante* transparent.

Since these nudges communicate directly with the nudgee's reflective capacities, Hansen and Jespersen call them 'type 2' transparent nudges.[52] In contrast to "heuristics-triggering"[53] nudges that tap into people's less-than-rational psychological mechanisms (shallow processing), these nudges are "informing" (if shallow processing is not affected) or "heuristics-blocking" (if they counteract the influence of shallow processing).[54] Because these nudges "respect the decision-making autonomy of the individual and enhance reflective decision-making,"[55] they do not pose any real concerns in terms of transparency. When they influence behavior through an 'update' of

[49] When we say 'by design', we assume that the nudge is a means of interacting with an agent who passes a minimal threshold of capacity for picking up on what is communicated. As we show later on, the effectiveness and transparency of a nudge always depend, to some degree, upon the nudgee's capacity. However, the threshold we assume here is very low. It is the capacity of an average Joe who has no insight into the character of heuristics and biases, does not know how behavioral techniques are used, or to what extent his choices are susceptible to their effects.
[50] Hansen and Jespersen, "Nudge and the Manipulation of Choice," 20–21.
[51] Ibid., 21.
[52] Ibid., 20. 'Type 2' is a nod to the reflective capacities of what dual-process theorists call 'System 2'.
[53] Adrien Barton and Till Grüne-Yanoff, "From Libertarian Paternalism to Nudging–and Beyond," *Review of Philosophy and Psychology* 6 (2015): 341–59, 343.
[54] Ibid.
[55] Baldwin, "From Regulation to Behaviour Change," 835.

people's beliefs, or make certain options more salient, these nudges work, but not 'in the dark' or 'behind people's backs.'

Sunstein refers to examples like the GPS, a type 2 transparent nudge, to argue that nudges do not threaten people's agency and autonomy.[56] While it is true that such nudges are not problematic in the relevant sense here, Sunstein's strategy does not address the worries critics have with other nudges. In contrast to Sunstein's broad definition of nudges, which includes strategies that merely induce reflection, nudges are often conceived more narrowly as tapping into the nudgee's shallow mental processing.[57] According to this conception, type 2 transparent nudges are comparable to billboards that communicate, by design, what course of action we may pursue. Note, however, that many instances of choice architecture design, such as saliently framing some piece of information, play into both shallow processing and reflective reasoning. If an intervention does tap into shallow processing, it certainly raises issues of transparency, which we tackle here.

Second, 'type 1' transparent nudges tap into the more automatic psychological mechanisms of System 1.[58] According to Sunstein, there is no principled reason why these should be ruled out as illegitimate or why they could not be made transparent: "so long as the initiatives are made public and defended on their merits, nudges should not be ruled off-limits merely because they work as a result of the operations of System 1."[59]

[56] Cass R. Sunstein, "Nudges, Agency, and Abstraction: A Reply to Critics," *Review of Philosophy and Psychology*, 6 (2015): 511–29, 512.
[57] Conrad Heilmann, "Success Conditions for Nudges: A Methodological Critique of Libertarian Paternalism," *European Journal of Philosophy of Science* 4 (2014): 75–94, 79.
[58] Hansen and Jespersen, "Nudge and the Manipulation of Choice," 21.
[59] Sunstein, *Why Nudge?*, 151.

Take the usage of shocking pictures of diseased lungs on cigarette packs or defaults and physical choice architecture.[60] The effect of these nudges is "more or less unavoidable to begin with, but transparent in a way that allows the influenced person to recognize the intention and means by which this is achieved as a direct consequence of the intervention."[61] Baldwin also talks about nudges that are detectable, despite their reliance on heuristics and shallow processing: "It is nevertheless the case … that the target of the nudge would be capable, *on reflection*, of realizing that a nudge has been administered and assessing its broad effect."[62] In our terminology, type 1 nudges typically become *ex post* transparent–so only after they have taken effect. Here, behavior is steered by a clear visual cue so that nudgees can reflectively notice being influenced as a by-product. While they typically do not or may not realize it while the influence is taking place, they can see it in the aftermath (*ex post*).

In contrast to the first category, the potential impact on people's autonomy of these nudges is real. People have reason to avoid heteronomous behavior that is "contrary to … their authentic will."[63] Type 1 shallow processing can arguably undermine people's autonomy because it bypasses their decision-making capacities. In this respect, *ex post* transparency may be insufficient to guarantee that nudgees act autonomously, if this depends on their ability to circumvent the nudge, which is diminished if the influence is not *ex ante* transparent. If transparency is meant to ensure that nudges do not divert agents from pursuing their goals and values, as Sunstein argues,[64] we ought to either rule out *ex post* transparent nudges as impermissible, or try to turn *ex post* into *ex ante* transparency.

[60] Baldwin, "From Regulation to Behaviour Change," 836.
[61] Hansen and Jespersen, "Nudge and the Manipulation of Choice," 21.
[62] Baldwin, "From Regulation to Behaviour Change," 836.
[63] Mills, "The Heteronomy of Choice Architecture," 497.
[64] Sunstein, *The Ethics of Influence*, 42.

This last strategy might work as follows. Sometimes *ex post* nudges become *ex ante* transparent to nudgees by mere exposure. A fake pothole may not have the same effect twice if the nudgee learns when and where to expect it, but he may fall for it if it appears elsewhere. The first time you are influenced by a picture of diseased smoker lungs, you only realize it *ex post*. However, after being exposed to the nudge more often, you can see the influence coming and it becomes easier to resist. Of course, the same technique can be used in new contexts where we have not learned to expect it. Empirical evidence shows how repeated exposure to a stimulus–like the kind on cigarette packs–diminishes the effectiveness over time, a phenomenon known in advertising as 'wear-out'.[65] Both anecdotal and more systematic evidence thus seems to suggest that agents indeed possess a general capacity for converting *ex post* into *ex ante* transparency, especially if the nudge runs counter to their goals and values.

All this convincingly shows that transparency does not only depend on the nudge's design but also on the circumstances and the capacities of the nudgee. If we want to achieve nudge transparency in a relevant sense, we should focus not only on the nudge (i.e., the design that can make nudges more or less visible) and the nudger (who needs to publicize his use of nudges), but also on the nudgee (who can be more or less experienced or vigilant).

Even with learning by exposure, a persisting worry remains that *ex post* transparency is insufficient for preserving autonomy. For decisions that carry particular weight, such as buying a house or making choices concerning education, nudges that are only transparent *ex post* seem problematic. It does not seem permissible in these weighty cases that nudgees fall for an *ex post* transparent nudge, even if it happens only once.

[65] Erin J. Strahan, Katherine White, Geoffrey T. Fong, Leandre R. Fabrigar, Mark P. Zanna, and Roy Cameron, "Enhancing the Effectiveness of Tobacco Package Warning Labels: A Social Psychological Perspective," *Tobacco Control* 11 (2002): 183–90, 186.

Third, when stakes are extremely high, one can wonder whether transparency is required, since autonomy is not the most important concern in these cases. Take nudges that aim at saving lives in traffic. Even in cases where they infringe upon people's autonomy and may arguably constitute a *pro tanto* wrong, they can still be justified overall because of the other values at stake.[66] Surely, making people comply with their enforceable duties and lowering casualty rates are amongst those values. Other vital concerns that outweigh autonomy have to do with human rights, liberties and justice more generally. Consider nudges that would make it less likely for people to exhibit racist behavior or violate fundamental rights of others. Given the importance of such ends, we do not think transparency should be required or regarded as a *conditio sine qua non* for democratic accountability and hence legitimacy. As such, we bracket nudges with such obvious and important advantages that these overrule the autonomy concerns raised.

There are two slippery slope worries lurking here. The first is the worry of sliding into perfectionism about the kinds of actions that can and cannot be valuable expressions of one's autonomy. Without taking this worry lightly, we can still claim that at least some trade-offs between autonomy and other values end up with autonomy getting the short end of the stick. While some governmental measures–speeding laws and tickets, red lights, speed bumps, but also the smart design of roads–can certainly be said to reduce people's autonomy, they are justified for good reasons. These measures are perfectionist only if their underlying rationale is that they are good for the person's autonomy, which is only one of the many possible justifications–the others being avoiding harm to others, avoiding health care costs, et cetera.

[66] Sarah Conly, *Against Autonomy: Justifying Coercive Paternalism*. (New York: Cambridge University Press, 2012); Thomas R. Nys and Bart Engelen, "Judging Nudging: Answering the Manipulation Objection," *Political Studies* 65 (2017): 199–214, 210–11; Wilkinson, "Nudging and Manipulation," 346.

The second worry is that, if autonomy takes second place, we might slide into justifying more aggressive behavioral techniques for achieving desired social ends. Why stop at nudging? Why not use subliminal messages, hypnotize or neuroenhance people to honestly fill out their tax forms? Now, as Sarah Conly argues, there are good reasons to believe that stronger forms of influence *are* permissible when other values clearly outweigh autonomy concerns, unless there are cases where we think it is particularly valuable that people autonomously act in blameworthy ways.[67] If using these stronger forms of influence is not feasible or if nudging is more effective in promoting the overriding value, we should resort to nudging without a second thought.

We reach two normative conclusions about transparency in cases of conflicting values. First, if some value clearly trumps autonomy, *ex ante* transparency is not required. Imagine a scenario where fake potholes are an affordable and highly reliable way of slowing down drivers ahead of a dangerous curve. The lives at stake justify putting these in place without disclosing this to drivers. Second, if there is uncertainty about which value should take precedence, an in principle transparency constraint suffices and should be respected. In fact, nudges are particularly useful in this regard, since they allow agents to balance out the values themselves. Take an intervention that nudges people into charitable giving, such as a clearly communicated automatic enrollment of employers into donation schemes with an easy opt-out option.[68] This strategy is sensitive to the nudgee's pursuits and allows him to sidestep the nudge. If a society is committed both to alleviating poverty and to respecting personal autonomy, people should be given leeway for balancing out the two as they see fit at different times.

[67] Conly, *Against Autonomy: Justifying Coercive Paternalism*.
[68] Cabinet Office Behavioral Insights Team and The Charities Aid Foundation (2013), "Applying Behavioural Insights to Charitable Giving," accessed online on September 17, 2018, at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/203286/BIT_Charitable_Giving_Paper.pdf.

Let us recapitulate. Since nudges are not a homogenous category, the purpose of this section has been to bracket cases where insisting on transparency is unnecessary: 1) when nudges are *ex ante* transparent by design and thus do not pose a threat to autonomy and 2) when other values obviously trump autonomy concerns. We will bracket these nudges and focus in what follows on 1) *ex post transparent* nudges and 2) *non-transparent* nudges. These are nudges that critics are concerned with when they argue, for example, that "nudges can be far less transparent than command."[69]

What are the typical examples of nudges that are non-transparent by design? Baldwin mentions framing, which arguably constitutes a serious intrusion into autonomy, because it is–by its very nature–difficult to detect: "Framing devices can be used to shape the decisions and preferences of an individual in a manner that is resistant to unpacking."[70] For instance, framing the risks of surgery in terms of dying (mortality rates) versus surviving (survival rates) has a predictable influence on people's choices.[71] When subjected to such a nudge, nudgees will not typically come to realize they are influenced (not even after the fact). These nudges typically tap into shallow processes, making them hard to detect, unpack and avoid.

Other examples are cafeteria food arrangements, decreasing the default size of food portions and drinks, or the use of social norms, the last of which were shown to be underappreciated by individuals as drivers of their behavior.[72] For most people (who are not well trained to detect them), these nudges really do work in the dark. The questions concerning their permissibility and whether and how people can overcome them are crucial.

---

[69] Baldwin, "From Regulation to Behaviour Change," 851.
[70] Ibid., 836.
[71] Barbara J. McNeil, Stephen G. Pauker, Harold C. Sox, and Amos Tversky, "On the Elicitation of Preferences for Alternative Therapies," *The New England Journal of Medicine* 306 (1982): 1259–62.
[72] Jessica M. Nolan, P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius, "Normative Social Influence is Underdetected," *Personality and Social Psychology Bulletin* 34 (2008): 913–32.

6. Watchfulness: importance and four conceptions

Perhaps the crucial aspect when it comes to assessing the transparency of nudge techniques is whether they are easily resistible. Yashar Saghai argues that a nudge is easily resistible if the nudgee "has the capacity to become aware of" the influence and "to inhibit her triggered propensity" to do as influenced.[73] *Ex ante* transparent nudges are clearly and by definition easily resistible. But what about *ex post* transparent and non-transparent nudges? According to Saghai, attention-bringing and inhibitory capacities can be activated even if influences are "'covert', that is, unannounced, and therefore not explicitly indicated to the influencee."[74] How can this work?

Apparently, nudgees can have a certain disposition that lies somewhere in between being fully oblivious (in which case the nudge is non-transparent) and fully aware (in which case the nudge is transparent). We call that disposition *watchfulness*, a capacity of individuals enhanced by resources at their disposal to detect *ex post* and non-transparent nudges that undermine their aims. Before we formulate a principle of watchfulness that allows citizens to navigate designed choice contexts and overcome threats to their autonomy in more detail, we lay out a number of reasons why watchfulness should be promoted throughout society.

Increasing watchfulness addresses the worry about nudges being one-size-fits-all policies in a world of diverse preferences and conceptions of the good life. In a society that lacks watchfulness, nudges indeed risk steering some people in ways they want to avoid. The potential of nudges to steer people towards the choices they would want to choose, had they not been burdened by cognitive biases, is realized best if people are watchful. Also, watchfulness enables citizens to avoid those choice architectures that are unlikely to facilitate reaching their goals

[73] Yashar Saghai, "Salvaging the Concept of Nudge," *Journal of Medical Ethics*, 39 (2013): 487–93, 489.
[74] Ibid.

while allowing them to set up or select their own choice architectures, which will steer their non-reflective behavior in ways they reflectively endorse.

While nudges have been said to reduce nudgees to passive followers,[75] watchfulness enhances their capacity to become 'planners' rather than mere 'doers.'[76] This both resembles and diverges from the idea of 'boosting.' According to 'boost' proponents,[77] nudgers assume that people inevitably fall prey to heuristics and biases and tap into these mechanisms to influence behavior. By contrast, boosts enrich and improve people's decision-making competences and skills. Making nudgees watchful by teaching them about heuristics and biases does not assume passivity but empowers people, quite like boosting does. But while watchfulness enables people to see nudges coming and circumvent them, boosts take it a step further, for example, by increasing peoples' statistical literacy and risk savviness.[78] Watchfulness is also less demanding than Peter John's 'nudge plus' policies, which encompass boosts, but also educative nudges and prompts to encourage slow thinking.[79] Our approach has similarities to John's, who also focuses on the need to democratize behavioral public policies through consultation and deliberation, but it is both more limited (we focus on more traditional nudges while John explicitly aims to broaden the set of policies) and more clearly focused on transparency (which John largely takes for granted).

Instead of dichotomizing between activating, empowering boosts and nudges that assume passivity, we instead suggest that nudges can be part and parcel of democratic self-government. As Schmidt convincingly argues, nudges that are implemented transparently and democratically

[75] Waldron, "It's All for Your Own Good."
[76] Thaler and Sunstein, *Nudge*, 42.
[77] Till Grüne-Yanoff and Ralph Hertwig, "Nudge Versus Boost: How Coherent Are Policy and Theory?," *Minds and Machines* 26 (2016): 149–83.
[78] Gerd Gigerenzer, *Risk Savvy: How to Make Good Decisions* (London: Penguin, 2014).
[79] Peter John, *How Far to Nudge? Assessing Behavioural Public Policy* (London: Elgar, 2018).

can democratize the control people have over their choice environments, thereby addressing the autonomy-related concern about nudges potentially increasing the extent to which people are subjected to alien control.[80] In a watchful society, doers (nudgees) are more likely to become planners (nudgers) and increase their control over their own behavior (via nudge-enhanced environments). Such a watchful society introduces a cooperative relationship between experts and laymen and blurs the distinction between them,[81] since these laymen can become aware of the influences and decide for themselves which to resist, which to endorse and which to create for themselves.

So, while the notion of 'watchfulness' seems promising in many respects, it is remarkably underdeveloped in the literature. We distinguish between four possible understandings of watchfulness. The first two stipulate fixed thresholds for the capacities that nudgees are required to have in order to be watchful. The third isolates the high capacities to an expert few who are entrusted with alarming the citizenry. The fourth conception, which we defend, provides conditions that are conducive to enabling people to spot and circumvent nudges as they see fit. Each of these conceptions has implications for which kinds of nudges are permissible.

First, according to a *minimalist* conception, the capacities of the nudgee–your 'average Joe'–are low so that ordinary people need not take special efforts to become watchful. While significant differences in terms of capacities can exist within the population, the focus of the minimalist conception is on establishing a minimal threshold that the vast majority of people supposedly cross. Nudges are then considered permissible as long as they can be unveiled by people of such a level of capacity. The attractiveness of this conception lies in its lack of demandingness.

[80] Schmidt, "The Power to Nudge."
[81] See also Nys and Engelen, "Judging Nudging: Answering the Manipulation Objection," 208–09.

Taking people as they are, it places the burden of responsibility on choice architects for designing easily detectable nudges.

The unfortunate consequence of the minimalist position is that it rules out *ex post* transparent and non-transparent nudges, even when these generate highly desirable effects. After all, your average Joe cannot pick up on such nudges (even if people with some training in behavioral techniques can). Since the minimalist conception of transparency regards only *ex ante* transparent nudges as permissible, it fails to cash in on the merits of nudges that work best in the dark. As we will explain later on, given that *ex post* transparent and non-transparent nudges play an important role in minimizing cognitive costs, the minimalist conception of watchfulness is too stringent.

The second conception of watchfulness is a *maximalist* one: it sets the threshold for watchfulness at a much higher level. The basic idea is that people should have some active awareness of the nudges that surround them. In order to keep on board the effective nudges that an average Joe fails to detect, this conception suggests that citizens should develop a high capacity for spotting nudges that are not easily detectable. As a result, they will detect nudges and adjust their behavior by either going along with them or rejecting them.[82] By contrast, low capacity individuals, like the average Joes, will "have very limited ability to adjust their behavior so as to reject messages that they disagree with and to act in ways that are inconsistent with such messages. They will, in turn, possess poor abilities to 'unearth' nudges such as defaults, and resist these."[83]

[82] Baldwin, "From Regulation to Behaviour Change," 840–42. Baldwin distinguishes high from low capacity individuals, but is not committed to the maximalist conception.

[83] Ibid. We believe Baldwin's use of the term 'message' is inappropriate here, as cognitive techniques become messages in the proper sense *only if* they are unearthed. They are messages for those who can spot them, but not for those who fail to understand their influence.

As mentioned before, moving from low to high capacity seems possible, as in cases where exposure helps turn *ex ante* into *ex post* transparency. Experiencing the same nudge repeatedly affects our capacities for picking up on its influence. However, citizens might need extensive training, which would require massive efforts from both their tutors and themselves. But, if feasible, this would relieve choice architects from being all too cautious as many nudges would become more easily spotted.

There are several problems with this view. First, it sets high demands on citizens, who are required to undergo extensive training and regularly update their knowledge about new behavioral techniques. Second, it is questionable whether individuals can be educated to surmount certain kinds of non-transparent nudges. In particular, empirical studies have shown mixed results regarding our possibilities to overcome framing effects. Robin LeBoeuf and Eldar Shafir, for example, found that inducing reflection does not make people less susceptible to framing: "framing effects are likely to persist even among careful thinkers."[84]

While the first two problems show that the maximalist conception violates feasibility constraints, a third problem arises if we assume people can actually become maximally watchful. Why, one may wonder, do we still need nudges if such levels of awareness and reflectiveness are attainable? This conception disregards that citizens may and indeed have good reason to welcome nudges exactly because they enable them to glide through some of their choices with little or no cognitive cost and thereby allow them to save up cognitive space for more important choices. We thus challenge the maximalist assumption that a permanent state of heightened attention would be a desirable state for autonomous individuals. Even if high capacities *can* be

[84] Robin A LeBoeuf and Eldar Shafir, "Deep Thoughts and Shallow Frames: On the Susceptibility to Framing Effects," *Journal of Behavioral Decision Making*, 16 (2003): 77–92, 89.

achieved, and individuals *could* effectively see through all non-transparent nudges, this is not what autonomy demands or how most people want to lead their lives.

A third conception of watchfulness goes back to Bovens's notion of in principle transparency, according to which a *watchful* person could identify the intentions behind nudges and blow the whistle on government transgressions. According to the *whistleblower* conception, some people should be particularly watchful in order to alert non-watchful others (the average Joes). Here, the burden of watchfulness does not fall on everyone, at least not equally: the proper way of fleshing out the government's duty of transparency is by effectively making it possible for a group of high-capacity individuals to publicly expose whenever the government engages in non-transparent and *ex post* transparent nudging.

While we acknowledge the strengths of this conception, it fails to reap the benefits of nudges working 'in the dark.' As Bovens rightly stresses, it would be advantageous to have "some watchdogs with sophisticated equipment keeping an eye on the government."[85] But Bovens also points out that having expert whistleblowers is insufficient: not only some experts but all citizens should be able to detect nudges: "We find it important that also *we ourselves* could decide to become watchful and unmask any manipulation."[86] Following Bovens, we believe that a conception of watchfulness should do more to empower lay citizens in being able to raise the alarm on nudges that work below the radar. Whistleblowing for laymen may indeed be possible when non-transparent nudges work against particularly weighty preferences. Think of seeing a family member die because of a 'do-not-resuscitate' default rule for advanced cancer patients. Nevertheless, the majority of non-transparent nudges remain detectable only to the expert few. The lay citizens' performance with such nudges depends on the successes or failures of

[85] Bovens, "The Ethics of Nudge," 217.
[86] Ibid.

designated experts to communicate with the laymen and help them navigate through such nudges.

There are two other reasons why the whistleblower conception falls short of a complete principle of watchfulness. The first is the worry of nudge stacking: the danger of too many nudges suffocating the nudgee in his attempts to navigate influences.[87] The whistleblower conception alone holds no constraints on how much nudging is to be permitted. Even if the whistleblowers end up competently revealing an endless sea of nudges, the nudgees nevertheless find themselves burdened in their attempts to grasp all the ways in which their choice environments have been altered. Secondly, the conception places whistleblowers at odds with the nudgers, quite like Wikileaks is a natural adversary to the US Government. While whistleblowers are indeed needed as a check on (the potentially sinister ends of) governments, we still miss out on the benefits of nudges in terms of their capacity to reduce cognitive costs. For this to be achieved, the relationship between whistleblowers and nudgers will have to go beyond an adversarial one and include a vibrant back-and-forth between all parties: government, experts, and lay citizens.[88]

To see the need to go beyond the whistleblower conception, recall the initial worry about nudges affecting the ability of people being in control of their lives. If this is what transparency ought to secure, then we should not put our trust only in the expert few to signal potentially problematic choice environments. If self-rule is at the core of autonomy, then citizens need to be able to govern themselves and thus exercise control over nudges. Nudges are promising, exactly because they can promote autonomy, by reducing cognitive costs and thus enabling citizens to focus on choices that matter. As Sendhil Mullainathan and Eldar Shafir put it, people operate with a

[87] Christian Coons and Michael Weber, "Introduction: Paternalism–Issues and Trends," in *Paternalism: Theory and Practice*, ed. Christian Coons and Michael Weber, 1–24, 21 (New York: Cambridge University Press, 2013).
[88] It could be argued that avoiding nudge stacking itself helps in overcoming this adversarial relationship. The more nudges there are, the easier it becomes for governments to hide illicit nudges from whistleblowers.

limited *cognitive bandwidth*. Faced with scarcity of time and resources, mired in pressing deadlines and worry, people often *tunnel*–use up their limited reflective capacity and executive control for urgent tasks at hand. At such times, they find themselves more prone to error in other tasks and decisions.[89] Operating with a limited bandwidth for reflection may well mean that people will often be unable to properly and effectively divide their attention to all the goals they wish to pursue, especially when they need to focus on pressing matters. Since non-transparent and *ex post* transparent nudges work below the radar, they do not incur costs by using up cognitive bandwidth. They can help in economizing attention, cognitive capacity and willpower and thus relieve people from having to divide their attention and aid them in pursuing some of their goals effectively at low capacity. Nudges offer cognitive relief in another sense as well. Since nudging is a second-order strategy that delegates the design of choice architectures to qualified nudgers, "it exports decision-making burdens to someone else, in an effort to reduce the agent's burdens both before and at the time of making the ultimate decision."[90]

We now come to the fourth conception of watchfulness, which we defend here, namely a democratic conception of watchfulness. It builds on the whistleblower conception, but incorporates opportunities not only for contestation, but also for deliberation and participation. Nudgers can indeed cooperate with nudgees in the process of steering behavior towards ends that the latter would endorse. Here, experts act not only as whistleblowers, but also as consultants: they advise both government and citizens and help the latter navigate through the nudges in place.

[89] Sendhil Mullainathan and Eldar Shafir, *Scarcity: Why Having Too Little Means So Much* (New York: Times Books, Henry Holt and Company, 2013).
[90] Cass R. Sunstein and Edna Ullmann-Margalit, "Second-Order Decisions," *Ethics* 110 (1999): 5–31, 16.

The idea is that citizens should have an *opportunity for watchfulness* so that they can willingly fluctuate between episodes of low and high capacity. The opportunity for watchfulness entails that citizens need not be conscious about and prepared for every nudge that influences their behavior. Instead, they should have resources readily available when nudges guide them towards actions they would want to avoid, given their reflective judgements about how they want to lead their lives. This means that they can have different levels of capacity for picking up on different types of nudges in different domains. Rather than being able to notice every single instance of nudging, they should be able to notice the behavioral techniques the purposes of which they disagree with.

Watchfulness is thus a disposition the actualization of which can be triggered at different times in different ways. First, a watchful nudgee with weighty preferences that go against some nudge can notice he is being steered in a direction he does not endorse. Saghai refers to evidence from cognitive psychology suggesting that "stimuli that … produce a feeling of dysfluency are more likely to trigger scrutiny … At least when individuals have strong and settled enough preferences, goals, or beliefs, they are likely to become aware of an anomaly". In such cases, people's "attention-bringing capacities" are activated and their disposition of watchfulness is actualized.[91] Second, a watchful nudgee who is familiar with some behavioral technique can recognize it in a similar or new choice setting (thus converting *ex post* transparent or even non-transparent nudges into *ex ante* transparent nudges), enabling him to more easily bypass it. While the first remark is good news for nudge enthusiasts, the second aspect shows that there is some need to educate citizens.

[91] Saghai, "Salvaging the Concept of Nudge," 489.

The main advantages of this conception are twofold. First, it does not automatically rule out potentially desirable nudges and cashes in on their benefits. Second, it is realistic and not overly demanding in its expectations towards citizens. So far, we have argued that the democratic conception of watchfulness should be promoted because: 1) heuristics-triggering nudges are permissible only if there is such watchfulness; 2) such nudges are desirable because they promote desirable outcomes while preserving autonomy (by reducing cognitive costs and allowing people to pursue their conception of the good). In the next section, we flesh out more fully what the implications are for the kind of nudge-enhanced but watchful society that we deem desirable.

7. Watchfulness: normative demands

What are the precise normative demands that our conception of watchfulness poses? We argue that four conditions need to be fulfilled: an educational, a legal, a democratic, and a societal one.

First, nudgees require a basic education in order to understand and acknowledge the impact of heuristics, cognitive biases, choice architectures and behavioral techniques. Unlike the maximalist, we require only a minimal understanding and acknowledgement of heuristics as important drivers of everyday human behavior. This is also less ambitious from what proponents of 'boosting' try to achieve, namely a significant improvement of people's decision-making competences. In addition, nudgees should be educated about the available resources for looking up nudges that would steer their behavior in what they consider undesirable directions.

That is why the second condition stipulates a legal requirement, namely that governments publish a nudge registry: a resource readily available to citizens which compiles the nudge

techniques implemented by governments.[92] For example, nudges and their goals could be published explicitly on a website such as nudge.gov.uk (similar to legislation.gov.uk–the online database of UK statute law) or in the Federal Register (the official journal that publishes legal rules and notices of the US Federal Government). Since transparency enables citizens to understand the justification for the policies they are subjected to, it has an important democratic role to play. This addresses the worry about nudges obscuring this process if they are not "visibly brought into the legal system."[93] It enables watchful citizens to figure out which behavioral techniques their governments are using and why, and helps them circumvent those that are likely to conflict with their goals. As Robert Lepenies and Magdalena Małecka suggest, such a nudge registry functions as a legal codification of nudges, increasing their visibility.[94]

Third, nudges need to be democratically legitimized. It is crucial that nudgees are able to figure out what aims nudging governments pursue. Nudge strategies are often inherently tied to particular kinds of behavior, but not necessarily to particular aims. Consider the cafeteria food arrangement example. Even if people do not agree with the government paternalistically helping them lead healthier lives, they might agree that there is a collective duty towards reducing health care costs. Knowing the aims pursued by nudges helps people in focusing exactly on those nudges related to behavior they find significant with respect to how they want to lead their lives. Our requirement on aims thus reintroduces the publicity principle, which Thaler and Sunstein[95] have adapted from Rawls,[96] as part of our principle of watchfulness.

---

[92] Robert Lepenies and Magdalena Małecka, "The Institutional Consequences of Nudging: Nudges, Politics, and the Law," *Review of Philosophy and Psychology* 6 (2015): 427–37, 435.
[93] Ibid, 433.
[94] Ibid., 430.
[95] Thaler and Sunstein, *Nudge*, 244–245.
[96] See also Hansen and Jespersen, "Nudge and the Manipulation of Choice."

Citizens usually acquire knowledge regarding policy aims not just through governmental disclosure, but through public debate. A public debate on nudge goals, we believe, helps citizens become aware of nudges and coordinate their preferences with others and in the face of expert information. On the side of the government, "citizen input can guide public decisions, which feeds into more responsive and efficient policies."[97] The exact institutional form of the communicative channels and deliberative fora in which nudge goals would be presented and discussed is uncertain, and will have to be determined in practice. The socially desirable form should succeed in gathering as many stakeholders as possible, tracking their long-terms goals and raising weighty concerns. For instance, it has been suggested in the UK that behavioral influences are geared toward localized problem-solving,[98] which implies that debates on nudge goals should also be organized locally. Still, given that people often give most attention to political deliberation on the national level, it is not obvious that a localized debate reaches the greatest possible number of those affected.

The democratic and the legal requirement (a nudge registry) complement each other, if certain practical considerations are taken into account. We mention two here. The first once again points to the problem of stacking, i.e., to circumstances in which an overwhelming number of nudges paralyzes citizens in attempting to figure out how these nudges work and why they are implemented. Establishing a partnership between nudgers, nudgees, and experts, our democratic conception of watchfulness curbs nudge implementation to what can reasonably be expected of a nudgee's attention span, and allows nudgees to veto further (stacking of) nudges through democratic debate and thus to limit the number of nudges being implemented and listed in the

[97] John, *How Far to Nudge?*, 125.
[98] Adam Burgess, "'Nudging' Healthy Lifestyles: The UK Experiments with the Behavioural Alternative to Regulation and the Market," *European Journal of Risk Regulation* 1 (2012): 3–16, 5–6.

registry. The second concern is that the registry should provide fairly stable and precise depictions of nudge techniques in practical contexts to enable nudgees to successfully identify any intervention as one that they would want to avoid. Moreover, if similar techniques are used to steer behavior for different kinds of purposes, nudgees may fail to successfully distinguish between them. In short, nudges in the registry have to be neatly illustrated and take fairly stable manifestations in public life.

Finally, the fourth condition, a societal one, incorporates the whistleblower conception into our account of watchfulness. In some cases, watchful citizens with superb knowledge of behavioral techniques or a particular interest in a specific domain can and should warn others that unannounced non-transparent nudges are operational. Unauthorized nudges might come to rise either spontaneously or as part of sinister scheming. Since a watchful person can unfold such nudges and their underlying aims, and blow the whistle when necessary, what we need in a watchful society is a sufficient number of such experts. These experts may be specialized in nudges within a particular field or with regard to particular forms of nudges.

Consider the fact that physicians, in their interactions with patients, can use framing techniques,[99] which are particularly resistant to watchful citizens. Even if people publicly debate on health policy aims and there is a health nudge registry in place, nudge-educated and watchful citizens may have a difficult time detecting and opposing such hard-to-resist framing techniques. Here, the whistleblowers might provide a helping hand. Health experts can help keeping a watchful eye on nudges in health care, political experts can focus on agenda setting and formulating referenda questions, and so on.

[99] Jennifer S. Blumenthal-Barby, "Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts," *Kennedy Institute of Ethics Journal* 22 (2012): 345–66, 361–63.

As with other policy measures and domains in well-functioning democracies, there need to be enough people with a strong interest in and knowledge of some particular issue to check, influence and contest what the government is doing. This provides a middle road solution to 1) a completely apathetic and uninformed society (where there is not enough actual watching going on) and 2) a completely active citizenry (where there is more than enough actual watching going on). In short, it is a middle road between 1) the minimalist conception of watchfulness, which is insufficient in light of the contestatory and participatory role citizens should have in democracies, and 2) the maximalist conception, which is overly burdensome to citizens and unnecessary for well-functioning democracies. In order to know whether a society is sufficiently watchful and vigilant, we need to look at the collective level, and check whether there are enough nudge experts and citizens active in different domains, and not just at the individual level, which is the focus of both the minimalist and maximalist conception.

Let us clarify the implications for nudges that are non-transparent (but in principle transparent) to non-experts. These are not ruled out due to their properties, provided there is an easily accessible nudge registry and a public discussion on the underlying intentions of particular nudge policies. Our principle of watchfulness thus allows non-transparent nudges only if 1) they result from proper democratic procedures, with citizens actively joining public debates, 2) the span of utilized nudge techniques is regulated, and 3) experts act as whistleblowers and aid citizens in detecting and navigating nudges. If there are nudges that nudgees find detrimental to the pursuit of their goals, they should be able to switch from low to high capacity and steer clear of their effects. As for areas where citizens agree with the intention, they can safely stay oblivious. In fact, citizens have an interest in staying at low capacity in these cases, as that enables them to minimize costs in the face of limited cognitive bandwidth. We have no interest in making all of

our choices reflectively and unearthing the nudges that steer our behavior towards ends with which we agree.

Bovens rejects type transparency since it is insufficient to make nudges permissible, quite like a general disclosure about the usage of subliminal messages fails to legitimize such a means. Instead, Bovens recommends in principle transparency, with watchful people being able to spot nudges. We believe that our principle of watchfulness adequately fleshes out what Bovens means with in principle transparency and why he (rightly) considers it to be valuable. In addition, however, we believe there might be little difference between in principle and type transparency in any watchful society.

After all, Bovens's type transparency can be understood in different ways, with regard to what and how much is disclosed. What may be disclosed and democratically agreed on are any particular nudge's aims, techniques and time or space frames. Consider, for example, a local government disclosing that it will start painting lines on roads on a particular date to minimize traffic collisions: this involves all of the components just mentioned. Type transparency may be compatible with remaining silent about any of these components, while still disclosing that some nudging will take effect and alarming nudgees to be on guard. But notice that, the more a nudge registry discloses, the more type transparency resembles Bovens's in principle transparency, provided our principle of watchfulness is incorporated.

8. Objections

Before concluding, let us tackle a number of objections that might be raised against our account of watchfulness.

The first objection argues that there is particular significance in agents reflectively acting on their long-term goals and bringing them to fruition by means of conscious choices and not via more automatic processes. For one, consciously acting on one's reasons might be constitutive of the value that we ascribe to pursuing our own conception of the good. We may also think that we need to act consciously in order to acquire a full range of reasons that may later lead us to revise our conceptions of the good, which is something that happens often over the course of our lifetimes. Using nudges as a walking stick is thus impermissible for the full realization of one's autonomy.

This objection mistakenly clings to only one aspect of how we realize our autonomy, and is thus too demanding. Three responses are relevant here. Firstly, many of our choices that strongly pertain to our conceptions of the good are acted on automatically whether we like it or not. If behavioral science has taught us anything, it is that we operate with only a limited capacity for reflective choices, while our remaining actions result largely from the workings of our cognitive heuristics. It is thus imperative–and democratically legitimized nudges can help here–to align our less-than-fully reflective behavior with our reflective judgments on how we want to live our lives. Secondly, our conception of watchfulness helps nudgees to reflectively consider how some choice environments lead them away from their autonomous pursuits. Consciously circumventing particular strands of behavior certainly helps agents in reflectively re-assessing

what is valuable about their conception of the good.[100] Thirdly, our conception also allows us to think of nudges as social pre-commitments to certain kinds of behavior. The analogy to individual pre-commitment is made stronger because individuals can easily opt out of the social pre-commitment when transparency is in place. And since we find nothing wrong with people individually pre-committing to certain kinds of behavior in the pursuit of their autonomous goals, we should think the same for social pre-commitments.

Granted, the analogy claim here is moving a bit too quickly. Collective pre-commitment runs into problems of monitoring, accountability and sinister ends, which do not arise with individual pre-commitment. Our account of transparency aims to alleviate or even resolve such problems. Note, however, that the analogy also breaks down in ways that favor collective pre-commitment in the form of nudging. Remember that nudging, as a second-order strategy, can be more effective at reducing cognitive costs overall, given that the cognitive work of planning our decision-making at low capacity is carried out by persons other than the decision-makers themselves.[101] Therefore, it is not the case that if everyone is effectively carrying out individual pre-commitments, collective pre-commitments can be eliminated at no cost.

The second objection raises the concern that making non-transparent nudges transparent by means of disclosure does not guarantee resistibility. Non-transparent nudges may be disclosed yet still remain irresistibly effective in steering behavior. Coons and Weber argue that disclosure in such cases may actually make things worse. If I were told that I would be coercively injected with a love potion, not only would the disclosure fail in making the act permissible, but it could

[100] If it is argued that agents need to pass at least some threshold of making a sufficient number of important choices reflectively, we need not disagree with this, nor do we think that nudges curb the passing of this threshold. Nudges do not hypnotize the agent nor do they block his assessment of his own behavior. It is also not in any way obvious that nudgees have less data to work with when they wish to revise their conceptions of the good.
[101] Sunstein and Ullmann-Margalit, "Second-Order Decisions," 13.

make it worse because I could notice my agency slipping away. Coons and Weber claim that some nudges could have this effect, like the (*ex post* transparent) distorting mirrors which make you look obese.[102]

Our response to this objection is straightforward. There is nothing in our conception that commits us to endorsing any nudge that cannot be resisted. Disclosing non-transparent and *ex post* transparent nudges only affects their permissibility if this makes them *ex ante* transparent, and thus resistible. With some techniques, like subliminal messaging (or love potion injections), this will not be the case. But just as we cannot draw general conclusions about the impact of transparency on the resistibility and permissibility of nudges from harmless cases, such as Sunstein's favorite example of the GPS, we cannot reach verdicts about the general resistibility of nudges from a handful of techniques which are not resistible. Nudges are a diverse category, and some are certainly such that they can be made easily resistible with transparency in place. For any techniques that do not conform to these characteristics, governments have weighty reasons not to put them to use.

Nevertheless, this objection should give us pause. In many cases, the resistibility of a nudge, which depends on triggering particularly strong heuristics, or generating substantial social and emotional costs, will be difficult to assess.

The third objection asks why we would still rely on nudges if we can make people watchful? And why add governmental nudges if there are already so many influences that a watchful society needs to unearth, ranging from random choice architectures to private nudges? Sure, the objection goes, it is a good idea to increase watchfulness, but this should primarily help people to

[102] Coons and Weber, "Introduction," 20.

avoid the errors they already make due to manifold influences, not serve as an excuse to add even more intentionally designed nudges and thereby increase the burdens of watchfulness.

We provide two responses here. First, increased watchfulness, knowledge and awareness about behavioral techniques and their impact can and should indeed help people more easily resist random choice architectures and unintentional or private nudges that go against their goals and values. It strengthens their capacity to navigate through whatever influences they encounter (be they well-intended, less well-intended or not intended at all). Second, we remind that in our conception, watchfulness is not the same as being completely reflective at any given time. Given the limits of human psychology, a world with nudges will always have benefits that a world without nudges will lack. When choice architectures are designed so as to successfully facilitate the choices people reflectively want to make, this frees up the cognitive space and effort that people can employ to focus on other choices.

The fourth objection stresses a distributive concern. Sure, democratically justified nudges may benefit the majority, whose choices are facilitated towards the goals they endorse, but they also impose cognitive costs on the minority, whose reflective preferences go against the implemented nudges.[103] Is it not unjust to disproportionately impose burdens 1) on dissenters who oppose being nudged altogether, and 2) on people who do not endorse the directions of most nudges?

In response, it is not a plausible or feasible ideal for governments to try to design some neutral choice architecture that imposes equal cognitive burdens on all citizens. If any one-size-fits-all nudge policy disproportionately advantages some citizens at the cost of others (in terms of cognitive burdens), then it makes sense for governments to focus on those nudges where they can

[103] This is a related but slightly different concern to the one voiced by Joel Anderson ("Review of Nudge," 373**)**, who stresses that "individuals vary in their capacities to resist … This already raises concerns (nowhere addressed in *Nudge*) about the equality of effects that nudges have."

safely assume what most people's goals and values are. With both high-stakes nudges (such as avoiding deaths in traffic) and low-stakes nudges (such as facilitating how to open doors by installing the right kinds of handles[104]), these goals and values are largely beyond doubt. In both cases, nudges have strong legitimacy. In cases where a large part of the population disagrees with the direction of the nudge,[105] nudges have low legitimacy, both in democratic terms, because they will lack consent, and in distributive terms, because they impose cognitive costs on many people.

9. Conclusion

In this paper, we have laid out a principle of watchfulness that is compatible with and expands on Luc Bovens's analysis of nudge transparency. We argue that it succeeds in securing personal autonomy, while cashing in on the merits of various nudge techniques that work best 'in the dark.' Instead of focusing exclusively on the specific design of nudges and the behavioral techniques involved, we argue in favor of specific measures that target both nudgees (making them more watchful) and nudgers (forcing them to disclose nudge strategies and goals). We make specific recommendations about the educational, democratic, legal and societal conditions that need to be fulfilled for *ex post* transparent and non-transparent nudges to be permissible. We show that, with these feasible background conditions for watchfulness in place, in principle transparency advocated by Bovens comes close to type transparency. Also, this allows for a nuanced discussion of the permissibility of different nudge techniques while keeping an eye on democratic principles and constraints.

[104] Even the innocuous nudges (like the door handles) matter, again because they facilitate most choices, reduce cognitive costs and thereby free up cognitive bandwidth for more important choices.
[105] Reisch and Sunstein, *Do Europeans Like Nudges?*

The normative suggestions that we provide here do not only sketch the contours of some utopian watchful society, but have implications for the legitimacy of nudges here and now. Our conception is fully compatible with modest nudge technique programs, the implementation of which can progress with equally modest steps. This modesty of our proposal can be understood in two ways: in numbers (it avoids nudge stacking) and in force (it rules out irresistible nudges). Our analysis thus has the advantage of striking an adequate balance between the complacency of minimalism and the demandingness of maximalism. By increasing watchfulness, people are made more nudge-savvy, which helps avoid impermissible nudges (which would be whistleblown or would not even make it through democratic processes), without having to give up on the benefits of permissible nudges. Watchfulness empowers nudgees and thus addresses the worries of critics, while salvaging the benefits nudge enthusiasts aim for.

Acknowledgements

We would like to thank TiLPS (Tilburg Centre for Logic, Ethics, and Philosophy of Science) for giving a visiting fellowship to Viktor Ivanković, which made this collaboration possible. We would also like to thank the following people for having read and commented upon previous versions of the paper: Luc Bovens, Andreas Schmidt, Andrés Moles, Lovro Savić, Chris Mills, Fay Niker, Matthew Clayton, and Mihovil Lukić. We are also grateful to two anonymous reviewers whose comments and suggestions improved the article significantly. Finally, we are grateful for the helpful comments received at the 'Nudging, Autonomy and Transparency' workshop, organized at WINK: The Nudge Conference, Utrecht, 2017.