**The Possibility of Epistemic Nudging**

Thomas Grundmann

University of Cologne

Department of Philosophy

Albertus Magnus Platz

50923 Köln

Germany

Email: thomas.grundmann@uni-koeln.de

phone: +49 221 470-4477

fax: +49 221 470-5006

ORCID ID: [https://orcid.org/0000-0002-7276-7345](https://orcid.org/0000-0002-7276-7345?lang=en)

**ABSTRACT**. Typically, nudging is a technique for steering the *choices* of people without giving reasons or using enforcement. In benevolent cases, it is used when people are insufficiently responsive to reason. The nudger triggers automatic cognitive mechanisms—sometimes even biases—in smart ways in order to push irrational people in the right direction. Interestingly, this technique can also be applied to *doxastic attitudes*. Someone who is doxastically unresponsive to evidence can be nudged into forming true beliefs or doxastic attitudes that are propositionally justified. When doxastic nudging uses non-rational mechanisms, the worry is that nudging cannot result in justified beliefs or knowledge, as the resulting doxastic attitudes lack the right kind of basis. In this paper, I will argue that given the right background views about knowledge, justified beliefs and the relevant processes, *epistemic nudging* is possible even in these cases. That is, all kinds of nudging can—in appropriate circumstances—produce justified beliefs or knowledge in the nudgee.

**Keywords**: doxastic nudging; nudging to knowledge; external methods

**Biographical note.** Thomas Grundmannis Professor of Philosophy at the University of Cologne, Germany. He has published numerous articles on general epistemology, applied epistemology, and philosophical methodology. Among his books are *Die philosophischen Wahrheitstheorien* (2018), *Analytische Einführung in die Erkenntnistheorie* (2017), and *Der Wahrheit auf der Spur* (2003). His current research addresses issues from social epistemology such as epistemic authority and experts, the limits of critical thinking, epistemic autonomy, disagreement, and post-truth. He is the co-editor of *The Epistemology of Fake News* (2021) and *Experimental Philosophy and Its Critics* (2012).

Nudging[[1]](#footnote-1) is a technique that is neither special nor new. It has always been used when people dress up their arguments in words that are particularly persuasive, or when they advertise products by associating them with emotions that motivate others to buy those products. As a first approximation, one may characterize nudging in this way: If A, the nudger, nudges B, the nudgee, then A intentionally steers B’s behavior or judgment by triggering automatic mechanisms in B’s mind. Everyone knows that people can be influenced using techniques of rhetoric or advertisement, and everyone uses this knowledge at times. What is new about nudging is that it has become a topic of systematic research and that the prospects of influencing people’s behavior through nudging have been explored more thoroughly in recent years (Thaler & Sunstein 2008). The background of this development is the discovery by behavioral economists and psychologists that people’s choices or judgments are systematically influenced by irrational factors, e.g., priming, anchor effects, confirmation bias, and many others (Kahneman 2011). There are two different ways of responding to this discovery. One can either try to train people in such a way that they become, to a certain extent, shielded from such irrelevant influences (Nisbett 2015), or one can try to influence people by triggering these biases in ways that lead to desirable behavioral consequences or judgments. Such *smart* use of cognitive biases constitutes the bulk of nudging.

What are the characteristic features of nudging? *First*, nudging typically influences the behavior of the nudgee, not by giving additional reasons but by triggering automatic, non-rational cognitive mechanisms. In this sense, nudging typically bypasses the nudgee’s reasoning capacity. In these cases, nudging seems to involve manipulation rather than the rational agency of the nudgee.[[2]](#endnote-1) *Second*, nudging is neutral with respect to whether it harms or benefits the nudgee. One can be nudged by a greedy seller to buy overpriced products, just as one can be nudged to buy products that are better for one’s health. If nudging serves the interest of the nudgee (or promotes a common good), it is an instance of *paternalism*. Paternalists improve the decisions of otherwise irrational people without their consent (Ahlstrom-Vij 2013: 39). *Third*, nudging influences behavior without taking away any options and without using physical force. Nor does it enforce choices through indoctrination or brainwashing. The influence must be resistible, such that free choice is not prevented. In this sense, nudges preserve the nudgee’s freedom of choice. Nudging is, therefore, a method of *libertarian* paternalism (Thaler & Sunstein 2008: 5). Originally, Thaler and Sunstein (2008: 10-11) defined nudges so broadly that every aspect of the choice architecture, whether intentionally created or purely accidental, may constitute a nudge. In this paper, I will focus exclusively on *intentional* nudging. If one understands nudging in this way, it is not unavoidable, even if irrelevant influences on the agent’s decisions are (Saghai 2013: 491). One might argue that it is better to put healthy food at eye level in the cafeteria, in order to increase the frequency of people eating healthy food, rather than let random factors determine which products are bought because it is unavoidable that this privileged position influences people’s choices anyway. However, only the former will be a case of *intentional* nudging.

When Thaler and Sunstein (2008) initiated the recent scientific and political discussion on nudging, their primary focus was improving people’s *behavior*. For example, people are not disposed to save a lot of money for their retirement. They irrationally prefer short-term goods (e.g., spending money now) to long-term goods (e.g., having more money later). However, if part of their salary is saved by default, this changes their behavior significantly. Since it takes at least some effort to decide against the default option, most people stick with the option of saving. Something similar happens in the case of organ donation. If donating is the default option and one must actively decide against organ donation to prevent it, significantly more people will donate organs. Or take the case of vital surgery. If physicians frame the decision to undergo some such surgery in terms of rates of success rather than failure (“95% survive this kind of surgery” instead of “5% die from this surgery”), the frequency of decisions in favor of the surgery rises significantly. Finally, consider the effect that the pictures of a refugee child’s corpse on the coast of Bodrum had when they appeared in the news. These pictures triggered empathetic emotions that substantially increased the European public’s readiness to help refugees (Mackey 2015).

How should one define “nudging” as applied to influences on behavior? According to Thaller and Sunstein (2008: 6),

[a] nudge […] is any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid.

Nudges thus influence decisions without, on the one hand, changing options or giving further reasons, and without, on the other hand, determining decisions. To more narrowly characterize *intentional nudges,* I suggest the following definition, inspired by Saghai (2013: 491):

A *nudges* B *intentionally* when A acts in a way that intentionally makes it more likely that B will φ, primarily by triggering B’s shallow cognitive processes, while A’s influence preserves B’s choice set and B’s freedom of choice.

Here, the phrase “shallow cognitive processes” refers to cognitive processes that are neither fully deliberative nor reflective. Typically, these automatic cognitive processes are called “S1-processes”, in contrast to deliberative and reflective S2-processes (Kahneman 2011: 20-24). Admittedly, there are at least some nudges that target S2 – rather than S1 – processes. For example, one can nudge people’s choices by placing warnings on the product or by attaching calorie labels or energy efficiency labels (Sunstein 2016: 80). In these cases, nudging engages with S2 – processes within the nudgee by providing her with information. However, the bulk of nudges are, as I will call them, *brute nudges*, i.e., nudges that are exclusively targeted at S1-processes within the nudgee’s cognitive system. Here I will only address *brute intentional* nudges that trigger shallow cognitive processes.

More precisely, in this paper, I will focus on an even more narrowly defined kind of nudging that I call “doxastic nudging”. Obviously, we can intentionally trigger certain automatic belief-forming mechanisms in order to make people believe (or disbelieve, or suspend judgment on) certain propositions. We thus can nudge people into adopting certain doxastic attitudes. In section 1, I will explore the scope, value and limits of doxastic nudging. In particular, I will define doxastic nudging and explore the epistemology of nudging that is based on it. At first glance, it seems perfectly plausible that brute nudging may result in true or propositionally justified beliefs, but it cannot generate justified beliefs or knowledge. In section 2, I will criticize Neil Levy’s recent claim that brute nudging typically does not bypass reasoning, but rather feeds our non-reflective cognitive mechanisms with additional reasons. In contrast to Levy, I will argue that a substantial part of brute nudging bypasses reasoning. In section 3, I will argue that brute nudging, even if it bypasses reasoning altogether, may still result in justified beliefs or knowledge. It follows that epistemic nudging, i.e., nudging that results in epistemic statuses such as justified belief and knowledge, is possible even in these cases. Interestingly, these epistemic states do not require epistemic agency or any epistemic competence on the part of the epistemic subject. If the nudger steers the doxastic states of the nudgee in the right way, she will thereby induce justified beliefs and knowledge. Finally, I will summarize the results of the paper in a brief conclusion.

Before going into details, I will clarify my relevant background assumptions and use of terms. First, I argue within a broadly reliabilist framework, according to which epistemic justification requires reliable belief-forming processes and knowledge requires perfect reliability, i.e., safety. Second, my use of the term “reasoning” deviates from standard usage. Usually, “reasoning” refers to belief-formation on the basis of genuine reasons. In contrast, I will use this term for cognitive activity that properly responds to genuine and apparent reasons. In this sense, a brain in a vat is able to reason, but does not respond to genuine reasons. My use of “reasoning” closely resembles that of “rationality.” In fact, I use these two terms interchangeably.

*1. From behavioral to doxastic nudging*

We can steer someone’s practical choices and behavior by triggering some of her automatic cognitive mechanisms, but we can also steer her doxastic attitudes by triggering automatic non-rational mechanisms of belief-formation. For example, we can make people believe certain propositions by rendering those propositions particularly salient or framing them in especially persuasive ways. We can increase the likelihood of a proposition being believed by mentioning it in the first place. We can persuade people that a particular action was morally wrong by directing their attention to its horrible details. We can make people believe that certain propositions are true by presenting them as common ground, as default assumptions or as methodological presumptions. And we can influence people’s numerical estimations by setting an appropriate anchor. In all these cases, doxastic states are influenced not by giving any reasons, nor by enforcing them in any direct way, but rather by triggering our biases (salience and framing effects, order effects, affective bias, status quo bias, anchor effects) in smart ways. Since these influences on our beliefs are neither rational nor cases of brute enforcement, they can be classified as “doxastic nudging.”

We can now turn these considerations into a definition of brute and intentional *doxastic* nudging. For this purpose, the definition of behavioral nudging must be adjusted in certain respects. Here is my suggestion:

A, intentionally and brutely, nudges B doxastically to have doxastic attitude DA =df. A acts in a way that intentionally makes it more likely that B will have DA by triggering B’s shallow cognitive processes, while A’s influence preserves both the represented facts and B’s rational control concerning DA.

What is clearly different from behavioral nudging is the target, i.e., doxastic attitudes instead of choices or behavior. Apart from this, two further aspects must be accommodated. In the case of doxastic nudging, the relevant factors comprise the facts instead of the choice options that must stay fixed. Moreover, we must substitute freedom of choice with rational control, since doxastic voluntarism is untenable (Alston 1989). What does it mean to say that the nudging of certain doxastic attitudes preserves the nudgee’s rational control concerning these attitudes? The main point here is that when doxastic attitudes are triggered by nudges, this does not remove the nudgee’s ability to revise these states in the light of further evidence. Thus, nudging does not isolate doxastic states from rational influences, as indoctrination or brainwashing would do.

Doxastic nudging can be evaluated from different perspectives. Here I will exclusively focus on the epistemic perspective. Due to its neutrality, doxastic nudging can be either epistemically harmful or epistemically beneficial. If the nudger nudges the nudgee to believe falsely, this will be epistemically harmful to the nudgee. But doxastic nudging can also be used to improve the epistemic situation of the nudgee. Suppose Judith is not properly responsive to relevant reasons. This may be so because she suffers from confirmation bias, believes only what her social identity requires her to believe, or falls prey to some other cognitive bias. In cases like these, giving further reasons is not an effective way of influencing Judith’s beliefs. But nudging might still be a way of improving Judith’s epistemic situation. Suppose that benevolent Arthur enters the scene. He knows the relevant truths about the world and also realizes that Judith does not believe these truths, even though her evidence sufficiently supports believing them. From Arthur’s point of view, it is not a promising strategy to give Judith further reasons to believe what is true. She seems to be unresponsive to such reasons. But he can nudge her to believe what is true and what is also supported by her evidence. Arthur may be able to counterbalance the irrelevant influences on Judith’s beliefs with other non-rational influences, in such a way that she ends up believing what is true and what is supported by her evidence.[[3]](#endnote-2)

True beliefs and beliefs that are propositionally justified, i.e., beliefs for which the believer possesses sufficient supportive evidence, are epistemic goods for the person who acquires these beliefs. Doxastic nudging that results in such beliefs is an act of epistemic libertarian paternalism. However, the epistemic goods that this nudging can bring about seem to be severely limited. There are two arguments that suggest such a limitation. First, this kind of nudging cannot generate justified beliefs if justified beliefs must be *based on* the reasons that support them. In the case of brute doxastic nudging, the resulting beliefs do not seem to be based on supporting reasons but rather on epistemically irrelevant factors, i.e., the triggers of the operative biases. Second, doxastic nudging does not seem to generate knowledge either. In order to support this claim, one has to assume that safety is at least necessary for knowledge.[[4]](#endnote-3) Accordingly, a true proposition is not known unless it is believed on the basis of a method that could not easily have resulted in a false belief. In the case of doxastic nudging, the relevant method is the operative bias that is triggered in such a way that it results in a true belief. But this does not mean that the same operative mechanism could not have easily resulted in a false belief. In the case of biases, the production of false beliefs seems to be an easy possibility. Hence such cases of doxastic nudging seem to be inapt to generate knowledge. Let us define *epistemic nudging* (in the strict sense) as doxastic nudging that results in justified beliefs or knowledge. Given this background, epistemic nudging seems to be out of reach for brute nudges.[[5]](#endnote-4)

*2. Is nudging typically a way of giving reasons?*

To sum up the results of the previous section: since brute doxastic nudging utilizes cognitive biases to influence beliefs, it seems generally inapt to produce justified beliefs or knowledge in the nudgee. Recently, Neil Levy has offered a revisionary account of nudging, according to which brute nudges do not trigger irrational cognitive mechanisms, but instead constitute genuine reasons of a special kind, which are effective by being fed into the nudgee’s reasoning mechanism. If this were true, then nudging would not bypass reasoning, and thus might be apt to generate justified beliefs or even knowledge in the nudgee (Levy 2019). In this section, I critically engage with Levy’s account, since I want to demonstrate that the bulk of brute nudges does not constitute reasons. In these cases, nudging is a smart way of using biases to induce beliefs. In the course of the paper, I will argue that even these instances of manipulative nudges can generate justified beliefs and knowledge in the nudgee.

Before taking a closer look at the details of Levy’s account, let me address a fundamental worry. Isn’t our definition of brute doxastic nudging incompatible with the idea that nudging is giving reasons?[[6]](#endnote-5) Not necessarily. According to the definition, brute nudges trigger “shallow cognitive mechanisms.” However, these mechanisms could be shallow and reasoning mechanisms at the same time. A shallow reasoning mechanism would then be one that does not involve the conscious recognition of reasons or any reflective deliberation process, but instead operates automatically, quickly, and unreflectively, like Kahneman’s system 1.[[7]](#endnote-6) Levy wants to insist that such a system can still be reason-responsive, just as perception and intuition are. So, it seems true that the definition of (brute) nudging does not by itself disallow that nudging is giving reasons. Here is Levy’s way of putting it:

Nudging doesn’t bypass our capacity to reason. When they are effective in changing behavior, manipulations of the context of choice typically […] work by giving us reasons. These reasons may not be consciously recognized or responded to by agents, but they are reasons nevertheless, and it is in virtue of being reasons that they alter behavior. The mechanisms that respond to nudges are reasoning mechanisms, and in most cases, at least, nudges no more bypass reasoning than do philosophical arguments. (Levy 2019: 282-83)

Why does Levy think that nudges can be reasons for decisions (or beliefs)? Levy presents two examples to illustrate this point. In the first example, the nudger makes a certain choice (a proposition) salient, or she sets it as a default option. If the nudger is benevolent, then setting the default or making something salient is “functioning as implicit testimony” (Levy 2019: 291). The nudger thus gives a trustworthy recommendation to the nudgee; and this recommendation is higher-order evidence for the choice being the right one (or the proposition being true). That is to say, in a good case, the nudge is a reason to make the recommended choice or believe in the proposition that was made salient. In his second example, Levy engages with corrections of false reports in the news (Levy 2017). Somewhat surprisingly, news consumers often do not revise their beliefs in accordance with public corrections. On the contrary, they often increase their confidence in the original propositions. This phenomenon is called the “backfire effect.” Under specific conditions, however, news corrections will be believed by consumers. Corrections are effective in this sense if they are made by partisan news outlets or if consumers recognize that the corrections are made by a source that does not take any special interest in the corrections being made. A case in point would be if Fox News were to report that the birther stories about Obama were fabricated. This suggests an effective nudging strategy: always make sure that corrections are published by those news outlets that are not interested in disclosing those corrections. Levy’s main point is now that these nudges are not only effective, but also constitute testimonial reasons for believing that the correction is true. This is so because testifying against one’s own interest can only be motivated by having overwhelming evidence for the testified proposition. In this case, the testimony is evidence of evidence for the asserted proposition, and thus constitutes a reason for the nudgee to believe what is testified.

Now, Levy not only claims that nudges can be reasons to decide (or believe) what the nudgee is nudged to decide (or believe); he also claims that to be nudged is to be properly responsive to these reasons (Levy 2019: 290 and *passim*). The latter claim involves the claim that when one is nudged, the nudge triggers a cognitive mechanism that has the *proper function* of appropriately responding to reasons like the nudge. Applying these general claims to Levy’s first example leads to the following result: considerations that are made salient to us (or have default status) are interpreted as testimonial recommendations; and it is clearly rational to follow testimonial recommendations. In the second example, testifying against one’s own interest is interpreted as indicating strong evidential support; and it is rational to follow assertions that are most likely strongly supported by the evidence. Of course, Levy is not arguing that these are explicit inferences that the nudgee performs when she is nudged. Rather, he believes that these cognitive mechanisms are intuitive, non-reflective processes whose operations are not transparent to the agent. But like many other heuristics, they have an evolved proper function of responding appropriately to the relevant reasons. Levy takes sides with Gerd Gigerenzer in claiming that cognitive biases are simply the result of actually misfiring reasoning processes (Gigerenzer 2015).

Levy’s account suffers from two substantial flaws. First, it overintellectualizes some of the addressed instances of nudging. Second, it is not applicable to core instances of brute nudging where reason-responsive mechanisms are clearly not involved. Let me address these worries in turn. First, the worry of over-intellectualization. Levy claims, e.g., that setting the default can be understood as implicitly recommending the default choice. However, even if the default setting is, in fact, evidence of such a recommendation, the default often has its effect on the nudgee without making any use of this evidential value. Default options are often effective because of inertia. It simply takes too much effort to change the default, so the nudgee simply goes along with it. In such a case, the explanation of accepting the status quo has nothing to do with the evidential value of the default setting.

Second, Levy addresses a few (brute) nudges that can be analyzed as testimonial reasons. It is true that he does not claim that every instance of nudging can be interpreted as a response to (genuine or apparent) reasons (Levy 2019: 282-283). Nevertheless, he seems to suggest that the bulk of brute nudging can be understood in terms of reason-responsiveness. However, once we consider brute nudging more broadly, it becomes obvious that the majority of instances lack (apparent or genuine) reasons or do not operate on reason-responsive mechanisms. Let me illustrate this with three examples.

*Social reference.* One way of nudging people into a certain kind of behavior is to inform them that this behavior is executed by the majority of their peers. The behavior of one’s peer group is not a reason to adopt such behavior, and conforming to this behavior is certainly not reason-responsive (Sunstein 2016: 92-93).

*Loss frame.* Since we have a loss aversion, framing certain choices as involving a loss will nudge people to stay away from this option. However, loss aversion is not rational. It is not more valuable to avoid certain losses than it is to gain the same amount. So, when nudges trigger the mechanism of loss aversion, they do not activate a reason-responsive mechanism (Sunstein 2016: 91).

*Affect heuristic*. When cognitive assessment is hard, we often (unconsciously) use our emotional reaction as a heuristic of the assessment. Suppose you have to vote, and there are a number of politicians to choose from. If you do not know much about the candidates, you will vote for the politician you like the most (Sunstein 2016: 94). A hard question (who is the best candidate?) is substituted by an easy question (who do you like the most?). It seems clear that this mechanism involves neither reasons nor reason-responsive mechanisms.

It has been found that brute nudging cannot generally be analyzed as triggering reason-responsive mechanisms in the nudgee. Characterizing it as smart use of non-rational or even manipulative mechanisms fits much better. Given this characterization, the fundamental challenge to epistemic nudging emerges particularly clearly. It is hard to see how nudges that trigger non-rationally biased processes in the nudgee are supposed to result in justified beliefs or knowledge.

*3. How epistemic nudging is possible after all*

Let us take stock. Brute doxastic nudging influences the nudgee’s beliefs without using reflective first-order reasons. Levy’s idea that, in general, brute nudging is a way of feeding further testimonial reasons into an automatic reason-responsive mechanism is unconvincing. In this section, I will argue that systematic nudging can compensate the nudgee’s lack of relevant reason-responsiveness. In other words, I will argue that systematic nudging can, in the right conditions, generate justified beliefs and knowledge without any rational agency on the part of the nudgee. If this is true, then epistemic nudging may work even if nudging cannot be understood as giving reasons to the nudgee. Consider the following case:

Political Loyality. John is the charismatic leader of a political party. In the context of his involvement in illegal activities, he commits a murder. Although John does not confess to his deed, his guilt is beyond reasonable doubt, and he is thus sentenced to life imprisonment. However, other party members are not responsive to the overwhelming evidence of his guilt. Firmly in the grip of loyalty, John’s charisma and confirmation bias, they do not trust the court’s decision. Alicia, the court’s public relations manager, does not want to leave it at that. She wonders how she might persuade the party members of John’s guilt. Finally, she comes up with an idea. Why not fight their cognitive biases with smartly deployed further biases? In particular, she frames the whole story of John’s trial in such a way that he appears unsympathetic and thoroughly suspicious. The official pictures of the trial are selected to make John appear unflattering, unsympathetic – even ugly. In the end, her strategy is successful. Most of the party members end up believing that John is guilty. When one of them, Mia, believes that John committed the murder, could we say that she knows this?

Let us analyze the epistemological situation in more detail. Alicia makes Mia truly believe that John committed the crime. Moreover, Alicia’s intervention is based on her knowledge that John is guilty. However, Mia’s belief is not based on the evidence of John’s guilt. Mia does possess such evidence but, as stipulated in the case description, Mia is not doxastically responsive to this evidence. Mia’s belief is also not virtuously formed, but instead results from her biases. The mechanism operative here is an affect heuristic (Kahneman 2011: 97-104). In order to answer a hard question (is John guilty?) Mia relies on her answer to an easy question (do I like John?). In this case, Mia’s negative feelings about John trigger her assessment concerning John’s guilt. If knowledge requires either the epistemic agent’s competent performance (as virtue theories of knowledge claim), or being based on adequate evidence (as evidentialist views have it), then Mia lacks knowledge in this case. But there are alternative views. My focus is on safety accounts of knowledge. For the sake of simplicity, I will assume that knowledge is identical to safe belief. A very promising account is method-safety. Accordingly, S’s belief that p is safe if and only if it is produced by a method M, such that all beliefs that are produced by M in the actual and relevantly nearby possible worlds are true (see Williamson 2000, Pritchard 2009, Grundmann 2020).[[8]](#endnote-7)

Is Mia’s belief that John committed the crime based on a method that produces only true beliefs in the actual and relevantly nearby possible worlds? The verdict crucially depends on the individuation of the operative method. Suppose the relevant method is internally individuated as the agent’s cognitive process responsive to the framing of reported contents. When contents are framed, credibility magnets are attached to the disclosed contents, which either attract or repulse belief in those contents. Being responsive to such content-unrelated amplifiers or attenuators of credibility cannot generally be safe, because credibility magnets will be attached not only to true but also to false contents, e.g., when people want to deceive their audience. Framing is used to benefit, but also to harm, people epistemically. Therefore, being guided by framing effects is neither safe nor reliable. So, if the relevant method is individuated internally, nudging cannot lead to knowledge in the nudgee.

There is, however, some reason to believe that belief-forming methods should be individuated externally, i.e., partly in terms of the epistemic agent’s environment (proponents of externally individuated methods are Williamson 2000, Goldberg 2010, Broncano-Berrocal 2014 and Grundmann 2020). The method can be externally individuated because justifying features need not be luminous (Williamson 2000); externally individuated methods help to explain why in hard cases the relevant method is safe after all (Broncano-Berrocal 2014, Grundmann 2020); and externally individuated methods underpin a reasonable anti-individualistic view of testimonial knowledge (Goldberg 2010). Moreover, consider the following version of *Russell’s Clock*:[[9]](#endnote-8)

Due to his perfectly attuned inner clock, Russell comes down the stairs every morning at exactly 8:15 a.m. Having arrived there, Russell relies upon reading his old grandfather clock. What Russell does not know is that there is a demon watching him who likes to play foolish tricks. So, on even days of the month, the demon does not interfere with the properly working clock. But on uneven days, the demon stops the clock at exactly 8:15 a.m. and reactivates it later. Intuitively, Russell knows that it’s 8:15 a.m. upon reading the clock on even days, but he does not know that it’s 8:15 a.m. upon reading the clock on uneven days.[[10]](#endnote-9)

Why does Russell know on even days but not know on uneven days? Obviously, he has a true belief about the time on both days. If the relevant method were individuated internally (e.g., forming one’s beliefs from one’s reading of the clock), it would be unsafe, and Russell thus wouldn’t possess knowledge on any day. This is because Russell’s reading of the grandfather clock might easily have led him to a false belief (if he had read the clock at other times on uneven days). But this does not agree with our intuitions. The only difference that can explain why Russell knows on even days but does not know on uneven days is that the consulted clock works properly on even days but is unreliable on uneven days. Now, properly working clocks implement safe methods, whereas unreliable clocks are unsafe. Notice, however, that the relevant difference is located in the external clock itself. So, we can explain the intuitive difference only if we treat the relevant methods as externally individuated.

Let us suppose that the relevant methods are externally individuated. In the case of nudging, the relevant method will then be something like *forming beliefs when steered by a particular nudger* rather than *being doxastically responsive to nudges no matter what*.[[11]](#endnote-10) Whereas the latter method cannot be safe, the former will be safe under specific conditions. To produce knowledge in the nudgee, the nudger must have a robust disposition to nudge the nudgee into holding true beliefs. If the nudger were to flip a coin in order to decide whether to nudge the nudgee into holding a true or a false belief, the method would quite easily have led to a false belief and hence failed to produce knowledge. Typically, this requires (i) that the nudger is knowledgeable about her domain of nudging,[[12]](#endnote-11) (ii) that she is epistemically benevolent towards the nudgee, and (iii) that her nudging is systematic and effective. If the nudger were ignorant rather than knowledgeable, she would not reliably steer the nudgee to true beliefs. If the nudger were not epistemically benevolent towards the nudgee, she might want to nudge the nudgee towards false beliefs. And if her nudging were unsystematic and ineffective, the nudger would only occasionally bring about true beliefs in the nudgee. This does not suffice for safe belief.

In Political Loyality, Alicia satisfies all three conditions: she knows that John is guilty, she wants to inform the party members about this fact, she is effective in her nudging and uses her nudging systematically in such a way that the group is safely steered to true beliefs about the murder. So, epistemic nudging is possible even in cases of manipulative nudges that bypass the nudgee’s reasoning altogether. However, keep in mind that this conclusion can only be derived under two further conditions. First, we must assume that safe belief is identical to knowledge, i.e., no further conditions (such as being based on evidence or being the product of virtuous epistemic agency) are required for knowledge.[[13]](#endnote-12) Second, we must also assume that the method that has to be assessed as safe is externally individuated in the right way. Since neither condition is uncontroversial, the result of my argument had better be presented as a conditional: *if* safe belief is identical to knowledge, and *if* the relevant method is externally individuated, then epistemic nudging is possible.

If brute nudging to knowledge is, under specific conditions, possible, then brute nudging to justified beliefs is possible as well. Of course, brute nudging to justified beliefs is not compatible with every account of justified belief. If a belief must be based, e.g., on sufficient evidence or on agential virtues for it to count as justified, then nudging to justified beliefs is out of reach. If, however, a (prima facie) justified belief requires nothing beyond being based on a reliable belief-forming method, then nudging to justified beliefs is, under specific further conditions, possible. As in the case of knowledge, this requires externally individuated methods of belief-formation and a reliable counterfactual correlation between nudged beliefs and the truth.

*4. Conclusion*

Brute doxastic nudging is possible. One can steer agents’ beliefs by smartly triggering their doxastic biases, even without their consent. When agents are biased against believing the truth or what is indicated by their evidence, one can use doxastic nudging to make them believe what is true or propositionally justified. It is still not entirely clear how effective specific nudges are.[[14]](#endnote-13) Nor do we have a complete list of effective mechanisms for doxastic nudging. What is clear is that such a list should include framing effects (using, e.g., risk and loss frames), social referencing, and affect heuristics. In this paper, I have argued that nudgers can achieve an even more ambitious epistemic goal: They can brutely nudge other people even to knowledge and doxastically justified beliefs. This is possible only under specific meta-epistemological assumptions, namely that the relevant methods are externally individuated, and that knowledge, or justified belief, requires nothing but a modal correlation with truth (e.g., safety or reliability). If this much is granted, then brute nudging to knowledge or justified belief is no longer miraculous.

It is a further and completely independent question whether we *should* use nudging to provide others with knowledge or justified beliefs. This practice might interfere with the value of the nudgee’s personal or epistemic autonomy (for discussion, see Ahlstrom-Vij 2013, Keeling 2018, Levy 2017). Moreover, the epistemically benevolent nudger misses the chance to make agents more resistant to manipulation by malign nudgers (Meehan 2020) or to the influence of cognitive biases. It might be preferable to turn the agent’s epistemic vices into virtues, rather than fighting her vices with other vices behind her back. Still, nudging might be an attractive option if education fails. I must leave answering the normative question for another occasion.

**References**

Ahlstrom-Vij, K. 2013. *Epistemic Paternalism: A Defence.* Basingstoke: Palgrave Macmillan.

Alston, W. 1989. “The Deontological Conception of Epistemic Justification.” In: Alston, W. (ed.) *Epistemic Justification: Essays in the Theory of Knowledge,* Ithaca and London: 115-152.

Broncano-Berrocal, F. 2014. “Is Safety in Danger?” *Philosophia* 42: 63-81.

Bullock, E. C. 2016. “Knowing and Not-Knowing for Your Own Good: The Limits of Epistemic Paternalism.” *Journal of Applied Philosophy* 35: 433-447.

Comesaña, J. 2005: “Unsafe Knowledge.” *Synthese* 146: 395–404.

Gigerenzer, G. 2015. “On the Supposed Evidence for Libertarian Paternalism.” *Review of Philosophy and Psychology* 6: 361-383.

Goldberg, S. 2010. *Relying on Others.* Oxford: Oxford University Press.

Goldberg, S. 2015: “Epistemic Entitlement and Luck.” *Philosophy and Phenomenological Research* 91: 273–302.

Goldman, A. 1991. “Epistemic Paternalism: Communication Control in Law and Society.” *The Journal of Philosophy* 88: 113-131.

Goldman, A. 1999. *Knowledge in a Social World*. Oxford: Oxford University Press.

Grundmann, T. 2020. “Saving Safety from Counterexamples.” *Synthese* 197: 5161-5185.

Hiller, A. and R. Neta. 2007: “Safety and Epistemic Luck.” *Synthese* 158: 303–313.

Hummel, D. and A. Maedche. 2019. “How Effective is Nudging? A Quantitative Review on the Effect Size and Limits of Empirical Nudging Studies.” *Journal of Behavioral and Experimental Economics* 80: 57-58.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Penguin.

Keeling, G. 2018. “Autonomy, Nudging and Post-Truth Politics.” *Journal of Medical Ethics* 44: 721-722.

Kelp, C. 2009. “Knowledge and Safety.” *Journal of Philosophical Research* 34: 21-31.

Lackey, J. 2006: “Pritchard’s Epistemic Luck.” *Philosophical Quarterly* 56: 284–9.

Lackey, J. 2007: “Why we don’t deserve credit for everything we know.” *Synthese* 158: 345-362.

Lackey, J. 2008. *Learning From Words: Testimony as a Source of Knowledge*, Oxford: Oxford University Press.

Levy, N. 2017. “Nudges in a Post-Truth World.” *Journal of Medical Ethics* 43: 495-500.

Levy, N. 2019. “Nudge, Nudge, Wink, Wink: Nudging is Giving Reasons.” *Ergo* 6: 281-302.

Mackey, R. 2015. “Brutal Images of Syrian Boy Drowned Off Turkey Must Be Seen, Activists Say.” *The New York Times. Open Source* (Sept. 2, 2015). https://www.nytimes.com/2015/09/03/world/middleeast/brutal-images-of-syrian-boy-drowned-off-turkey-must-be-seen-activists-say.html. Accessed: Oct. 5, 2020.

Meehan, D. 2020. “Epistemic Vices and Epistemic Nudging: A Solution?” In: Amiel, B. and G. Axtell, (eds.), *Epistemic Paternalism: Conceptions, Justifications and Implications*, Rowman and Littlefield: 247-259.

Nisbett, R. 2015. *Mindware: Tools for Smart Thinking.* Penguin Books.

Pritchard, D. 2009. “Safety-Based Epistemology: Whither Now?” *Journal of Philosophical Research* 34: 33-45.

Saghai, Y. 2013. “Salvaging the Concept of Nudge.” *Journal of Medical Ethics* 39: 487-493.

Sunstein, C. R. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge University Press.

Thaler, R. H. & C. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth and Happiness*. Yale University Press.

Williamson, T. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

1. **Acknowledgements.** An earlier version of this paper was presented at the CONCEPT Epistemology Brown Bag (Cologne) in October 2020. I am grateful for the extensive and extremely helpful discussions with the audience. Moreover, I received written comments to a draft version of the paper from the following colleagues: Anna-Maria Eder, Eve Kitsik, Jakob Lohmar and Jasper Lohmar. Their valuable feedback enabled me to work out an improved version of this paper. I would also like to thank two anonymous reviewers for their substantial criticism that helped me to avoid significant mistakes. Finally, my thanks go to the guest editors of this special issue on post truth for including my contribution: Filippo Ferrari, Anna Maria Lorusso, Sebastiano Maruzzi and Giorgio Volpe. No potential competing interest was reported by the author. [↑](#footnote-ref-1)
2. Sunstein (2016: 88) defines manipulative nudges as not sufficiently engaging with or appealing to people’s capacity for rational choice. He claims that some but not all nudges are manipulative in this sense (ibid., 114). [↑](#endnote-ref-1)
3. There are other, less substantial interventions that might also help Judith. Her attention might be directed to the relevant evidence, so as to make her reason-responsive, or someone might prevent her from being exposed to evidence that would predictably trigger certain biases in her (see Goldman 1991 or Ahlstrom-Vij 2013). Although at least the latter kind of intervention is controversial (for criticism, see, i.e., Bullock 2016), it does not bypass Judith’s rational deliberation. [↑](#endnote-ref-2)
4. Here I rely on a notion of knowledge that is more substantial than mere true belief. For the latter notion of “weak knowledge”, see Goldman 1999: 23-26. [↑](#endnote-ref-3)
5. Let me point out here already that both arguments rely on premises that can (and, later in this paper, will) be disputed. First, one does not have to accept that justified beliefs must be based on appropriate (genuine or apparent) reasons possessed by the epistemic agent. In particular, reliabilism drops this assumption. Second, we need not individuate the relevant belief-forming mechanism internally. However, if we consider the nudgee’s bias to be merely a part of some socially extended process, this larger process might be safe even if it includes a bias as one of its parts. [↑](#endnote-ref-4)
6. Levy does not consider the real possibility that nudges are sometimes targeted at system 2 and therefore do not always trigger shallow cognitive processes. However, we can fix this blind spot by pretending that his account is concerned with *brute* nudges only. Let me emphasize that Levy himself does not use this term. [↑](#endnote-ref-5)
7. Whereas, according to Kahneman, system 1 operates automatically, quickly and without voluntary control, system 2 involves reasoning processes that operate consciously and reflectively. See Kahneman 2011: 20-24. [↑](#endnote-ref-6)
8. That being based on safe methods is necessary and sufficient for knowledge is highly contested in the literature. Many epistemologists currently believe that safety is not sufficient for knowledge, since, in addition, a true belief must be creditable to the epistemically virtuous agent for it to count as knowledge. See, e.g., Pritchard 2015. But there are also critics of this extra virtue-requirement for knowledge. See, e.g., Lackey 2007 and Grundmann 2020. Others object to the claim that safety is necessary for knowledge. See, e.g., Comesaña 2005, Lackey 2006, Hiller & Neta 2007, Coffman 2010, Goldberg 2015. But even this criticism has not remained unchallenged. See Grundmann 2020. The latter is one of the rare proponents of safety as necessary and sufficient for knowledge. [↑](#endnote-ref-7)
9. Although this case is inspired by Kelp 2009: 27-28, it is substantially different from his version. Whereas in Kelp’s case, Russell could easily have formed a false belief, in my case, he could easily have acquired a true belief based on an unreliable device. [↑](#endnote-ref-8)
10. I have to admit that the general verdict about this case may be more ambivalent than expected. I am grateful to an anonymous referee who made me aware that my verdict about the case is contested. [↑](#endnote-ref-9)
11. That this is the correct individuation of the relevant method is, of course, not fully determined by the fact that some methods are externally individuated. An externally individuated method may not involve the particular nudger. This is why a full defense of this typing would require further argument. [↑](#endnote-ref-10)
12. Cases such as Lackey’s Creationist Teacher (Lackey 2008: 48) suggest that knowledge on the part of the testifier is not necessarily required in order to generate knowledge in the hearer. For this reason, the qualification “typically” is used above. [↑](#endnote-ref-11)
13. Worry: Isn’t the assumption that knowledge is identical to safe belief hopeless, because knowledge is defeasible by misleading evidence? In this way it could be argued that a belief based on a safe method does not constitute knowledge if the agent possesses evidence against the believed proposition, or against the safe formation of that belief. Reply: This is a serious worry. But it can be answered by adding one further condition. Epistemic nudging is possible only if the nudgee possesses no (undefeated) evidence that defeats the belief that she is nudged into. [↑](#endnote-ref-12)
14. A recent meta-analysis by Hummel & Maedche 2019 argues that defaults and warnings have the largest effect size. [↑](#endnote-ref-13)