Journal of Academic Ethics
https://doi.org/10.1007/s10805-024-09568-x

Danger of Slippery Slopes in Nudge Research

Helena Siipi[1]

Accepted: 16 September 2024
© The Author(s) 2024

Abstract
Nudges are a way to steer people's behavior through changes in how choices are presented. Nudge research has been incorporated into public policy in many countries, and nudge research, thus, has the potential to directly influence societies and individuals. As a result, research ethics for nudge research is needed to ensure that nudges developed are not instances of unethical manipulation of people. In this paper, I argue that two types of slippery slopes from ethically fine nudges to ethically problematic ones can take place in nudge research. The conceptual slippery slope follows from (1) the broad way of defining nudges, (2) the multitude of different ways of understanding manipulation of people, (3) many manipulation definitions implying that some nudges are manipulative, and (4) some forms of manipulation of people being *prima facie* wrong. The empirical slippery slope rests, first, on the possibility of small steps from ethically acceptable to wrong nudges, and second, on the cognitive limits typical to human beings (including nudge researchers). Both slippery slope arguments imply the need to of nudge researchers to create so-called firebreaks. The firebreaks can take the form of ethical justifications. To avoid slippery slopes, certain types of arguments should be excluded from these justifications.

Keywords Nudge · Slippery slope · Manipulation · Cognitive limit · Firebreak · Justification

Background: Nudges and Objections

In addition to individuals' mental states, their outer circumstances influence their choices. Among other things, the choice architecture–the way the different alternatives are presented to individuals–has an impact on their behavior. Intentionally arranging the choice architecture to promote a certain choice is commonly referred as nudging.

✉ Helena Siipi
helsii@utu.fi

[1] University Lecturer, Philosophy, FI 20014 University of Turku, Turku, Finland

Published online: 10 October 2024

One of the most commonly repeated examples of nudging concerns the arrangement of food choices in a cafeteria (for the original example see Thaler & Sunstein, 2009). Healthy eating can be promoted by putting vegetables and low-fat foods at eye level at the beginning of the line and less healthy alternatives further back on small serving dishes (Ensaff, 2020). Additionally size of the plates influences the amount of food taken (Kallbekken & Sælen, 2013). These arrangements of choice architecture enhance healthy consumption without changing prices or excluding any dishes from the menu.

People are often quite ready to accept nudges (Sunstein et al., 2019; Congiu & Moscati, 2021). Moreover, many nudges are relatively easy to arrange and cost-effective (Congiu & Moscati, 2021; Benartzi et al., 2017; for critical views see e.g., Ridder et al., 2020; Lades & Delaney, 2022). As a result, they are subject to keen research activities. Research groups around the world are developing nudges to be used as policy instruments for various fields of life including, but not limited to, retirement saving, active commuting, organ donation, saving electricity, charitable giving, decisions regarding privacy and safe driving.

At the same time, nudges remain controversial. While very few want to forbid all nudges, most agree that some nudges are ethically better than others and that some nudges can be ethically problematic.[1] As a result, there is active academic discussion regarding ethically relevant features of nudges (see e.g., Siipi & Koi, 2022; Lades & Delaney, 2022; Kuyer & Gordijn, 2023), and various frameworks have been developed for evaluating the ethical aspects of particular nudges (see e.g., Hansen & Jespersen, 2013; Clavien, 2018; Engelen, 2019).

To date, research ethics for nudge research has not been extensively discussed. This may be due to various general guidelines for research on human beings (such as ethical the standards of the Office of Human Research Protections of the U.S. Department of Health and Human Services), which also apply to nudge research. The aim of this paper is to spell out a danger of two slippery slopes that can take place in nudge research–especially in nudge research with the aim of developing new types of nudges. Both slippery slopes are research ethically noteworthy, as they may lead nudge researchers to develop ethically problematic nudges (or ways of influence that come close to nudging). The current research ethical guidelines and practices fail to capture these slippery slopes; thus, they imply the need to nudge researchers to create and use so-called firebreaks.

Slippery slope arguments are not foreign to the ethics of nudging (see e.g., Rizzo & Whitman, 2009; Frischmann, 2021; Hansen, 2016; Ivanković & Engelen, 2017). The "fathers" of nudge thinking Richard H. Thaler and Cass R. Sunstein (2009) acknowledge them already in their classic *Nudge: Improving Decisions About Health*, *Wealth*, *and Happiness*. However, most of this discussion has concentrated on slippery slopes from nudges to other types of influences based on coercion, punishments and bans. The slippery slopes I am discussing take place within the context of nudging and nearby. I aim to show that in nudge research (for new types of nudges) there is a risk of sliding from ethically fine nudges to nudges (and ways of influence that come close to nudging) that are ethically undesirable. Even though most nudge scholars, as noted above, agree that some nudges may be ethically problematic, this possibility has not yet been discussed much (however, see Schubert, 2017).

Nudging as a policy instrument will be discussed in the section "*What are nudges?"*. The basic idea of slippery slope arguments is presented in the following section. The fourth

[1] Even the developers of nudge thinking, Thaler and Sunstein (2009) acknowledge that "[s]ome nudges are bad or just unwelcome".

section focuses on conceptual slippery slopes. It is argued that nudging conceptions of the second section overlap with some central understandings of the 'manipulation of people'. It is followed by the section in which I spell out empirical slippery slopes regarding nudging. In the final section, the ethical implications of these arguments for nudge research are discussed. This section will include recommendations concerning firebreaks.

What are Nudges?

Original Nudge Description and Its Three Components

Nudges come in a variety of different ways of influencing human behavior. Thaler and Sunstein (2009) describe nudges as follows:

> A nudge is any aspect of 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.

This description consists of three components (Congiu & Moscati, 2021). The first two components are negative in excluding some influences from the sphere of nudges (Hansen, 2016). First, nudges do not forbid any options or reduce the choice sets of the nudgees (Saghai, 2013; Zimmermann, 2023).[2] This requirement sets nudges apart from coercion, and because of it, nudges are often described as freedom retaining (Hansen & Jespersen, 2013; Siipi & Koi, 2022) or allowing their targets to go their own way (Sunstein, 2015). Thinking back to the cafeteria example, nudging does not exclude unhealthy alternatives from the menu; they are just made to appear less tempting. Moreover, coming to the second negative component of nudging, the unhealthy alternatives are not made less attractive by pricing. Nudges are not (dis)incentives (Hilton et al., 2018). (Dis)incentives, such as high prices for environmentally problematic products (e.g., gasoline) and low prices on the environment friendly ones (e.g., public transport), can efficiently guide behavior, but they are not nudges. Even though the original definition mentions only economic (dis)incentives, nudges are often understood not to significantly change (dis)incentives regarding time or trouble. Moreover, they do not create substantial social, physical or psychological sanctions (Hansen, 2016; Hausman & Welch, 2010; Zimmermann, 2023).

The third component of the original description describes an inclusion criterion: nudges are easy and cheap to avoid. In other words, nudges do not hinder individuals from choosing what they prefer. Those who want to eat a salami pizza can easily avoid the cafeteria nudge and eat according to their preference. This third requirement implies that a nudge can effectively influence only someone who does not have a strong preference between the alternatives or someone who, for one reason or another, does not engage in a deliberation regarding the choice between them (Congiu & Moscati, 2021). Thus, in practice, the third requirement means that a person who has a preference against the nudge can easily resist it (Saghai, 2013; Nys & Engelen, 2017).

[2] The term 'nudgee' refers to the targets of the nudge, that is the individuals and groups who are nudged toward certain kind of behavior. The term 'nudger' refers to those who develop nudges and do the nudging.

Clarifications and Additions to the Original Description

Thaler and Sunstein (2009) describe nudge to be "any factor that significantly alters the behavior of humans, even though it would be ignored by Econs". By 'Econ' they refer to *homo economicus*–a being that is motivated to bring about greatest benefit and makes no mistakes in calculating the outcomes of different alternatives. Humans are not Econs. We are weak willed, and cognitive boundaries, biases, implicit attitudes and flaws compromise our ability to perform rationally (Hansen, 2016; Nosek, 2007). Sometimes this dependence on cognitive boundaries is taken to be the defining feature of nudges:

> They [nudges] are called for because of flaws in individual decision-making, and work by making use of those flaws (Hausman & Welch, 2010).

> A nudge is a function of (I) any attempt at influencing people's judgment, choice or behaviour in a predictable way (1) that is made possible because of cognitive boundaries, biases, routines and habits in individual and social decision-making posing barriers for people to perform rationally in their own declared self-interests and which (2) works by making use of those boundaries, biases, routines, and habits as integral parts of such attempts (Hansen, 2016; see also Engelen & Nys, 2020).

This paper does not rest on the view that all nudges work through cognitive limits and flaws in human thinking. Rather, it is sufficient to note that some nudges utilize them.

The second clarification concerns the intentionality of nudges. It is often noted (sometimes as a defense for nudging) that choice architecture is everywhere and that it may have an impact on human behavior even when it is not intended (Thaler & Sunstein, 2009; Levy, 2017; Sunstein, 2015; Zimmermann, 2023). The arrangement of a cafeteria, for example, influences customer choices also when the cafeteria manager randomly, without further reflection, places the food offers on display. In this paper, accidental and haphazard influences are not taken as nudges. Following various scholars, only choice architectures that are intentionally created to guide behavior in a certain direction are understood as nudges (Hansen, 2016; Hansen & Jespersen, 2013; Hausman & Welch 2010; Zimmermann, 2023).

Sometimes the criteria for an influence to be a nudge are further enriched by requirements regarding the nature of the intentions of nudgers. According to Thaler and Sunstein (2009):

> A policy is paternalistic if it dries to influence choices in a way that will make choosers better off, *as judged by themselves*. […] Individuals make pretty bad decisions–decisions they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control.

This addition limits the use of nudges to situations where the nudgers and nudgees agree on the desirability of the goal of nudging. According to a strong interpretation of this addition, changing plates of the cafeteria into smaller ones would fail to be a (ethically good) nudge if a climate action critical person followed it. It is easy to see that many common nudges are not according to this strong interpretation (Siipi & Koi, 2022). Moreover, many scholars are ready to accept as policy instruments also other-regarding nudges – that is nudges that

do not directly benefit the nudgee (M'hamdi et al., 2017; Sharif & Moorlock, 2018; Conqui & Moscati, 2021). Thus, it is best to interpret the addition weakly as follows: Nudges do not (or *prima facie* should not) (a) mislead a person to act against their deeply held values and beliefs (Siipi & Koi, 2022) or (b) aim to increase the nudger's well-being exclusively (Congui & Moscati, 2021).

The three components of nudges as well as the additions and clarifications of this section will be later utilized to form a conceptual slippery slope.

Slipery Slope Arguments

Slippery slope arguments are an instance of consequentialist arguments. According to them, some course of action *A1* is very likely to lead into another course of action *An*. Even though the course of action *A1* is morally neutral as such, the course of action *An* is morally bad. Thus, either *A1* should be avoided or a firebreak that hinders *An* resulting from *A1* should be built. As Fumagalli (2020) puts it,

> SSAs argue against a proposed action, practice or policy *A1* on the alleged ground that allowing or implementing *A1* will (likely) lead to some unacceptable (i.e., morally impermissible or otherwise objectionable) action, practice or policy *An*, typically through a number of intermediate stages *A2*, …, *An-1*.

Slippery slope arguments have sometimes been presented to be fallacies (see e.g., Spielthenner, 2010). However, they are an argument type that includes instances of both strong and weak arguments (Jefferson, 2014; Whitman, 1994; Fumagalli, 2020). Thus, evaluation of the slippery slope arguments should be done case by case.

Slippery slope arguments consist of the following components: (a) the course of action under interest *A1*; (b) the course of action, which is claimed to be ethically bad *An*; (c) the claim that *A1* will lead to *An*; and (d) the conclusion that A1 should (*pro tanto*) not be carried out (Fumagalli, 2020). The strength of a slippery slope argument rests on whether the moral evaluation regarding *An* is taken to be correct and on the justification given to the claim of component c. The slippery slope arguments can be divided into different types that differ with respect to the interpretation of component c– that is, how *A1* is seen to lead into *An.*

The so-called empirical slippery slope arguments see *A1* to lead into *An* through social, legal, political and psychological mechanisms such as habituation or changing values (Jefferson, 2014; Evans, 2020). Getting used to *A1* makes step *A2* seem reasonable, which makes *A3* seem acceptable and so on (Jefferson, 2014). For example, slippery slope arguments against euthanasia often rest on the claim that people will eventually become accustomed to the idea of medical doctors killing patients. This will slowly lead into accepting new groups of patients being euthanized until "extension of permission for euthanasia for competent and terminally ill patients with unbearable suffering to incompetent or vulnerable people" (Terkamo-Moisio, 2016). The strength of an empirical slippery slope argument rests on prediction regarding the mechanisms of the component c: How likely is it that the described changes take place? Typically, empirical slippery slope arguments do not describe inevitabilities but rather (high) probabilities of unacceptable practices or policies (Rizzo & Whitman, 2009). Thus, they can be answered–not only by forbidding *A1*–but also by build-

ing firebreaks, in other words, by changing circumstances (e.g., social arrangements) so that the connection between *A1* and *An* breaks down (Buchanan et al., 2000; Evans, 2020).

The so-called logical slippery slope arguments take the connection between *A1* and *An* to rest on consistency. According to them, if one accepts *A1* one must logically also accept *An* (Fumagalli, 2020). The point of the logical slippery slope arguments is that (even though we may first fail to notice this) *A1* and *An* actually belong to the same class (Whitman, 1994). Thus, taking the course of action *A1* commits us to accept that the course of action *An* can also be taken. In other words, if we allow *A1*, there is no good reason not to allow *An* because they are similar with respect to their morally relevant properties (Jefferson, 2014).

Conceptual slippery slope arguments are instances of logical slippery slope arguments (Whitman, 1994). They rest on the idea that *A1* and *A2* and, eventually, *An* cannot be conceptually distinguished from each other. From *A1* to *An*, each step down the slippery slope is the result of the lack of a sharp line between the cases. This may be due to so-called "continuity vagueness", where the cases from *A1* to *An* are on continuous measure (Evans, 2020; Rizzo & Whitman, 2009). "Similarity vagueness" as a source of slippery slope refers to cases in which measurement is impossible or at least very imprecise (Evans, 2020). For example, measurements of infantilization (which is sometimes claimed to result from nudging; see Kuyer & Gordjin, 2023) may be prone to suffer from similarity vagueness.

All slippery slope arguments rest on the idea that once one takes the first step *A1*, one cannot stop sliding down the slippery slope until *An*. The counterarguments for slippery slope arguments typically aim to show that the slide does not exist or that it can be avoided by building a firebreak (Buchanan et al., 2000; Evans, 2020). For conceptual slippery slope arguments both options can follow from pointing out relevant conceptual differences between *A1* and *An* (Evans, 2020). The vagueness of boundaries or inability to distinguish the cutoff point is often insufficient for showing that *A1* and *An* belong to the same class. This is famously the case, for example, in moral discussions regarding the status of the human fetus.

If building a firebreak is possible, it should often be favored over forbidding *A1*. This is especially the case when *A1*, as such, is ethically desirable (e.g., in having potential to bring about health or environmental benefits). Thus, the "need for firebreak" interpretation of slippery slope arguments is especially suitable in the contexts of new technologies, social arrangements, policies or behavioral insights such as nudging. In what follows, I present a slippery slope argument for the moral necessity of building firebreaks in nudge research. In short, I claim that nudging has the potential to lead–through both conceptual and empirical slippery slopes–the nudge researchers to accept morally problematic behavioral insight policies. Thus, nudge researchers should, as part of their nudge development work, build firebreaks. With appropriate firebreaks, it is possible to simultaneously stop sliding down the slippery slope and retain the possible benefits following from ethically fine nudges.

Conceptual Slippery Slopes, Nudges and Manipulation

Conceptual Slippery Slope Argument

In what follows, I present a conceptual slippery slope argument from ethically acceptable forms of nudging to nudges (and ways of influence that come close to nudging) that are

instances of ethically *prima facie* wrong manipulation of people. The slippery slope follows from the combination of (1) the broad way of defining nudges[3], (2) the multitude of different ways of understanding manipulation of people[4], (3) many manipulation definitions implying that at least some nudges are instances of manipulation, and (4) agreement that some forms of manipulation of people are *prima facie* wrong. The argument is taken to imply that nudge researchers should invest in creating firebreaks. To put it short, the definitions commonly given to nudges and manipulation overlap, which implies the possibility of a conceptual slippery slope from ethically fine nudges to ethically undesirable manipulation. This creates a research ethical judgment that firebreaks should be developed to stop this slide.

Nudges being a form of manipulation is a common claim (see e.g., Lades & Delaney, 2022; Ridder et al., 2020; Noggle, 2018; Hansen & Jespersen, 2013; Coons & Weber, 2014; Engelen & Nys, 2020). Jongepier and Klenk (2022), for example, state that "[n]udging, at least in some forms, seems to be manipulative", and Gorin (2014) presents the cafeteria nudge as an example of manipulation of people. Whether nudges are manipulative depends both on what is taken to be a nudge and what counts as manipulation. The common descriptions of nudging (presented above) do not explicitly exclude nudges from the sphere of manipulation. At the same time, it seems intuitively hard to accept that nudges, such as a default to print on both sizes of paper (Egebark & Ekström, 2016) or a sticker prompt at the lids of rubbish bins reminding of proper recycling of waste (Shearer et al., 2017), were manipulative. However, this is not to say that nudges are never manipulative. In order to find out whether some nudges are instances of *prima facie* wrong forms of manipulation and whether one should, therefore, worry about conceptual slippery slopes in nudge research, one needs to spell out the criteria for an influence to be an instance of manipulation. Thus, a closer look at manipulation conceptions is needed.

Limits for Defining Manipulation

Few clarifications are necessary before a closer look at different manipulation concepts. First, things other than people–for example currency, variables in research experiments and sport competitions–can also be manipulated (Barnhill, 2014). On some wide interpretations of the term, not only people but also technologies or other entities can be manipulative (Belohrad, 2019). In this paper, the focus is on the manipulation of human beings done by human agents–either by individuals or by groups. Second, for some interference to be manipulative, it must be intentional in the sense that the agent aims to influence the behavior of someone else. However, it is not necessary that the manipulator has an intention to manipulate. The manipulator may well be unaware that their way of influencing is an instance of manipulation (Belohrad, 2019; Pepp et al., 2022). Third, in this paper, the use of the term 'manipulation' is limited to altering the behavior and choices of other humans. However, in its broader scientific sense, it may also include techniques and tasks that allow researchers to assess the subconscious and hidden mental processes of other people (Nosek,

[3] Also, Hansen (2016) argues that the concept of nudge is so loosely defined that it can lead into slippery slopes.

[4] In regard the manipulation of variables in research experiments, the term 'manipulation' is certainly well-defined in many the research ethical guidelines such as the ethical standards of the Office of Human Research Protections of the U.S. Department of Health and Human Services. However, in this paper I discuss the manipulative nature of research outputs (the nudges and nudge-like policies). The term 'manipulation' then refers to psychological manipulation of people's choices. In that context the precise definition is missing.

2007). These techniques can sometimes even be taken to be paradigmatic cases of manipulation. Fourth, manipulation is often associated with the type of influences that aim to benefit the manipulator. However, manipulation can also be paternalistically motivated or aim to benefit a third party (Noggle, 2018; Barnhill, 2014; Jongepier & Klenk, 2022). All these clarifications are compatible with the idea that some nudges are manipulative.

Defining manipulation is a complicated task. Manipulation can happen in numerous different ways. Even though there are clear cases–influences that are generally agreed to be instances or noninstances of manipulation–there are also borderline cases. Practically all manipulation definitions are judged to be either under-inclusive or over-inclusive (or both) by some experts. Under-inclusive concepts fail to encompass some instances of manipulation, whereas over-inclusive concepts also cover ways of influence that are not manipulative (Barnhill, 2014). As long as intuitions regarding what should be taken as manipulation differ, we lack clear criteria for over- and under-inclusiveness. The concepts of manipulation to be discussed in what follows are not perfect. (If we already had a perfect conception discussion on the matter would not be necessary.) However, they are influential in the philosophical literature on manipulation as well as useful for the task in hand–that is, for showing the danger of conceptual slippery slopes in nudge research.

Manipulation as Pressure

In what follows I, following Robert Noggle's (2022) categorization, discuss three different types of manipulation concepts and determine what they imply with respect to the question of whether some nudges are manipulative. The aim is to show that some manipulation definitions overlap with those given to nudges and that this implies the possibility of the types of conceptual slippery slopes presented above.

Like nudging, manipulation is also a way of influencing the behavior of others. The goal of manipulation concepts is to identify features that set manipulation apart from other influences–typically from rational persuasion and coercion (Coons & Weber, 2014; Blumenthal-Barby & Burroughs, 2012; Todd, 2013). Coercion can occur either by eliminating some alternatives or by penalizing them (Wood, 2014). For example, a road can be closed from traffic either by concrete barriers or by placing penalties for driving on it. Both force people to change their behavior. Rational persuasion consists of giving people good reasons for their choices. For example, people might be persuaded to give up driving on a road by providing them with facts about the negative climate and health effects of driving.

Manipulation is sometimes seen as pressure that falls between rational persuasion and coercion. Coercion is an extreme pressure and usually very hard (if not impossible) to resist, whereas rational persuasion at the other end of the continuum forms no pressure (Wood, 2014; Noggle, 2018). As Noggle (2022) puts it:

> It seems plausible […] to suppose that there is a continuum between rational persuasion and coercion with regard to the level of pressure being exerted, with rational persuasion exerting no pressure, coercion exerting maximum pressure, and the middle region, manipulation, exerting pressure that falls short of being coercive. In this way, we might arrive at the idea that manipulation is a form of pressure that does not rise to the level of coercion.

According to this *pressure view*, manipulation differs from coercion in degree. Manipulation does not exclude options or penalize them. A manipulated individual can choose differently. Yet, manipulation can be an effective way of getting someone to act in a certain way (Feinberg, 1989; Jongepier & Klenk, 2022; Blumenthal-Barby & Burroughs, 2012). According to Wood (2014), manipulation operates "by making other options less attractive without absolutely removing them or making them inacceptable".

Wood's description of manipulation comes close to the original description of nudging: nudging is all about rearranging the choice architecture so that some options appear more attractive and others less attractive. None of the components of nudge description or clarifications (presented in the second section of this paper) set all nudges apart from manipulation understood in this way. Regarding the latter, the pressure view of manipulation is compatible with pressure being intentional, having an effect on the behavior of humans but not of Econs, utilizing flaws in decision-making and making individuals behave in ways that are (in the broad sense of the term) judged good by themselves. The three components of nudge description–nudges as freedom retaining, not being (dis)incentives and being easy and cheap to avoid–set some forms of manipulation apart from nudging. When manipulation comes close to coercion, it fails to fulfill the third (and under some readings also the first) component. Yet, nudges can still be understood as forms of manipulation that are closer to the end of rational persuasion (Noggle, 2018).

The pressure view is problematic in being far too over-inclusive. For example, offering a good salary from an easy job or singing praises for children who learn multiplications do not seem to be instances of manipulation, even though they fall somewhere between rational persuasion and coercion. At its best, the pressure view can offer one (but not the only) necessary condition for an influence to be manipulative. As a result, even though many nudges may fall within the scope of this definition, this is not enough to show that nudging is manipulative. Thus, it also fails to support the slippery slope argument from fine nudges to undesirable manipulation.

Manipulation as Bypassing Reason

Manipulation is sometimes characterized as bypassing its target's rational deliberation (Noggle, 2022; Blumenthal-Barby & Burroughs, 2012; Jongepier & Klenk, 2022; Gorin, 2014). This idea is tempting since, as noted in the last section, manipulation is often contrasted with rational persuasion. Since rational persuasion influences behavior by engaging rational capacities, manipulation must influence its target's behavior in some other way. The connection to nudging is easy to see. When understood in this *bypassing reason* way, manipulation, similarly to nudging, does not influence Econs even though it alters the behavior of humans (Nys & Engelen, 2017).

A paradigm case for this understanding of manipulation is subliminal advertising–or a folk view of subliminal advertising–which (if it worked as sometimes is presented) influences human behavior through unconscious mechanisms (Noggle, 2022). The folk view of subliminal advertising, however, fails to be an instance of nudging as it does not fulfill the criteria of being easy to avoid. Thus, the crucial question becomes the following: Are there instances of nudging that are also instances of manipulation understood this way? The answer depends on how manipulation is taken to bypass reason.

Sometimes, the claim that manipulation bypasses reason is understood to imply that manipulation is a form of *hidden or covert influence*. According to this line of thought, manipulation takes place when A intentionally influences B's choice without B noticing this or at least without B understanding how A influences B's choice (Todd, 2013; Barnhill, 2022). This kind of understanding of manipulation would imply that some–but certainly not all–nudges are instances of manipulation. Some nudges are transparent with respect to both of their goals and motives, whereas others are hard to notice (Ivanković & Engelen, 2019). The latter are sometimes taken as instances of manipulation:

> This sort of view [of manipulation] […] would also appear to apply to other sorts of psychological influences that bypass conscious attention, such as priming effects and the decision-making biases and other processes that involve unconscious, fast processing made famous by the work of Daniel Kahneman, Amos Tversky, and colleagues (Noggle, 2022).

The hidden or covert influence view of manipulation seems underinclusive. There are several clear cases of manipulation, such as guilt trips and emotional blackmail, which can be easily noticed by their targets (Barnhill, 2022). However, regarding nudging and the conceptual slippery slope argument I am defending the relevant question is whether the view is also overinclusive. Can an intentional opaque influence fail to be an instance of manipulation?

If an intentional hidden influence is always manipulative, then the cafeteria nudge is (unless the intention to influence is revealed to the customers) an instance of manipulation. So is any other nudge which is opaque with respect to its goals and motives. This seems harsh. Full transparency is not usually seen necessary for the ethical acceptability of nudges. Rather, it has been seen to require so-called *in principle token interference transparency*, meaning that a watchful individual could, without unreasonable effort, notice that nudging is taking place (Ivanković & Engelen, 2019; Bovens, 2009; Schimdt, 2017). Thus, this understanding of manipulation could be met by claiming that nudges that fail to meet *in principle token interference transparency* criteria are manipulative. This view could nicely serve the conceptual slippery slope argument under interest.

The idea of "bypassing reason" can also be interpreted as intentional influence which involves *misleading the target about the true motives* behind it. In such cases, the intention to influence is noticeable to the target. Manipulation takes place if and only if the target is intentionally caused to have false beliefs regarding the motives of the influencer–that is, if manipulator intentionally creates an impression that their motives are different from what they actually are (Belohrad, 2019). A manipulation takes place, for example, when someone gives an impression of promoting the well-being of the target even though their true motive is to benefit the influencer's own economy.

All nudges do not create false impressions about the nudgers' motives, but some nudges do. Transparency regarding the intention to nudge does not imply transparency regarding why one wants to nudge or what the aims of nudging are (Ivanković & Engelen, 2017). Some nudges are not just silent about nudgers' motives but actually mislead nudgees regarding them. For example, environmentally motivated nudgers have created nudges that appear to the nudgees as ones benefiting their health, economy or social relations (see e.g., Filippini et al., 2021; Riggs, 2017). This is not uncommon. Rather, constantly seeking the most effi-

cient motivators is an integral part of nudge research. Classifying these nudges as manipulative could serve the conceptual slippery slope argument under interest.

Manipulation as Trickery

According to manipulation as *trickery view* manipulation is closer to deception than to coercion. Manipulation contains an element of leading astray (Belohrad, 2019). As Mills (1995) notes,

> a manipulator tries to change another's beliefs and desires by offering her bad reasons, disguised as good, or faulty arguments, disguised as sound–where the manipulator himself knows these to be bad reasons and faulty arguments.

Manipulation as *trickery view* differs from the *misleading the target about the true motives* view in allowing the misleading to concern wide variety of different kinds of beliefs as well as other mental states, such as desires and emotions (Noggle, 2018; Belohrad, 2019). The idea of the trickery view is that there are certain norms regarding how we should come to have different mental states. Manipulation happens when someone intentionally brings it about that someone else comes to have mental states in ways that are not according to these norms (Barnhill, 2022). According to Noggle (2018), manipulation is any "deliberate attempt to induce–trick–a person into relying on a faulty mental state in her deliberation".

This trickery view allows manipulation to occur in many different ways. False beliefs can be produced not only by lying but also by intentionally leaving some relevant facts out. Similarly, when emotions and desires are influenced in a way that leads to emotions and desires that the influencer sees as inappropriate in that particular situation, then we are talking about manipulation (Barnhill, 2022). Desires fail to be ideal if they do not conform to an agent's beliefs about what they has reason to do. Ideal emotions are based on true beliefs and have adequate intensity (Belohrad, 2019).

The trickery view implies that some salience nudges are instances of manipulation. They are not manipulative when they bring the salience of some fact into closer alignment with its actual importance. However, when they make the salience of some fact more distant from its actual importance, they are instances of manipulation. In manipulative nudges, either a non-important issue is made to influence a choice to a high degree or an important matter is blocked from the decision-making process (Noggle, 2018).

Noggle (2018) also states that the cafeteria nudge is manipulative in this sense. The rearrangement of the options does not make the nudgees pay attention to issues that should matter to their decision-making. It merely makes certain dishes easier to choose; the rearrangement does not point into anything that makes the favored option more choice-worthy. If the cafeteria nudge is an instance of manipulation, then are other nudges similar to it manipulative as well.

Manipulative Nudges – so What?

The above presented manipulation views imply that some (but not all) nudges are manipulative. This partly follows from the broad way of understanding nudges. When nudges are described in a way that covers a great variety of different kinds of influences, some manipu-

lative influences are likely to be included. The manipulation claims do not, however, concern only some non-typical instances of nudging. Quite contrary, some manipulation views imply that even the most paradigmatic nudges–like the cafeteria nudge–are manipulative (Noggle, 2018; Gorin, 2014).

What is crucial to the conceptual slippery slope argument under interest is that the original description of nudging and some central manipulation conceptions overlap. Some ways of influencing human behavior are instances both of nudging and of manipulation. As a result, a clear distinction cannot be drawn between nudging and manipulation. Thus, there is a possibility for conceptual slippery slope from nudging to manipulation.

The slippery slope arguments are meant to be normative. Showing that some nudges are manipulative does not, as such, imply that they are ethically wrong. The conceptual distinctions made above are silent regarding the moral status of manipulation. A convincing slippery slope argument requires introducing ethics to the picture. This can be done either by presenting an argument for the (*prima facie*) ethical wrongness of manipulation or by showing 'manipulation' to be a so-called thick concept. In other words, the distinction between ethically desirable manipulation and non-desirable manipulation needs to be spelled out and conceptualized.

The *pressure view* as well as *the hidden or covert influence view* (the first one of the bypassing reason views) can be understood as morally neutral manipulation views. They do not imply that manipulation is ethically wrong. Neither do they imply it being morally all right (Jongepier & Klenk, 2022). Thus, they cannot alone–without additional arguments for the (*prima facie*) moral wrongness of manipulation–support the conceptual slippery slope argument.

Thick manipulation conceptions take moral status to be a defining feature of manipulation (Jongepier & Klenk, 2022; Todd, 2013). The *misleading the target about the true motives* view (second of the bypassing reason views) and the *trickery* view are thick manipulation concepts. Both take manipulation to involve deception, which implies that it is not morally all right–at least without a justifying excuse (see e.g., Belohrad, 2019; Noggle, 1996). Thick manipulation conceptions do not exclude the possibility that many (or in practice even all) instances of manipulation are morally permissible. This could be the case if (a) manipulation is not absolutely but only *prima facie* wrong and (b) many (or all) actual cases of manipulation are accompanied by good excuses–other countervailing moral factors that justify the manipulative action in question.

The view that manipulation is always morally wrong is hard to accept. Obvious counterexamples for this view are cases in which manipulation of one person has extremely good consequences, such as saving thousands of lives (Noggle, 2022). At the same time, it seems that many clear cases of manipulation–such as Iago's actions in Shakespeare's Othello[5]–contain morally dubious elements (Todd, 2013; Nys & Engelen, 2017). A common solution is to judge that manipulation is *prima facie* wrong (see e.g., Nys & Engelen, 2017; Belohrad, 2019; Mills, 1995). The *prima facie* understanding acknowledges that even though the manipulative nature of an action gives a reason not to do it, there may also be other stronger moral reasons that override the *prima facie* rule not to manipulate. Thus, even

[5] Iago, without straightforwardly lying, makes Othello to believe that Othello's wife Desdemona has been unfaithful. Iago's insinuations make Othello to take as evidence some minor issues (such as a misplaced handkerchief) he would have normally paid no attention to. The outcome is tragedy: Othello kills Desdemona.

though manipulation is always *prima facie* wrong, many instances of manipulation may still be *all things considered* acceptable or even the morally best things to do. In some cases, the moral harm from manipulation may be quite minor, at least in comparison to its benefits. As Noggle (2018) writes:

> Carolyn's cafeteria nudge is a relatively minor form of manipulation that is directed toward a benevolent end. It thus stands in stark contrast to Iago's manipulation of Othello, which is serious, malevolent, and ultimately tragic. But the cafeteria nudge is trickery nonetheless. All things considered, we might think that it is justified or at least excusable because it is a minor and well-meaning form of trickery. Moreover, if Carolyn's cafeteria is in a school, or hospital, or health club, we might argue that we should expect encouragement to make healthy choices in those settings, even if some of that encouragement is manipulative. But even if we are inclined to give Carolyn a pass on this basis, we should keep in mind that we are giving her a pass on something that is a form of trickery […]. The fact that her nudge employs trickery constitutes a strong case for thinking that it is manipulative.

The justifications given to particular manipulative actions can be stronger or weaker. Typically, as in the above quote regarding the cafeteria nudge, the justifications refer to the good consequences of manipulation. The acceptability of a manipulative act, however, does not depend merely on its good consequences but also on the nature of the manipulative act (the type of influence) itself, its possible risks or bad consequences, the target group, the manipulator's status and the context of action (Belohrad, 2019; Noggle, 2018; Jongepier & Klenk, 2022). As a result, manipulative acts form a continuum with respect to how good justifications can be given to them. Some manipulations, such as the cafeteria nudge, have strong justifications, whereas others, such as Iago's acts, lack justifications. Others fall somewhere between them. Importantly to the slippery slope argument, as a result of the continuity vagueness (Evans, 2020; Rizzo & Whitman, 2009) we lack the sharp line between "strong enough" and "not strong enough" justifications for manipulation.[6]

Taken together the conceptual overlap of 'nudge' and 'manipulation', the *prima facie* wrongness of manipulation (inherent in some manipulation conceptions), the continuum of the strength of manipulation justifications and the vagueness of that continuum form a slippery slope argument from innocent nudges to ethically wrong manipulation. In short, nudging cannot be conceptually distinguished from manipulation and a sharp line between acceptable and non-acceptable forms of manipulation cannot be drawn. Thus, a conceptual slippery slope from ethically acceptable forms of nudging to ethically objectionable forms of manipulation is possible. Nudge researchers should pay attention to this possibility and as part of their work develop firebreaks. Firebreaks are discussed at the end of this paper.

[6] There may also be similarity vagueness when evaluation of the degree of justification is imprecise (Evans, 2020).

Empirical Slippery Slopes

The empirical slippery slopes from ethically fine nudging to ethically wrong manipulation rest on nudge researchers' cognitive limits–limits they share with all human beings (but not with Econs). As Nosek (2007) notes, "[m]ental process and mental experience are not the same thing. The former is the operation of the mind; the latter is the subjective life that emerges from that operation". All operations of the mind are not conscious to the subject, which makes them vulnerable to cognitive mistakes. Following the empirical slippery slope model presented above (Jeffersson 2014; Evans, 2020), when researchers develop nudges, they become accustomed to the practice of intentionally influencing the behavior of others. Through habituation, the *A1* type of influencing starts to feel usual, normal, natural and–importantly–free of ethical questions. Once researchers become accustomed to the *A1* type of nudging, the *A2* type of nudges do not seem to constitute a large step, as the difference between *A1* and *A2* is small. After some time, researchers become accustomed to *A2* type nudges, and as a result, a small step to *A3* seems appropriate. Eventually, researchers are developing *An* type of nudges, which they, at the beginning of the process, would not have approved and which (according to the presenter of the slippery slope argument) are ethically problematic (Rizzo & Whitman, 2009).[7]

This empirical slippery slope can be summarized as follows:

(1) Nudge researchers have similar flaws in thinking as other humans do.
(2) Nudges can in nudge research be developed by small steps, meaning that many successive minor changes eventually accumulate into a large difference.
(3) Because of certain flaws typical to human thinking and because of the possibility of small steps in nudge research, there is a risk of empirical slippery slope from ethically fine nudges to ethically undesirable ones.
(4) Thus, firebreaks should be developed to hinder this slippery slope.

The success of empirical slippery slope arguments rests on the likelihood of the claimed social, legal, political or psychological mechanisms taking place. In what follows, I argue that certain common biases in human thinking are prone to lead nudge researchers to the kind of slippery slope just described.

First, so-called *status quo bias* refers to a common human tendency toward an irrational preference of current states of affairs (Bostrom & Ord, 2006; Ölander & Thøgersen, 2014). Status quo bias implies a reluctance to see moral problems in how things are currently carried out, and it also relates to seeing non-action (not making changes) ethically prior to altering the current state of affairs (Hofmann, 2020; Bostrom & Ord, 2006). Status quo bias concerns not only behavior but also justifications given to it. People tend to verbalize customary entities as natural and normal and, thus, morally unproblematic (Mill, 1969; Harris, 1985; Siipi, 2008; Hofmann, 2020). This kind of moral reasoning is, of course, mistaken. Being customary does not imply moral acceptability (Bostrom & Ord, 2009). The status quo bias is intimately related to habituation. As such, it means an unwillingness to

[7] The empirical slippery slope may also concern the justifications given to manipulative nudges. The habituation, then does not concern the nudge itself, but what is taken as a sufficient justification for carrying out *prima facie* wrong manipulative nudges. Step by step smaller and smaller as well as less certain benefits are taken to justify manipulation.

make changes. However, it also feeds to the empirical slippery slopes in which changes are enabled by making very small steps from *A1* to *A2* and finally through various small steps to *An* (instead of making one large change).

Second, *conformity bias* supports empirical slippery slope arguments from nudging to ethically wrong manipulation. Conformity refers to interpersonal influence. People tend to accustom their behavior on the basis of what they think others are doing or expecting. In other words, an individual's awareness of social norms and knowledge about choices made by others is likely to influence their behavior. We also follow others and predict their expectations without being aware of it. Conformity bias refers to the tendency of individuals to quite blindly follow the behavior or views of others even when they should question the practices (Lazaric et al., 2020). Conformity and conformity bias are often utilized in nudging (see e.g., Demarque et al., 2015; Ölander & Thøgersen, 2014; Engelen & Nys, 2020; Thaler & Sunstein, 2009). However, they influence not only nudgees but also nudgers. Nudge researchers typically work in environments in which people are accustomed to different ways of influencing the behavior of others and have positive hopes regarding nudging–after all, the goal of their work is to develop efficient nudges. As a result of this working environment and conformity bias, nudge researchers may fail to question ethics of what they are doing.

Finally, nudge researchers' success in their work depends on their ability to create new effective nudges. It is in their interest to see the nudges developed by them as ethically fine ones. Thus, they are prone to fall into *confirmation bias* regarding them. Confirmation bias refers to a human tendency to favor and more easily accept information that supports their previously existing beliefs and views (Peters, 2022). Moreover, all people have a strong *tendency to view themselves as ethical and competent*. This prevents them from recognizing their own biases (Sah, 2017). In regard to professionals, they tend to think that their "professionalism" makes them immune to being influenced by conflicts of interest (such as one between their success and following ethical requirements). However, in practice people are quite poor in remaining objective–even when they are motivated to be impartial (Sah, 2017). Finally, even without conflicts of interest, nudge researchers may be overconfident regarding their ability to identify ethically sensitive issues in nudging. *Optimism and overconfidence regarding one's own abilities* are also common biases (Thaler & Sunstein, 2009).

In addition to biases, (other) subconscious mental states and processes may lead researchers into slippery slopes. People's attitudes can be divided into implicit and explicit attitudes. The implicit attitudes are subconscious and they can sometimes be contrary to person's explicit (conscious) attitudes. The implicit attitudes, just like biases of thinking, influence our behavior without us noticing it (Nosek, 2007). In the context of nudge researchers, the implicit attitudes might influence, for example, choices regarding targets of nudges (less well-off or privileged groups) and what are seen as suitable forms of nudging for different groups. As an example, implicit attitudes of nudge researchers might well lead them to apply different standards of transparency to nudges for different groups.

Taken together, these flaws in human thinking make the empirical slippery slope from ethically fine nudging to less acceptable ones possible and somewhat likely. According to Rizzo and Whitman (2009), the presented kind of slippery slopes are not uncommon. They state that Thaler and Sunstein are in their *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2009) actually presenting a continuum (and maybe also a slippery slope) from intuitively acceptable nudges to more questionable ones. *Nudge* begins with the

cafeteria nudge and through various small steps ends into cases that, from the point of view of ethics, compromise a lot. As Rizzo and Whitman (2009) put it:

> [W]e see that the leading new paternalists themselves believe that soft and hard paternalism can be connected by a series of small steps. [They] present public and private, and coercive and non-coercive, paternalistic activities alongside each other with little or no recognition of when they are crossing the line from one to the other.

Firebreaks

If the slippery slope arguments presented above are convincing, then what? If one is not willing to give up all nudging, then the arguments imply the necessity of developing firebreaks in nudge research (for similar views see Rizzo & Whitman, 2009). Firebreaks refer to procedures and practices that stop one from sliding down the slippery slope–that is they work against unintentional development of ethically problematic nudges (or influences close to it). Firebreaks can take several forms. Some of them can block both conceptual and empirical slippery slopes whereas others can stop only certain specific slippery slopes.

In theory, the conceptual slippery slopes regarding nudge research could be tackled by defining 'nudge' and 'manipulation' in ways that exclude nudges from the sphere of (*prima facie* wrong forms of) manipulation. However, given the wide academic discussion regarding both terms, consensus regarding the matter seems unlikely (Barnhill, 2022). There is no single commonly accepted definition for 'nudge'[8], and all manipulation definitions presented this far have been found either over- or underinclusive by some academics (Barnhill, 2014). It is not advisable for nudge researchers to adopt one manipulation conception and evaluate their nudges on the basis of it. That a nudge fails to be manipulative under one conception tells little about its ethical acceptability and is unlikely to form a sufficient firebreak.

According to Barnhill (2022), real and possible manipulation charges could be utilized for pinpointing ethically problematic nudges. Charges of manipulation can be understood analogously to the so-called "yuk factors" discussed by Mary Midgley (2000) as "moral traffic lights". The manipulation charges then work as indicators of the need for ethical analysis before moving on. In other words, when people complain that a nudge is an instance of manipulation, researchers should stop and ask what they are objecting to. Some manipulation charges may turn out to be normatively weak or misplaced. However, others may succeed in pointing out morally relevant factors. To distinguish between the two, reasons for calling an influence manipulative should be spelled out. Is it because the nudge is covert? Does it include false claims, or is it misleading? Is it based on targeting psychological weaknesses (Barnhill, 2022)?

As presented with respect to conceptual slippery slopes, even a manipulative nudge may–just like Carolyn's cafeteria–turn out to be ethically all right if one can give a strong justification for it. Thus, in addition to analyzing the manipulation charges, building firebreaks consists of offering justifications for particular nudges–that forming arguments that show that the nudge in question is all things considered good even though it is *prima facie* wrong in being manipulative.

[8] For different understandings see e.g., Congiu & Moscati, 2021; Hausman & Welch 2010; Hansen, 2016.

In the case of empirical slippery slopes, the morally problematic feature of an influence *An* may have nothing to do with manipulation. However, in their case forming a justification may also serve as a firebreak. A good justification is based on a thorough ethical analysis. Various ethical tools and frameworks have been developed and suggested (e.g., Meske & Amojo, 2020; Hansen & Jespersen, 2013; Lades & Delaney, 2022; Clavien, 2018), but there is no consensus regarding which method of ethical evaluation is best. Most researchers consider the ethics of nudging to be multidimensional, meaning that the moral status of a nudge does not depend on any single factor. Rather, ethical evaluation of nudges consists of weighting and balancing various ethically relevant factors against one other (Siipi & Koi, 2022; Lades & Delaney, 2022). As a result, a nudge that is not ethically excellent with respect to one factor may still be *all things considered* ethically acceptable or even the best alternative. Conversely, a nudge being ethically good in one respect does not imply that it is ethically acceptable overall. In practice, that a new nudge *An* is less transparent than *A1* and *A2*, does not imply nudge *An* to be all things considered ethically worse than *A1* and *A2*. *An* may also be far more just than *A1* and *A2* and guide people efficiently to actions that are the best among the choices available.

To avoid slippery slopes, certain common justification methods should be excluded from firebreak building. First, to avoid the small steps typical of both empirical and conceptual slippery slopes, the justifications should not take the form of analogy arguments. Analogy arguments would justify an influence by referring to its great similarity to some other (already accepted) nudge.[9] Omitting analogical reasoning is also useful for avoiding *status quo* bias.

Second, even though studies on social acceptability may be important for the overall evaluation of nudging, they may–especially when a high degree of social acceptability is implied–also build into conformity bias. Thus, the results regarding social acceptability (as well as results regarding expert views) should, in the evaluation of particular nudges, be used only for negative purposes–that is, for implying the need for further ethical consideration. Finally, to avoid confirmation bias and biases based on overconfidence, the ethical analysis should be done systematically and openly.

There probably is no single way of using the ethical analysis for the development of firebreaks that can serve all research groups equally well. What is important is that they take the danger of slippery slopes seriously and choose and use methods that suit their and their target populations' needs for building firebreaks.

Conclusions

Some nudges are ethically better than others. While some nudges are ethically unproblematic, others can be ethically unacceptable. In nudge research there is a danger of both empirical and conceptual slippery slopes from ethically fine nudges to ethically problematic ones.

The presented slippery slope arguments do not need to be objections to all nudging. However, they highlight the need to develop firebreaks in nudge research to hinder sliding from the development of ethically fine nudges to ethically problematic ways of influencing human behavior. Nudge researchers can develop firebreaks by providing justifications for nudges under development. These justifications can be developed with the help of different

[9] For more discussion on analogy arguments see Spielthenner, 2014.

kinds of ethical tools and frameworks from which the argument types that are likely to support slippery slopes are excluded.

Acknowledgements This contribution was written in the context of a research project that received funding from the Strategic Research Council at the Academy of Finland (grant number 335186). I want to thank the anonymous reviewers from their valuable comments regarding the earlier version of this article.

Funding This contribution was written in the context of a research project that received funding from the Strategic Research Council at the Academy of Finland (grant number 335186).
Open Access funding provided by University of Turku (including Turku University Central Hospital).

Declarations

Conflict of Interest I declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Barnhill, A. (2014). What is Manipulation? In C. Coons, & M. Weber (Eds.), *Manipulation: Theory and practice*. Oxford University Press.

Barnhill, A. (2022). How philosophy might contribute to the practical ethics of online manipulation. In F. Jongepier, & M. Klenk (Eds.), *The Philosophy of Online Manipulation*. Routledge.

Belohrad, R. (2019). The Nature and Moral Status of Manipulation. *Acta Analytica*, *34*(4), 447–462. https://doi.org/10.1007/s12136-019-00407-y

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., & Galing, S. (2017). Should governments Invest more in nudging? *Psychological Science*, *28*(8), 1041–1055. https://doi.org/10.1177/095679761770201

Blumenthal-Barby, J. S., & Burroughs, H. (2012). Seeking Better Health Care Outcomes: The Ethics of Using the "Nudge". *The American Journal of Bioethics* 12(2), 1–10. https://doi.org/10.1080/15265161.2011.634481

Bostrom, N., & Ord, T. (2006). The Reversal Test: Eliminating Status Quo Bias in Applied Ethics *Ethics* 116(4), 656–679. https://doi.org/10.1086/505233

Bovens, L. (2009). The Ethics of Nudge. In T. Grüne-Yanoff, & S. O. Hansson (Eds.), *Preference change: Approaches from philosophy, economics and psychology*. Springer.

Clavien, C. (2018). Ethics of nudges: A general framework with a focus on shared preference justifications *Journal of Moral Education* 47(3), 366–382. https://doi.org/10.1080/03057240.2017.1408577

Congiu, L., & Moscati, I. (2021). *A review of nudges*. Definitions, justifications effectiveness *Journal of Economic Surveys*, 1–26. https://doi.org/10.1111/joes.12453

Coons, C., & Weber, M. (2014). Manipulation: Investigating the Core Concept and Its Moral Status. In C. Coons & M. Weber (Eds.) *Manipulation: Theory and Practice*. Oxford Academic.

De Ridder, D., Feitsma, J., Van Den Hoven, M., Kroese, F., Schillemans, T., Verweij, M., Venema, T., Vugts, A., & De Vet, E. (2020). Simple nudges that are not so easy. *Behavioural Public Policy*, 1–19. https://doi.org/10.1017/bpp.2020.36

Demarque, C., Charalambides, L., Hilton, D. J., & Waroquier, L. (2015). Nudging sustainable consumption: The use of descriptive norms to promote a minority behavior in a realistic online shopping environment. *Journal of Environmental Psychology*, *43*, 166–174. https://doi.org/10.1016/j.jenvp.2015.06.008

U.S. Department of Health and Human Services. Office for Human Research Protections (2024). Ethical Codes and Research Standards. https://www.hhs.gov/ohrp/international/ethical-codes-and-research-standards/index.html

Egebark, J., & Ekström, M. (2016). Can indifference make the world greener? *Journal of Environmental Economics and Management*, *76*, 1–13. https://doi.org/10.1016/j.jeem.2015.11.004

Engelen, B. (2019). Ethical criteria for health-promoting nudges: A case-by-case analysis. *The American Journal of Bioethics*, *19*(5), 48–59. https://doi.org/10.1080/15265161.2019.1588411

Engelen, B., & Nys, T. (2020). Nudging and Autonomy: Analyzing and Alleviating the Worries *Review of Philosophy and Psychology* 11, 137–156. https://doi.org/10.1007/s13164-019-00450-z

Ensaff, H. (2021). A nudge in the right direction: The role of food choice architecture in changing populations' diets. *Proceedings of the Nutrition Society*, *80*(2), 195–206. https://doi.org/10.1017/S0029665120007983

Evans, J. H. (2020). New barriers on the Slippery Slope? *The American Journal of Bioethics*, *20*(8), 19–21. https://doi.org/10.1080/15265161.2020.1781961

Feinberg, J. (1989). Harm to self (the Moral limits of the Criminal Law 3). *Oxford University Press*. https://doi.org/10.1093/0195059239.001.0001

Filippini, M., Kumar, N., & Srinivasan, S. (2021). Nudging adoption of electric vehicles: Evidence from an information-based intervention in Nepal. *Transportation Research Part D: Transport and Environment*, *97*, 1–18. https://doi.org/10.1016/j.trd.2021.102951

Frischmann, B. (2021). Nudging humans. *Social Epistemology*, *36*(2), 1–24. https://doi.org/10.1080/02691728.2021.1979121

Fumagalli, R. (2020). Slipping on slippery slope arguments. *Bioethics*, *34*(4), 412–419. https://doi.org/10.1111/bioe.12727

Gorin, M. (2014). Towards a theory of interpersonal manipulation. In C. Christian, & M. Weber (Eds.), *Manipulation: Theory and practice*. Oxford University Press.

Hansen, P. G. (2016). The definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove. *European Journal of Risk Regulation*, *7*(1), 155–174. https://doi.org/10.1017/S1867299X00005468

Hansen, P. G., & Jespersen, A. M. (2013). 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*(1), 3–28. https://doi.org/10.1017/S1867299X00002762

Harris, J. (1985). *The value of life: Introduction to Medical Ethics*. Routledge & Kegan Paul.

Hausman, D. M., & Welsch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, *18*(1), 123–136.

Hilton, D., Treich, N., Lazzara, G., & Tendil, P. (2018). Designing effective nudges that satisfy ethical constraints: The case of environmentally responsible behaviour. *Mind & Society: Cognitive Studies in Economics and Social Sciences*, *17*(3), 27–38. https://doi.org/10.1007/s11299-019-00201-8

Hofmann, B. (2020). Progress bias versus status quo bias in the ethics of emerging science and technology *Bioethics* 34(3), 252–263. https://doi.org/10.1111/bioe.12622

Ivanković, V., & Engelen, B. (2019). Nudging, transparency, and Watchfulness. *Social Theory and Practice*, *45*(1), 43–73. https://doi.org/10.5840/soctheorpract201917

Jefferson, A. (2014). Slippery Slope arguments. *Philosophy Compass*, *9/10*, 672–680. https://doi.org/10.1111/phc3.12161

Jongepier, F., & Klenk, M. (2022). Online manipulation: Charting the field. In F. Jongepier, & M. Klenk (Eds.), *The Philosophy of Online Manipulation*. Routledge.

Kallbekken, S., & Sælen, H. (2013). Nudging hotel guests to reduce food waste as a win–win environmental measure. *Economics Letters*, *119*(3), 325–327. https://doi.org/10.1016/j.econlet.2013.03.019

Kuyer, P., & Gordijn, B. (2023). Nudge in perspective: A systematic literature review on the ethical issues with nudging. *Rationality and Society*, *35*(2), 191–230. https://doi.org/10.1177/10434631231155005

Lades, L. K., & Delaney, L. (2022). Nudge FORGOOD. *Behavioural Public Policy*, *6*(1), 75–94. https://doi.org/10.1017/bpp.2019.53

Lazaric, N., Le Guel, F., Belin, J., Oltra, V., Levaud, S., & Douai, A. (2020). Determinants of sustainable consumption in France: The importance of social influence and environmental values. *Journal of Evolutionary Economics*, *30*, 1337–1366. https://doi.org/10.1007/s00191-019-00654-7

Levy, N. (2017). Nudges in the Post-truth World. *Journal of Medical Ethics*, *43*, 495–500. https://doi.org/10.1136/medethics-2017-104153

M'hamdi, H. I., Hilhorst, M., Steegers, E. A., & de Beaufort, I. (2017). Nudge me, help my baby: On other regarding nudges. *Journal of Medical Ethics*, *43*, 702–706. https://doi.org/10.1136/medethics-2016-103656

Mill, J. S. (1969). Essays on Ethics, Religion and Society. In J. M. Robson (Ed.), *Collected works of John Stuart Mill, part 10*. Toronto University.

Mills, C. (1995). Politics and manipulation. *Social Theory and Practice*, *21*(1), 97–112. https://doi.org/10.5840/soctheorpract199521120
Noggle, R. (1996). Manipulative actions: A conceptual and Moral Analysis. *American Philosophical Quarterly*, *33*(1), 43–55. http://www.jstor.org/stable/20009846
Noggle, R. (2018). Manipulation, salience, and nudges. *Bioethics*, *32*(3), 164–170. https://doi.org/10.1111/bioe.12421
Noggle, R. (2022). The Ethics of Manipulation. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition). https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/
Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, *16*(2), 65–69.
Nys, T. R. V., & Engelen, B. (2017). Judging nudging: Answering the Manipulation Objection. *Political Studies*, *65*(1), 199–214. https://doi.org/10.1177/003231716629487
Ölander, F., & Thøgersen, J. (2014). Informing Versus Nudging in Environmental Policy. *Journal of Consumer Policy*, *37*, 341–356. https://doi.org/10.1007/s10603-014-9256-2
Outcomes The Ethics of using the Nudge. *The American Journal of Bioethics* 12(2), 1–10. https://doi.org/10.1080/15265161.2011.634481
Pepp, J., Sterken, J., McKeever, M., & Michaelson, E. (2022). Manipulative machines. In F. Jongepier, & M. Klenk (Eds.), *The Philosophy of Online Manipulation*. Routledge.
Peters, U. (2022). What is the function of confirmation Bias? *Erkenntnis*, *87*, 1351–1376. https://doi.org/10.1007/s10670-020-00252-1
Riggs, W. (2017). Painting the fence: Social norms as economic incentives to nonautomotive travel behavior. *Travel Behaviour & Society*, *7*, 26–33. https://doi.org/10.1016/j.tbs.2016.11.004
Rizzo, M. J., & Whitman, D. G. (2009). Little brother is watching you: New Paternalism on the Slippery slopes. *Arizona Law Review*, *51*(3), 685–739.
Saghai, Y. (2013). Salvaging the concept of nudge. *Journal of Medical Ethics*, *39*, 487–493. https://doi.org/10.1136/medethics-2012-100727
Sah, S. (2017). Policy solutions to conflict of interest: The value of professional norms. *Behavioural Public Policy*, *1*(2), 177–189.
Schubert, C. (2017). Exploring the (behavioural) political economy of nudging. *Journal of Institutional Economics*, *13*(3), 499–522. https://doi.org/10.1017/S1744137416000448
Sharif, A., & Moorlock, G. (2017). Influencing relatives to respect donor autonomy: Should we nudge families to consent to organ donation? *Bioethics* 32, 155–163. https://doi.org/10.1111/bioe.12420
Shearer, L., Gatersleben, B., Morse, S., Smyth, M., & Hunt, S. (2017). A problem unstuck? Evaluating the effectiveness of sticker prompts for encouraging household food waste recycling behavior. *Waste Management*, *60*, 164–172. https://doi.org/10.1016/j.wasman.2016.09.036
Siipi, H. (2008). Dimensions of naturalness. *Ethics & the Environment*, *13*(1), 71–103. http://www.jstor.org/stable/40339149
Siipi, H., & Koi, P. (2022). The Ethics of Climate nudges: Central Issues for applying Choice Architecture interventions to Climate Policy. *European Journal of Risk Regulation*, *13*(2), 218–235. https://doi.org/10.1017/err.2021.49
Spielthenner, G. (2010). A logical analysis of Slippery Slope arguments. *Health Care Analysis*, *18*, 148–163. https://doi.org/10.1007/s10728-009-0117-0
Spielthenner, G. (2014). Analogical reasoning in Ethics. *Ethical Theory and Moral Practice*, *17*, 861–874. https://www.jstor.org/stable/24478718
Sunstein, C. R. (2015). Nudges, Agency, and abstraction: A reply to critics. *Review of Philosophy and Psychology*, *6*, 511–529. https://doi.org/10.1007/s13164-015-0266-z
Sunstein, C. R., Reisch, L. A., & Kaiser, M. (2019). Trusting nudges? Lessons from an international survey. *Journal of European Public Policy*, *26*(10), 1417–1443. https://doi.org/10.1080/13501763.2018.15319
Terkamo-Moisio, A. (2016). *Complexity of Attitudes Towards Death and Euthanasia*. Publications of the University of Eastern Finland. Dissertations in Health Sciences, 363. http://urn.fi/URN:ISBN:978-952-61-2198-7
Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decision about Health, Wealth, and happiness*. Penguin.
Todd, P. (2013). Manipulation. In H. LaFollette (Ed.), *The International Encyclopedia of Ethics*. Blackwell Publishing. 10.1002/ 9781444367072.wbiee585
Whitman, J. P. (1994). The many guises of the Slippery Slope Argument. *Social Theory and Practice*, *20*(1), 85–97. 130.232.128.99.
Wood, A. W. (2014). Coercion, Manipulation, Exploitation. In C. Coons, & M. Weber (Eds.), *Manipulation: Theory and practice*. Oxford University Press.
Zimmermann, V. (2023). Privacy nudges and informed consent? Challenges for privacy Nudge Design. In N. Gerber, A. Stöver, & K. Marky (Eds.), *Human factors in privacy research*. Springer. https://doi.org/10.1007/978-3-031-28643-8_8

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.