




ARTICLE

Manipulation in politics and public policy

Keith Dowding¹  and Alexandra Oprea²

¹School of Politics and International Relations, Research School of the Social Sciences Building, Australian National University, Acton, Canberra ACT 2601, Australia and ²Department of Philosophy, University at Buffalo, 190 Founders Prom, Amherst, NY, 14068, USA

Corresponding author: Keith Dowding; Email: keith.dowding@anu.edu.au

(Received 20 February 2023; revised 16 November 2023; accepted 27 November 2023; first published online 06 March 2024)

Abstract

Many philosophical accounts of manipulation are blind to the extent to which actual people fall short of the rational ideal, while prominent accounts in political science are under-inclusive. We offer necessary and sufficient conditions – Suitable Reason and Testimonial Honesty – distinguishing manipulative from non-manipulative influence; develop a ‘hypothetical disclosure test’ to measure the *degree* of manipulation; and provide further criteria to assess and compare the morality of manipulation across cases. We discuss multiple examples drawn from politics and from public policy with particular attention to recent debates about the ethics and politics of nudge.

Keywords: Manipulation; nudge; hypothetical disclosure; public policy; persuasion

Introduction

Manipulation in politics comes in many forms. Consider a politician intentionally spreading false information in order to improve her chances of winning an election – perhaps malicious rumours about opposition candidates, fabricated statistics supporting the incumbents’ policy agenda, or claims of electoral fraud to discredit election results. She is directly manipulating voters through deceptive communication. Politicians may also manipulate voters indirectly by engineering favorable situations – gerrymandering political districts through partisan electoral commissions, selectively filibustering to prevent opponents’ proposals progressing, or setting voting rules that guarantee a favourable outcome.

While voters and the media are constantly on the alert for manipulation by politicians and elected officials, concern about manipulation by nonpartisan government agencies is more recent and has largely been caused by the increasingly common commission of experts in psychology and behavioural economics to design ‘nudges’ in public health, environmental policy, tax policy and education.

Some nudges communicate directly with the target of the policy by providing reasons that are evident to the target (hereafter evident reasons). Examples include

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

information about norm compliance (for example, a tax bill reporting that the majority of citizens pay their taxes on time); ratings (for example, 1–5-star Australian food health ratings; Canadian fuel-consumption ratings by vehicle range); reminders (for example, social media app push notifications; text-message appointment reminders); and emotional appeals (for example, public service announcements that use fear to deter drink driving). We call these types of influence *communicative*.

Other nudges influence choice architecture without providing such evident reasons to the target. The best-known examples involve (1) changes to the default rules (for example, automatic retirement-plan enrolment; automatic voter registration; opt-out rather than opt-in policies on organ donation) and (2) changes to the infrastructure or environment (for example, rearranging supermarket or cafeteria products so healthy items are at eye level; painting flies in urinals to improve hygiene; designing street markings to encourage drivers to slow down before dangerous turns; changing the order in which candidates are displayed on the ballot). We call these types of influence *situational*.¹

Some situational influence can easily be made communicative with accompanying messaging – such as a statement pointing out what the default rules are. Others might be considered communicative without such statements, provided enough people understand the implicitly given reasons. The fly in the urinal might seem to have obvious communicative content — or at least its content would become clear after repeated exposure. Other non-verbal signals such as traffic lights are directly communicative, since everyone is aware of what they mean. So, some types of regulation and nudges combine both situational and communicative aspects; the two are not mutually exclusive.

Other types of situational influence are more obviously non-communicative. The fact that the order of names on a ballot has a non-trivial effect on vote choice is not evident to everyone. Nor is the planting of trees closer together on bends in the road to encourage drivers to slow down. This does not mean that they do not provide reasons at all. To the extent they motivate or change behaviour, then we can discern reasons for that behaviour (Levy 2019). We further interrogate that notion below. We introduce the distinction between communicative and situational influence merely to highlight how our account can be applied to each, even to those with little direct communicative content, without claiming there is some natural disjunction between the two types.

Most forms of manipulation aim to change some aspect of the behaviour of the targets. Some forms of manipulation, however, aim not to change behaviour as such, but rather to change the outcome given expected behaviour. People might vote in the same manner over a series of binary votes, but the agenda-setter can manipulate the outcome by the order in which those votes are taken (Plott and Levine 1978; Riker 1986; Dowding and van Hees 2008). Gerrymandering district boundaries need not change the way people vote, just the outcome of the election. We set aside this form of manipulation until section 3 (on situational influence), where we suggest our tests for manipulation and degree of manipulation can be applied to them as well.

¹Our classification into communicative and situational nudges draws on the distinction in Noggle (2021) between psychological and situational manipulation.

We offer here an account of manipulation that can help us identify instances in both politics and public policy. While the account itself has wider applicability to other cases of interpersonal manipulation, our narrower focus allows us to consider complications of dealing with collective decisions and collective action. In developing criteria for distinguishing manipulative from non-manipulative influence, we build on existing accounts of manipulation from both philosophy and public policy. Notably, we modify and repurpose Keith Dowding's (2016, 2018) account of what he calls non-coercive persuasion to address a broader range of contexts and cases. Our account explains why manipulation often purposefully bypasses conscious deliberation, while recognizing that much behaviour does not involve conscious deliberation in the first place. It recognizes that manipulation has a covert element, while acknowledging that manipulative influence itself can be overt. Our account recognizes a division between manipulative and non-manipulative influence, but also allows us to grade the former. Some cases involve mild manipulation (analogous to 'white lies'), while other examples are highly manipulative. The degree of manipulation will figure in the assessment of the normative status of any given policy.

Section 1 of the article engages with recent literature to help identify important desiderata for an account of manipulation to normatively assess the use of influence in politics and public policy. It discusses some limitations of existing conceptions and points the way to overcoming most of those problems. Sections 2 and 3 develop a set of necessary and sufficient conditions to distinguish between manipulative and non-manipulative influence, drawing on Dowding (2016, 2018). Section 2 applies them to communicative influence; Section 3 to situational influence and mixed cases. Section 4 discusses the moral significance of labelling some forms of influence as manipulative, providing criteria to judge their severity. It also argues for a symmetrical application of manipulation criteria across private and public sector nudges. Section 5 concludes.

1. Manipulation in Philosophy and Politics

People influence each other every day. Some forms of influence, such as reasoned persuasion, are morally permissible. Others, such as bargaining, are morally acceptable, unless unequal bargaining is exploitative. Other forms of influence, such as coercion, are seen as morally unacceptable in most circumstances. Generally, manipulation is a type of influence where the target is moved to action without overt threats or coercion, but that falls short of persuasion. There are many accounts in philosophy and political science that attempt to identify the phenomenon and provide necessary and sufficient conditions for manipulation. Some, though not all, rely upon an overly idealized conception of human rationality and deliberation. In this section, we note some of the shortcomings of existing models, arguing that we need a realistic, applicable and more precisely specified model that can be applied to a range of cases in politics and public policy, including nudges.

There is no universally agreed-upon definition of manipulation. Noggle (2022), for example, identifies three distinct conceptions of manipulation, alongside a series of hybrid or disjunctive views aiming to bridge the gaps between them. Some

scholars even argue that manipulation is too multifaceted or vague to be rigorously defined (Ackerman 1995: 337–338; Wilkinson 2013).² The most prominent accounts tend to fall into one of three categories: (1) manipulation as bypassing rational deliberation, (2) manipulation as a form of trickery and (3) manipulation as a form of non-coercive pressure (Noggle 2022).

Manipulation as a form of influence bypassing or subverting rational deliberation (Gorin's (2014a) 'bypassing or subverting view', henceforth BSV) is especially prominent in debates about whether nudges are manipulative. Critics claim some nudges operate 'behind the back of choosers' (Waldron 2014), 'circumvent deliberative faculties' (Grüne-Yanoff 2012), 'undermine that individual's control over her own deliberation' (Hausman and Welch 2010), or otherwise 'pervert' people's reasoning capacities (Wilkinson 2013). Even nudge defenders concede that a nudge can be manipulative 'to the extent that it does not sufficiently engage or appeal to their capacity for reflection and deliberation' (Sunstein 2016: 216, emphasis removed), and that '[m]anipulation occurs when one influences another by bypassing their capacity for reason, either by exploiting nonrational elements of psychological makeup or by influencing choices in a way that is not obvious to the subject' (Blumenthal-Barby and Burroughs 2012: 5).

Nevertheless, BSV accounts offer no consensus on the threshold for such bypassing or subversion. As Noggle (2022) argues, too high a threshold – where one would have to bypass one's rational capacities fully through (unrealistically) effective subliminal advertising or brainwashing – would rule out numerous behaviours commonly considered manipulative. On the other hand, too low a threshold – where any non-rational influence can count as manipulative – would lead to overinclusive accounts, condemning uses of non-verbal or emotional cues not commonly considered manipulative. Another problem with BSV accounts is that neither bypassing nor subversion is necessary for manipulation to occur across a range of plausible BSV accounts; Gorin (2014a) offers a series of counterexamples to show that one can manipulate a target while simultaneously engaging their rational faculties.

Given that defining the right threshold will always be controversial, it would be advantageous for an account of manipulation to avoid requiring an account of how rational capacities should be engaged. Much of what we believe and do is based on testimony, hearsay or habit; but that does not mean we have been manipulated (Dowding 2016). We can be influenced by poor reasons or irrational drives without being manipulated. We might act in a deliberately reasoned manner yet be manipulated – if we have been lied to, for example. An account that does not rely on an account of rational autonomy is an advantage so long as it captures the examples that BSV accounts desire.

The second most common account of manipulation views it as a type of trickery (Mills 1995; Noggle 1996; Barnhill 2014; Hanna 2015). Noggle describes manipulation as a more general form of deception that 'tricks the target into adopting a faulty mental state' and can influence any range of mental states, leading to false beliefs, but also to faulty desires or inappropriate emotions (Noggle 2020: 243).

²Sunstein (2016) argues for a piecemeal approach, distinguishing between canonical/easy cases and contestable/hard cases as an alternative to a fully developed theoretical account of manipulation.

The third account is manipulation as a form of positive or negative pressure that does not rise to the level of coercion (Baron 2003; Wood 2014; Noggle 2020). Examples include emotional blackmail, nagging, flattery and the ‘charm offensive’. Noggle (2020) unifies the second and third account by arguing that both induce the target to make a mistake or engage in flawed reasoning. The deceptive manipulator takes advantage of cognitive biases to influence the target’s decision. The pressure manipulator takes advantage of akratic tendencies, such as the propensity to overvalue short-term obvious costs and benefits over long-term less obvious ones. In all such cases, the manipulator intends the target to fall below certain norms that ought to govern the formation of one’s beliefs, desires or emotions (see also Noggle 1996, 2022).

While there is much to commend in such accounts, some aspects are currently underspecified in ways that leave open how they can be applied to cases. First, one must determine which norms ought to govern non-manipulative influence. Second, one needs to determine whose perspective on the relevant norms ought to take priority. Noggle’s account is open-ended about whether the manipulator intends to get the target to fall below the manipulator’s norms, the target’s norms, or the prevailing social or moral norms. While these tasks are often feasible, a more parsimonious set of conditions is preferable. Our account provides two such conditions.

A fourth account of manipulation is especially prominent in political science. While a subset of the manipulation-as-deception camp, political theorists present it independently with specifically political applications. It derives originally from Goodin (1980), who describes manipulation as involving ‘unknown interference’ (that is, a covert and deceptive exercise of power over the target) and ‘unwelcome interference’ (that is, power exercised in ways contrary to the known or assumed will of the target). Whitfield’s (2022) recent account shares these key features, while aiming to expand the scope of manipulation beyond direct manipulation in interpersonal cases. He offers the following definition:

An act of manipulation is any intentional attempt by an agent (A) to cause another agent (B) to will/prefer/intend/act other than what A takes B’s will, preference or intention to be, where A does so utilizing methods that obscure and render deniable A’s intentions vis à vis B.³ (Whitfield 2022: 786)

Our account is largely compatible with this definition, while bringing out important features of such attempts. It does so without relying on what Klenk (2022) calls the ‘covert thesis’ which equates manipulation to ‘hidden influence’ (Klenk 2022: 86). To be sure, manipulatory acts *themselves* need not be hidden, only (some of) the intentions of the agent.⁴ There are many examples where the manipulatory act is not itself covert (Barnhill 2014; Gorin 2014b; Klenk 2022), but the agent wishes to conceal some key motivating aspects of their intentions. The conditions we develop

³Whitfield (2022) disagrees with Goodin in a few regards. First, he denies manipulation must be direct; he provides some examples of indirect manipulation. Second, he argues that manipulation does not necessarily have to involve deception, since strategically revealed truthful information can also be manipulative. Despite these differences, the shared emphasis on covertness and going against the will of the target warrant discussing them together.

⁴Klenk (2022) demonstrates that manipulative acts themselves are not always covert, but some of the articles he references as making this error do not do so. They only suggest there is something covert within an act.

rule out these forms of concealment without requiring a further discussion of manipulative versus non-manipulative methods à la Whitfield. We should also note that while unseen influence can be manipulative, this does not mean that if the target uncovers the attempted manipulation, the attempt was therefore non-manipulative. As we argue in section 3, disclosure and hypothetical disclosure by the agent can affect the degree of manipulation involved (regardless of the method deployed) in ways that discovery or hypothetical discovery by the target cannot. Whitfield's definition thus needs weakening to admit A's methods aiming to render their intentions obscure or deniable.

For Dowding (2016, 2018), what matters for distinguishing persuasion from manipulation is how the agent attempts to influence the target. Persuasion occurs when reasons for some proposition P endorsed by the agent are offered to the target, and the agent displays deliberative honesty. That is, the agent is herself motivated by or identifies with those reasons and does not deliberately hide contrary considerations. In discussion, the agent is willing to change her views.⁵

So far, we have suggested that we need an account of manipulation that (1) is realistic (that is, does not rely on an idealized account of rational deliberation), (2) is sufficiently precise to apply across a range of cases in politics and public policy, and (3) pays attention to the intentions of the putative manipulator rather than just the target of the manipulation. Reviews of the literature suggest this last point has been relatively neglected (Engelen and Nys 2020; Noggle 2022), although some accounts aim to redress this (Mills 1995; Baron 2014; Gorin 2014b; Klenk 2022). The account which centrally features the agent's intentions is Dowding's (2016, 2018) – although he defines persuasion rather than manipulation. He provides two 'reliability' conditions that are jointly sufficient for non-coercive persuasion: (1) common reasons (roughly, that the agent persuades the listener using reasons that she herself holds for her beliefs) and (2) the agent's intention is to learn the truth together (or otherwise come to an agreement regarding the truth). Unfortunately, Dowding's account suffers some limitations.

First, it is narrowly specified to cases of interpersonal communication in the context of democratic deliberation. The intention condition, in particular, is too narrow to be applicable in cases where a speaker aims to convince an audience or where a government agent aims to influence the behaviour of the citizenry through public policy. In sections 2 and 3, we will show how suitably modified versions of Dowding's original conditions allow us to distinguish between manipulative and non-manipulative influence in a range of contexts within politics and public policy – deliberative as well as non-deliberative.

Second, the original conditions are too demanding. For example, reasons that an agent finds plausible or persuasive, but that are not shared with the target, will fail to pass the non-manipulation threshold. According to Whitfield (2021), Dowding's account rules out non-manipulative cases of 'conjectural reasoning' (that is, cases of persuasion through reasons that the target accepts, but the agent does not). This is partly because Dowding's conditions are meant to be sufficient rather than necessary for non-coercive persuasion. In this paper, we modify and expand the

⁵It thus discounts the 'careless reasons' that Klenk (2022) thinks characterizes communicative acts of manipulation.

common-reason condition to account for a broader range of non-manipulative cases, including cases of conjectural reasoning allowing us to provide necessary as well as sufficient conditions, further improving the original account.

Finally, our account moves beyond the original Dowding conditions to introduce a way in which we can make judgements about the degree of manipulation, helping assess its normative status across different situations labelled manipulative. This allows us to develop a more general account of manipulation that can speak to the concerns of multiple literatures, including the ethics and politics of nudge, the philosophical literature on interpersonal manipulation, and the role of manipulation within politics.

2. Conditions for Non-manipulation: Communicative Influence

In the following sections, we introduce two conditions to distinguish persuasion from manipulative influence. We begin by looking at *communicative influence*, where an agent i attempts to influence a target agent j by providing evident reasons for j to believe P or choose option x over alternatives. We then proceed (in section 3) to investigate *situational influence*, where an agent i arranges the situation that j finds herself in, so as to get j to believe P or choose x . These forms of situational influence are often undisclosed and require separate analysis.

Communicative influence involves evident communication between the agent i and the target j . This communication could be verbal and bidirectional, or indirect and involve a recorded message sent to j by i . In addition to words and arguments, such communications sometimes involve images, sounds and emotional signals, where those signals are evident (such as traffic lights).

We should clarify that, when we say that communicative influence involves i providing reasons to j , we rely on a minimal conception of reasons. A reason is any explicit or implicit statement that can be taken by j to cause, explain or justify belief in P or choice x . These do not have to be morally good, philosophically rigorous, or prudentially sound reasons. They simply have to influence j to assent to P or to choose x , even if the epistemic practice leading j is flawed or inadequate. Biased evidence, the status of i , or even feelings described by i can count as reasons.

Reasons are often thought of as conscious decisions or beliefs. Certainly, those accounts suggesting manipulation involves bypassing rational deliberation seem to suggest so. However, as we have already pointed out, much of our behaviour is habitual (though still propositionally justifiable). On this externalist account we view any affect as a potential reason, even when people are not consciously motivated. We can (sometimes) infer that j is choosing x for reason E ,⁶ but that doesn't mean that j is necessarily aware that E is the reason for their choice. In a famous study, Californian residents were exposed to one of five possible door hangers inviting them to conserve energy (Nolan *et al.* 2008). The control provided (truthful) information about how one could conserve energy. The four treatment

⁶'E' stands for 'evidence' is the dictionary sense of facts or information indicating whether a given proposition is true or valid or in the case of actions justifies the action, though as detailed below one might act on 'content-independent reasons' where the interpretation of E is more complex. It also connects up to the idea of the reason being 'evident'.

door hangers included messages about how conserving energy: (1) protects the environment; (2) benefits future generations; (3) saves money; and (4) is already done by most of one's neighbours. The largest effect on behaviour, measured by household electricity use before and after the intervention, came from the social norm treatment (4). Yet individuals claimed the habits of their neighbours were the least important reason, showing a mismatch between stated reasons and behaviourally revealed reasons. Our externalist account will identify both stated and revealed reasons as reasons, even if the agent may only consciously deliberate about the first category.

Drawing on Dowding (2016), we give two conditions that separate non-manipulative from manipulative influence. These conditions build on the paradigmatic understanding of interpersonal persuasion, where an individual *i* offers suitable reasons to a target *j* in a truthful manner in order to influence *j*'s beliefs, actions or choices in ways that align with *i*'s own beliefs, actions and choices. By contrast, a paradigmatic form of manipulative communication involves individual *i* offering any reason to a target *j* that will influence their beliefs, actions or choices, irrespective of whether the reasons given are shared, truthfully presented or plausible to *i*. In this case, *i*'s goal is only to successfully influence *j*'s behaviour by whatever means necessary. The underlying intuition is that while *i* causes *j* to believe *P*, where *i* and *j* share the same common cause, *E*, leading them to believe *P*, then *i* has not manipulated *j*. Where *i* has used any means whatsoever to lead *j* to believe *P*, then this is not persuasion, but some form of manipulation. Common cause by *E* is weakened (see below) to suitable reasons. We thus redefine non-manipulative persuasion, recognizing that manipulation comes in various forms.

We specify the two reliability conditions for non-manipulative influence: Suitable Reasons (SR) and Testimonial Honesty (TH) conditions. These conditions are weaker than Dowding's original conditions, which were constructed in the context of deliberation and persuasion in order to reach common agreement. Political and public policy communications are not always so deliberative in form.

Suitable Reasons (SR). The reasons (*E*) that *i* offers to *j* to assent to proposition *P* are also suitable for *i* as reasons for *i* to assent to *P*, even if *i* has other reasons not to so assent.⁷ (In choice contexts, proposition *P* constitutes a reason for choosing *x* rather than *y*.)

When we say that certain reasons are 'suitable', we mean that these reasons support the claim that the agent wants to make *in the opinion of the agent i*. These do not have to be objectively true or objectively fitting reasons, but merely reasons that the agent finds suitable as grounds for a given proposition *P* or as reasons to choose *x*. They do not need to support *P* on their own (that is, they need not be decisive reasons) and may require significant supplementary or background assumptions to do so. They need not even constitute the most important reasons why the agent

⁷Dowding's original condition is stronger. His Common Reason entailed that, in persuading *j*, *i* had to provide reasons common for both to believe *P*. Suitable Reason is weaker as *i*'s reasons must only constitute reasons for accepting *P*, even if *i* has other reasons for doing so. 'Suitability' is also relevant to *i*'s role with regard to *j* (see below).

could come to believe *P*, though we will subsequently include a separate condition about truthfulness and relevance in the opinion of the agent.

Ideally, both *i* and *j* believe that the evidence that constitutes the reasons *E* establish a suitable reason to assent to *P* (even if *i* or *j* does not assent to *P* for other contrary reasons). These grounds *E* might constitute content-dependent reasons for assenting to *P*. Proposition *P* does not have to be true, nor even to logically follow from or be demonstrated by *E*. It must only be the case that both *i* and *j* recognize that *E* roughly leads to a belief in *P*. For example, *i* might convince *j* that vaccination reduces the risk of becoming infected with COVID-19 on the grounds that *i* doesn't have any vaccinated friends who have tested positive for COVID-19. Such anecdotal evidence based on *i*'s availability heuristic would not be sufficient to demonstrate the truth of *P*, but if both *i* and *j* believe *E* and take it to be a reason for believing *P*, then *i* has persuaded rather than manipulated *j*. In this example, *i* finds reasons *E* to be suitable reasons to vaccinate, but she might have other reasons – such as some specific medical condition – not to get vaccinated herself. If, on the contrary, *i* was arguing that *j* should get vaccinated based on *E*, but *i* herself found *E* not to be a suitable reason, failing to pass the level of scrutiny that *i* generally uses when making decisions that affect her (or her family), or because it is based on known false data, then *i* is attempting to manipulate.

The condition also does not rule out real-world circumstances where both *i* and *j* come to believe *P* based on content-independent reasons *E* (that is accepting the judgement of another without fully knowing their reasons for their judgement), because of testimony by a third-party *k* whom they deem trustworthy on the subject. Neither does it rule out *j* believing *P* for reasons *E* because *j* sees *i* as a trustworthy source on the subject, provided the second condition below is simultaneously met. Following the example above, *i* might convince *j* to believe *P* on the basis of *E* (for example, testimony on vaccine efficacy from the World Health Organization or national public health officials in their country). Or *i* might be regarded by *j* as having more authority on the subject and therefore *j* adopts the reasons *E* given by *i*. (This claim to authority might be well grounded – *i* is an infectious diseases specialist – or poorly grounded – *i* took a biology course in college. Provided *i* is not deceiving *j* about their credentials or qualifications in ways ruled out by our testimonial honesty condition, the SR condition can be met.)

We can see in other examples that belief in *P* will often have opposite consequences for individual choice. Telling two people that the beverage before them is unsweetened tea can lead one to drink and another to reject it. Such true communication is non-manipulative, even if it predictably leads different people to opposite (often highly predictable) choices. In other words, SR does not require that *i* and *j* come to the same conclusions given *P*, or that they make the same choices based on *P*. This applies to many communicative nudges through which *i* informs *j* about the behaviour of others through proposition *P* (for example, about how much energy *j*'s neighbours consume, how often *j*'s fellow citizens vote in elections, or how much people *j*'s age drink in an average week). As long as *i* communicates suitable reasons for *P*, the nudge is not manipulative. Many people respond to information about the behaviour of others by choosing to conform to the descriptive norm. But others will not. Not only that, but a given individual could oscillate from conformist to contrarian, depending on the context and their mood on a particular occasion.

The revised SR condition responds to Whitfield's (2021) concern about 'conjectural reasons'. Reasons can be suitable even if they are not decisive in the particular case. For example, *i* might convince *j* that choosing a beer from a local microbrewery is morally preferable to choosing one from a larger manufacturer even if *i* is a recovering alcoholic who does not drink or a teetotaler who chooses not to. This means that *i* can appeal to suitable reasons even in cases where *i* does not meet the supplementary conditions for reason E to be decisive in her own case. This gives us a different condition for when conjectural reasons are manipulative compared with Whitfield (2021). The decisive factor is not whether *i* is covert or open about being opposed to drinking alcohol. The decisive factor is whether *i* considers a given reason to be suitable. (Lying and claiming to drink when one does not is manipulative, according to the second condition below, as would failure to disclose when the issue is particularly relevant; but one is not under a general obligation to disclose one's AA membership when discussing alcohol.)

Testimonial Honesty (TH). The agent *i* must intend to provide relevant and truthful information to *j* and not to hide information pertinent to his intentions.⁸

TH focuses on the intentions of *i* in persuading *j* of *P* on grounds *E*. So long as *i* is aiming to provide relevant and truthful information (even if *i* holds false or irrelevant beliefs that *i* wrongly considers to be true), the TH condition can be met. We would not expect *i* to disclose any and all information to *j*. Not only would this be unrealistic, but it might even be counterproductive, since information overload can lead *j* to make worse decisions than following the presentation of less information that is pertinent to the question at hand. Instead, we would expect *i* to use common sense as to which information is relevant. One helpful heuristic would be to ask oneself, 'What kind of information would I want in order to make an informed decision in this matter?'

Statistical information is particularly problematic in this regard. The deliberate misuse of statistical visual representation is well known (see, for example, Tufte 1997: 55–72; Calzon 2021). However, using correct but irrelevant statistics is also problematic. Stating the increased risks of heart disease or dying from an activity, without providing the baseline risk, is ubiquitous in newspaper reports of healthy eating and drinking, for example. Nonetheless, statisticians often disagree on what the relevant statistics are. What matters in our account is the integrity and the beliefs of those providing the information.

The two conditions are jointly necessary and sufficient for influence to be non-manipulative. Their role is to identify unambiguous ways in which tools of persuasion can be used without manipulating the target.

We can see how TH can hold, yet manipulation still occurs by breaking SR, in the imaginary case of *Lucrative Suicide* (Gorin 2014a). In this case, agent *i* is aiming to

⁸Again, this is a weakened version of Dowding's (2016) original TH condition, which includes *i* being open to *j* changing *i*'s mind. This matters in truly deliberative contexts (Dowding's original framework). Nudges are a form of communication, but are not deliberative in that sense. Some of the cases below might be considered deliberative, in which case the stronger condition might be considered more pertinent.

persuade target j to commit suicide by using information, that i considers truthful and relevant, about the non-existence of God. In this case, the reasons presented by i for the proposition P (there is no God) are reasons that i finds suitable for determining the truth of P . However, i himself does not think that these reasons are suitable *as reasons to commit suicide*. So i is persuading j to choose x (that is, to commit suicide) on the basis of reason E (the case for there being no God), which violates the SR condition.⁹

We can also construct examples where the SR condition holds, but the TH condition is violated. These include cases where i considers E to be a suitable reason for believing P or choosing x if E was true, but at the same time believes E to be false. For example, i might be selling ‘junk food’, claiming it is nutritionally valuable when it is not. Although i might believe that E is a suitable reason for choosing their product if true, i can still violate TH if they do not believe that E is true or do not believe that E is relevant for the majority of customers (for example, for those who are not endurance athletes and therefore do not need to ‘carbo-load’).

Real-world examples may only approximate our two conditions in the sense that most real interactions might involve some form of manipulation. We do not consider this as problematic for our account. While we have defined what non-manipulative communication is, beyond the ‘on/off switch’ manipulation comes in degrees. Some forms of communication are more manipulative than others. After all, many of our actions, including persuasion, emanate from multiple reasons. *Ceteris paribus*, the more manipulative a given instance, the more normatively problematic it should appear to be. We have provided ideal conditions for non-manipulation, and as far as the real world approximates those conditions, we do not consider the interaction to be manipulative. As the interaction departs further and further from these conditions, we will consider it to be manipulative; however, some forms of manipulation might seem rather mild, warranting little concern in either personal or public policy contexts. The further away from our conditions the interaction is, the more manipulative it is and the more concerned we should be about it. Finally, though, we note that while manipulation at any level can be considered *pro tanto* wrong, in some contexts even severe manipulation might be normatively justified. In the final section, we introduce several considerations that might mitigate the normative concerns about manipulation in particular cases. Here it suffices to say that the more distant a communicative influence from our non-manipulation conditions, the more it is manipulative.

Our account provides a more precise understanding of manipulation in politics and public policy. In a recent US example, Fox News commentators, such as Tucker Carlson, compared vaccine mandates to apartheid, and undermined their efficacy on an almost daily basis. Yet the company itself mandated that all unvaccinated employees (comprising less than 10% of the total workforce) must undergo daily covid tests – a regime far stricter than that suggested by the Biden administration and subject to daily attacks from Fox broadcasters (Bauder 2021; Ecarma 2021). This seems clearly to be manipulation. First, Carlson seems to be violating TH by failing to disclose relevant information such as the stricter vaccination mandates of the Fox

⁹Incidentally, i would not be manipulative if he believed that the absence of God is a suitable reason for believers to commit suicide, but we are assuming here that he does not.

organization and the high vaccination rate among Fox employees. Second, he and others seem to be violating SR by presenting reasons against COVID vaccination that their revealed choices suggest they do not believe. Of course, further information is required for a final determination in this (or any real-world case), including further information about the putative manipulators' intentions and beliefs.

Another example concerns elected representatives manipulating the public with regard to climate change. The Center for American Progress (CAP) identifies 'climate deniers' as those who deny there is any climate change, or deny it is caused by human activity, or deny scientific consensus on the issue (Drennan and Hardin 2021). The CAP classification is based on interviews or official or informal communications about climate change from each official. To be examples of manipulation, these communications would have to violate either the SR or the TH condition. An SR violation could be a politician who denies global warming because E (for example, there was some unusually cold weather last winter), but who knows that E is a bad reason for denying climate change. A TH violation could be a politician claiming there is no scientific consensus on global warming, which he knows to be false or misleading.

It seems that at least some politicians are violating either SR or TH, or both. Former Senator Richard G. Lugar claims that some of his Republican colleagues are not expressing their true beliefs on climate change: 'So even if they privately believe we ought to do something about it, they're reticent, especially with the Republican president taking the views he is now taking' (Davenport and Lipton 2017). Energy reporter Anthony Adragna (2014) conducted interviews with dozens of former senior congressional aides, nongovernmental organizations, lobbyists and others:

In stark contrast to their party's public stance on Capitol Hill, many Republicans privately acknowledge the scientific consensus that human activity is at least partially responsible for climate change and recognize the need to address the problem.

Adragna lists economic crisis, the limited popularity of policies addressing climate change among Republican voters, the Tea Party's influence over Republican primaries and the hostility of environmental groups towards Republican candidates as reasons for violating TH.

Some of these politicians, of course, might be truthfully conveying reasons that persuaded them to deny climate change and might themselves have been manipulated by others. Climate misinformation is intentionally spread through organizations funded by oil companies. Keith McCoy, a former top Exxon lobbyist, acknowledges Exxon's funding of climate denial in a secret recording:

Did we aggressively fight against some of the science? Yes. Did we join some of these 'shadow groups' to work against some of the early efforts? Yes, that's true. But there's nothing illegal about that. We were looking out for our investments. We were looking out for our shareholders. (Negin 2021)

The extent of such manipulation needs more investigation, but our account points to the key conditions for assessing it: (1) Do elected officials find their own arguments

suitable for denying climate change? (2) Are they conveying truthful and relevant information, at least to the best of their knowledge? A negative answer to either question means the politician is manipulating the public. Such manipulation matters because many people are influenced by politicians, news organizations and social media posts.

We can apply the same conditions to public policy examples. UK civil servant Nick Down was responsible for chasing down £600 million in unpaid personal income taxes and approached David Halpern's Behavioural Insights Team (BIT) for help (Halpern 2015: Ch. 5). The tax office subsequently changed the text of the letter sent to taxpayers with overdue tax liabilities to include the truthful information that 'nine out of ten taxpayers pay on time'. This nudge increased the payment rate by 4.5%. Although multiple agents were involved in designing and implementing the nudge, how far it may be deemed manipulative can be assessed by examining the intentions of key decision-makers, particularly Nick Down and the more senior civil servants involved. To see whether the nudge meets the SR condition, one should investigate whether Down considers that nine out of ten taxpayers paying taxes on time is a suitable reason for paying one's taxes. We could infer the truth of the SR condition based on Down's published statements, responses in interviews and his own behaviour when presented with such social proof. To see whether the nudge meets the TH condition, we should see if the information is true and believed by Down and assess its relevance. Although we would require further evidence for a definitive assessment, we find it plausible that the nudge can meet both SR and TH conditions.

Another communicative nudge example comes from Australia, where the government Behavioural Economics team designed an intervention to reduce online gambling. In a field experiment, they tested whether showing the gambler a statement summarizing their gambling history and transactions would reduce how much they gamble. The statement and the visual elements highlighting losses visually in red is presented in Figure 1. Online gamblers who were shown the statement ended up reducing the amount they bet by 7.6% (Commonwealth of Australia 2020).

Once again, the nudge can meet the two conditions for non-manipulation, provided that the government agents approving its use (1) find the information provided to be a suitable reason to reduce one's online gambling (SR condition) and (2) have selected information that is true and relevant to the gambler (that is, that the account balance actually reflects the individual's gambling history) (TH). Again, both conditions could plausibly be met upon further empirical investigation.

A final communicative nudge example comes from the US Office of Evaluation Sciences (2018). The intervention involved mailing letters to seniors over 65 years of age, informing them of their potential eligibility for a Supplemental Security Income (SSI) managed by the Social Security Office. The letters informed seniors of their eligibility, mentioned the maximum benefit that they could be entitled to and noted that the application process is simple (see Figure 2) (Hemmeter *et al.* 2020). As a result, SSI awards increased by up to 340%. Once again, assessing the degree of manipulation requires evaluating whether the reasons given meet the SR and TH conditions; in this case, too, it seems the conditions are met.

We have discussed the conditions to attend to in determining whether a given communication is manipulative. Of course, real-world examples of communication involve the exchange of multiple propositions, some of which will meet the relevant conditions and some of which may not, and we may wish to focus on the

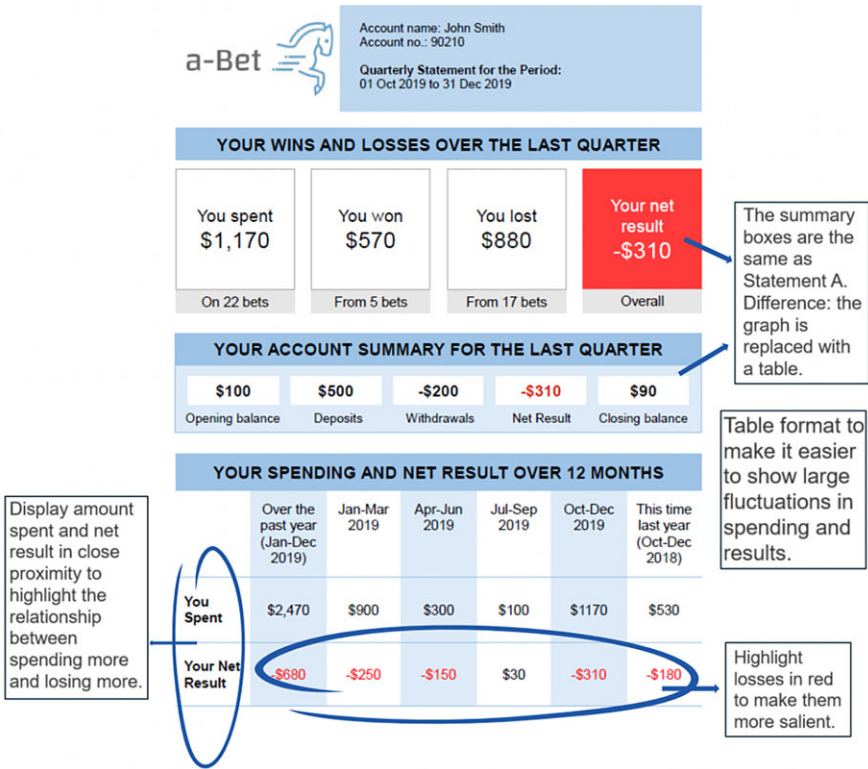


Figure 1. Australian gambling nudge source: Behavioral economics team of the Australia Government, “Better Choices: Applying Behavioural Insights to Online Wagering” (December 2020), Figure 4, p.16 : <https://behaviouraleconomics.pmc.gov.au/projects/applying-behavioural-insights-onlinewagering>.

preponderance and strength of one rather than another. Regardless of the complexities of any given case, however, a proper investigation of such communications should focus on the reasons given and the intentions of the presenter.

While these nudge examples from public policy are arguably not manipulative by our criteria, that does not, of course, mean others may not be manipulative, and we consider more examples below. Now, however, we turn to situational influence, which supplies the examples that are usually provided for manipulative nudges.¹⁰

3. Conditions for Non-manipulation: Situational Influence

We can deploy the two conditions to draw distinctions between persuasion and manipulation. However, many forms of influence in politics and public policy do not provide evident reasons, but instead operate indirectly by influencing the choice

¹⁰Indeed, some writers do not consider giving information to constitute a nudge, so would not consider these examples manipulative. We have taken the broader perspective of most nudge activists and for completeness.

Address Line 1
Address Line 2
Address Line 3
Date:
Claim Number:

[RP NAME FOR, if applicable]
[NH NAME]
[ADDRESS]
[CITY, STATE ZIP]

Our records show you may be able to get Supplemental Security Income (SSI) benefits because you are age 65 or over.

If you are eligible, you may be able to earn up to \$735 (single) or \$1,103 (married) per month in SSI benefits.

Applying is simple! Call to schedule an appointment to apply in person or by phone. A Social Security representative will help you apply.

What is SSI?

SSI is a monthly cash benefit that is in addition to regular Social Security retirement benefits.

How do you apply for SSI?

Call Social Security for more information on how to apply: toll-free at 1-800-772-1213, or call your local Social Security office at *F2. We can answer most questions over the phone.

If you are deaf or hard of hearing, you may call our TTY/TDD number *F3. For general information about Social Security we invite you to visit our website at www.socialsecurity.gov on the Internet. If you do call or visit an office, please have this letter with you. It will help us answer your questions.

Social Security Administration

Suspect Someone Else of Committing Social Security Fraud? Please visit <http://oig.ssa.gov/r> or call the Inspector General's Fraud Hotline at 1-800-269-0271 (TTY 1-866-501-2101).

Figure 2. Supplementary Security Income Nudge Source: US Office of Evaluation Sciences, "Increasing SSI uptake among a potentially eligible population" <https://oes.gsa.gov/projects/increasing-ssi-uptake/>, based on Jeffrey Hemmeter, John Phillips, Elana Safran, and Nicholas Wilson, "Communicating Program Eligibility: A Supplemental Security Income (SSI) Field Experiment" (November 2020), p. 36.

context of targets. Many of the best-known nudges are situational. Take the following four paradigmatic cases:

Opt-out Organ Donation. All citizens are considered organ donors upon their death unless they explicitly opt out of this status.

Cafeteria. Healthy products in a cafeteria are displayed at eye level to make them more salient to consumers buying lunch.

Slow down. Trees on the side of the road are planted closer together around sharper turns to get drivers to instinctively slow down.

Fly in the urinal. A fly painted in the urinal helps men aim and keeps the bathroom cleaner.

Situational influence primarily relies on changes to the default rules or arrangement of alternatives to make certain choices more salient. They are not explicitly communicative between i and j because no overt reasons are provided to the target of the nudge. In some of these cases, the reason for the choice architecture might be considered evident. These days the fly in the urinal is well known and people understand its purpose (though that was not the case when it was first introduced in the Netherlands in the mid 1990s). Drivers are now well aware of lines painted closer together on roads to encourage them to slow down but might be less aware of the effects of trees. Most people do not understand the effects of placing items at eye level until it is pointed out to them. All these examples are designed to change behaviour and the reasons why the targets' behaviour changes may not be evident to the targets themselves. In Organ Donor, an individual may never learn about this policy, particularly as it only affects them after their death.

The reasons behind other situational nudges are even less evident. Here are some noteworthy political examples:

Ballot Order. The names of one party's candidates are listed before the others.

Voters with limited information tend to vote for the names listed higher up the ballot.

Strategic Agenda-setting. The chair of a legislative session decides not to bring an agenda item to a vote since she expects the vote to go against her preference. She postpones the vote until a more favourable opportunity.

Gerrymandering. Members of party i redraw the district map in a way that maximizes the number of seats won by their party given likely voting patterns from previous elections. They are not directly influencing any individual's vote, but rather influence the vote count.

Ballot order is designed to change the behaviour of some voters. For most voters, the choice architecture is not evident; they have no idea the ordering of candidates is deliberate.¹¹ In the other two cases, as we noted above, the intention of the agent is not to change behaviour, but to change the outcome given expected behaviour. Some may notice the changes to the voting rules, but without necessarily understanding their effects on the collective decision.

In general, situational influence is manipulative, even if not all situational influence is equally manipulative. Without evident reasons, situational nudges fail the TH condition of our account. Because no reasons are disclosed to the target, the agent i cannot intend to provide relevant and truthful information to j . However, situational influences can be made non-manipulative by adding a communicative element. If the explicit or implicit reasons are made public, then we can assess the reasons in terms of the TH and SR conditions. For example, a letter could be sent to citizens informing them of their enrolment as organ donors and justifying the

¹¹In Party List systems and occasionally in STV, the parties can order their candidates to advantage their most favoured. In many countries, traditionally candidates were ordered alphabetically. Given there was no intention to affect the result, there was no manipulation. Now aware of ordering effects, some countries randomize the order. Using electronic voting, the order can be randomized – for example, in the ACT (Australia) – so that voters will not all see the candidates in the same order.

change in default rules. (The government of Singapore, in fact, sends such letters to all citizens and permanent residents over the age of 21 (Government of Singapore 2013). One can interpret this move as making the recommendation sufficiently discernible to transform the situational influence into a communicative one.)

Even with situational influence, there are cases where the reasons are (or eventually become) sufficiently clear and the two conditions are met. Then we might say there is no manipulation. In these cases, the influence can be read as a recommendation (Levy 2019). The clearest examples of defaults that entail recommendations come from online purchases where one is often asked ‘Would you like to protect your purchase with supplementary insurance/warranty etc.?’ and one of the choices (usually ‘yes’) is pre-selected, coloured green, or otherwise explicitly endorsed. In these cases, the default clearly comes with a discernible recommendation. To determine whether the influence is manipulative, we need to investigate whether the agent himself endorses the recommendation. For supplementary insurance, the answer is often no. In the organ case, if it is made clear that organ donation is recommended as the default and if the government agents endorse the recommendation, then we are dealing with non-manipulative influence. If the café makes clear that healthy options are put at eye level, then there is no manipulation. If, however, the café does not make clear that healthy options are at eye level or the organ-donation default is undisclosed, then the nudge is manipulatory. But how manipulatory?

Levy (2019) suggests that all situational nudges, including ballot order, can be seen as providing reasons. On his account, all external reasons, no matter that they are non-conscious or exist due to our social psychology or evolutionarily developed neural routines, constitute reasons. In his argument, such nudges therefore do not ‘bypass reason’. Of course, BSV proponents will be unimpressed by this argument since they mean that ‘conscious deliberative reason is bypassed’. Similarly, applying our conditions requires the presence of evident reasons that are given by the agent to the target (even if the reasons are external reasons). However, one way of closing the gap between the desideratum of the BSV proponents and our account in terms of external reasons is to consider how mild or strong manipulation is in such nudge cases. We cash this in term of a hypothetical disclosure test.

In our account, what matters for whether an act is one of manipulation is the intentions of the agent. If their evident communicative recommendation is based on suitable reasons and testimonial honesty, then it is not manipulative. If their non-evident situational recommendation is based on suitable reasons and testimonial honesty, then it is only mildly manipulative. Furthermore, the attitude of the target to those reasons can also be considered a test of how mild the manipulation is. We can consider these two aspects by way of a hypothetical disclosure test.

The hypothetical disclosure test is designed to expose the reasons why situational recommendation was made in relation to what the agent considers to be suitable reasons. We define such a test as the statement that could be made by *i* in accordance with the TH condition. In other words, the hypothetical disclosure statement represents what the agent *i* would say if she intended to provide relevant and truthful information to *j*. The extent of the manipulation depends on what we might call the ‘distance’ between the reasons in the hypothetical disclosure and reasons acceptable to *i*. To be sure, we are using the term ‘distance’ metaphorically.

Measuring ‘distance’ precisely is difficult, so we rely on a psychological shortcut. We can think of this in terms of how unabashed the agent would be to disclose their honest and full reasons for the nudge.¹² For the healthy food example, the disclosure would read: ‘We have placed healthy food at eye level as you are more likely to choose food at eye level and we wish to encourage healthy eating’. For ‘slotting’ behaviour in supermarkets: ‘since you are more likely to choose products at eye level, we place products at eye level according to the highest bidders’. The degree of the agent’s embarrassment at making such statements is one test of how manipulative the nudge is.¹³

The second way of thinking about the hypothetical disclosure test is what effect it would have on the target if the disclosure were to be made. Would the target consider themselves to be manipulated if they were given the full facts? We surmise that, if the target identifies or agrees with the reasons behind the recommendation, they would not feel they were being manipulated, and if they disagree, they would feel manipulated. With nudges, once one knows what the agent is attempting to get one to do, the target can reject it. With healthy food, the target can see the suitable reason, but decide that on balance they prefer the unhealthy food.¹⁴ With the slotting example, the target might feel they are being manipulated, even if they can still override the recommendation. They will know that, even as they try to override the recommendation, they might still be subject to its force. Thus, we can see why, in the example of Krstić and Saville (2019) discussed by Klenk (2022), the Twitter user knows the Twitter algorithm is promoting content that will ‘push her emotional buttons’, but still feels the force of that content. Her knowledge of the intentions of the agents setting up the algorithm does not make it feel any less manipulative, because she cannot override it despite disagreeing with the reasons behind setting up the algorithm.¹⁵

Real disclosures of situational influence are more likely where the nudge is meant to help the customer or citizen make better decisions. The four nudge examples above seem to fit this description and therefore represent cases of minimal manipulation. Real disclosures are highly unlikely where the situational influence is designed to help the agent at the expense of the targets. For example, a political party is unlikely to issue the following disclosure, characteristic of gerrymandering:

¹²We could similarly measure such distance as a subjective assessment by the agent *i* regarding the credence they have in the suitability of a given reason *E*. A suitable reason is a reason that supports the claim *i* wants to make, in the opinion of *i*. Up until now, we have largely treated this concept as a known binary for *i*. However, once we introduce some doubt in the mind of the agent *i*, we can distinguish between reasons that *i* believes to be suitable with a high degree of credence and reasons about which *i* has moderate to strong doubts about their suitability. The higher the credence regarding the suitability of a given reason, the closer to reason is to meeting the SR condition and therefore, the less manipulative. The lower the credence, the more manipulative the reasons given.

¹³Our hypothetical disclosure test can be compared with similar ideas for assessment, such as the publicity condition in Rawls (2005: 68–69), or Pettit’s (2012: 84–87) ‘eyeball test’ in his account of freedom as non-domination.

¹⁴In some examples, such as the evident reasons behind traffic lights, drivers are not manipulated, even if they are coerced.

¹⁵What if the algorithm has that effect, but was not deliberately set up? This is a difficult question we do not go into here. In part it depends upon what we consider to be agents, and issues of commission and omission.

We have redrawn the electoral districts to maximize the number of seats we will win in the upcoming election. Districts that would have otherwise been contested have now been turned into secure victories for us by concentrating the majority of the other party's voters into a few 'packed' districts.¹⁶

The hypothetical disclosure test provides a conceptual way to get around the absence of explicit reasons. But all sorts of evidence can turn the hypothetical test into a real one. Leaked corporate memos, regulatory oversight, investigative journalism, public legal proceedings and scientific investigations are some of the many ways in which further information about the reasons behind certain policies, including nudges, eventually become publicly available. Moreover, democratic governments are usually quite open about their nudge policies, precisely because they are paternalistic: that is, they are designed to enhance citizen welfare.

The above helps explain why both critics and defenders of nudge are preoccupied with transparency conditions (Thaler and Sunstein 2008; Bovens 2009; Schmidt 2017).¹⁷ According to Bovens (2009: 217), therefore, 'every Nudge should be such that it is in principle possible for everyone who is watchful to unmask the manipulation'. However, this is a necessary, but not sufficient, condition for non-manipulation. Mere awareness of the influence is not enough without the reasoning behind it. Thaler and Sunstein also advocate the inclusion of a publicity condition: 'In its simplest form, the publicity principle bans government from selecting a policy that it would not be able or willing to defend publicly to its own citizens' (Thaler and Sunstein 2008: 244). Their publicity principle approximates our public relations test. However, Wilkinson (2013) points out that this principle is insufficiently precise, and some governments seem comfortable defending extremely manipulative policies. It also leaves a lot of work to be done by the ability condition ('able to defend') that requires further normative theorizing to properly explicate. Our hypothetical disclosure statement better reveals what the analyst considers to be manipulative, recognizing that without disclosure there is always a degree of manipulation, however slight.

4. When is Manipulation Morally Justified?

We have argued that, without supplementary conditions regarding disclosure, situational influence will generally be classified as manipulative. However, the degree of manipulation varies, and some manipulation is only petty. Is such manipulation sometimes justified? An important advantage of our less moralized definition of manipulation is that it allows us to separate when influence is manipulative and then to what degree it is so, allowing us to assess the all-things-

¹⁶Where parties do admit gerrymandering, they justify it with the claim 'the other side are worse than we are'.

¹⁷Bovens (2009: 217) distinguishes between *type interference transparency* (where an agent is generally aware they are being nudged) and *token interference transparency* (where an agent recognizes a particular nudge). He also distinguishes between *actual* and *in principle* token interference transparency. According to Bovens, nudging agents do not have to ensure actual token interference transparency provided that, at least in principle, attentive agents can identify the nudge. This is sufficient to distinguish nudge from subliminal messaging (which is entirely unidentifiable). We agree that at least in-principle token interference transparency is required to avoid manipulative situational nudges.

considered moral status of such influence. While our approach to whether influence is manipulative is focused on the agent, our assessment of the morality of manipulation focuses on the target. By first assessing whether a nudge is manipulative through the intentions of the influencer, and then determining how morally unacceptable the act is through the target's perspective, our approach integrates both parties better than earlier accounts.

Much recent discussion of manipulation focuses on the nudge agenda. Here the primary moral concern about manipulative influence is that it reduces the autonomy of the target, where autonomy is generally understood as 'the control an individual has over his or her own evaluations and choices' (Hausman and Welch 2010: 128).¹⁸ Non-manipulative influence allows targets to retain control over their own evaluation and choices. However, many argue that even manipulative nudges are morally acceptable under certain conditions where they either (1) do not reduce autonomy (Wilkinson 2013; Schmidt 2017) or (2) the autonomy reduction is compensated for by countervailing welfare gains (Thaler and Sunstein 2008: 244).

Wilkinson (2013) argues that manipulative nudges to which the target consents do not reduce autonomy and are therefore morally acceptable. Imagine, for example, giving permission to a coach or physical trainer to nudge you into exercising more frequently or eating more healthily. Our account will handle such cases by considering the role of the agent *i*. For example, if *i* is acting as the agent for the principal *j*, then *i* might well give advice in terms of *j*'s world view and not *i*'s. For example, a financial adviser might believe that investing in fossil fuels is the most lucrative form of investment, but knowing agent *j*'s environmental concerns will not suggest such investments. Our approach allows all such agents to be handled in this way, since the communications they give are given with the known interests of their principal in mind. They act as the representative of the principal.¹⁹

The more complicated question involves weighing autonomy considerations against welfare ones. Instead of proposing an objective assessment of the relative weights of these two types of considerations, we prefer to use the target's subjective assessments. Our account of manipulation enables people to identify and assess manipulatory practices. The types of judgements people make are likely to be heterogeneous. Some 'libertarian' citizens will weigh autonomy considerations more heavily, requiring a much larger degree of welfare gains to compensate for even small reductions in autonomy. Other 'welfarist' citizens will weigh welfare considerations more heavily and so be comfortable with more manipulation provided it results in welfare gains. We can expect the same heterogeneity to apply at the level of various constituencies or electorates. The median voter in the US will likely have a different autonomy-to-welfare tradeoff from the median voter in Sweden. Hence minimally or moderately manipulative nudges will be morally acceptable when implemented in countries where the average citizen weighs welfare

¹⁸For further discussions of the different ways to understand autonomy and its implications for whether nudges are autonomy-preserving or not, see Engelen and Nys (2020) and Schmidt and Engelen (2020).

¹⁹Republican accounts see government as such an agent acting on behalf of citizens to promote their 'common interests'. Under this republican model, the agent would not reduce the liberty or autonomy of the population when appealing to these interests as reasons. However, given the heterogeneity of the principal in these relationships we prefer not to view government in this republican manner. That is, we do not assume that 'common interests' ensure the government does not manipulate. We take an individualist stance.

more highly, but morally unacceptable in countries where they more strongly prefer autonomy. We consider this to be a helpful and heretofore overlooked consideration in the ethics of nudging.

In addition to our general principle of respecting the weights citizens assign to autonomy and welfare, we offer two further considerations as part of a normative assessment of the justification or appropriateness of situational influence. These are considerations that should guide the moral assessment of citizens and public officials, both with respect to policy tools and to regulating private-sector influence. Such assessment should involve these questions: (1) is this public policy more manipulative than the alternatives? And (2) are the manipulation criteria applied in (unjustifiably) asymmetric ways across cases?

We can transform these questions into conditions to determine when a manipulative situational influence is morally acceptable. Call the first *the comparative manipulation condition*. For example, for a nudge policy to be problematically manipulative, it would have to be more manipulative than the most likely non-nudge alternative. Some examples of regulation are transparent and meet the relevant conditions. A paradigmatic example of non-manipulative regulations is clearly presented and honestly communicated traffic signs. Not only do citizens have clear signs along every road communicating the relevant traffic laws, but all drivers receive mandatory training in interpreting such signs correctly, meeting the TH condition. Moreover, the reasons behind such rules are often straightforward and endorsed by the relevant decision-makers.

Contrast this with the gerrymandering example we discussed in section 3. New district maps can be drawn for plausibly suitable reasons (for example, to accommodate a growing population) or for plausibly unsuitable reasons (for example, to secure partisan political gains). Adding computer-generated 'fair maps' as recommendation nudges to influence the decision of redistricting committees would seem to constitute a move away from manipulation. Similar cases can be found across a range of regulations in domains as diverse as housing, health, education, family law and environmental politics. When scholars criticize nudges as being problematically manipulative, it is important to compare the proposed nudge with alternative regulatory mechanisms. In some cases, nudges might be more manipulative and therefore a matter of moral concern. In other cases, they might be less manipulative. *Ceteris paribus*, a manipulative nudge should be morally acceptable if it replaces a more manipulative regulatory alternative.

Call the second condition the *symmetry condition*, requiring the moral assessment of manipulation to be applied symmetrically in private- and public-sector contexts. Some believe that government manipulation is worse than private manipulation. The argument sometimes offered is that government has the monopoly on the legitimate use of violence, hence its acts are backed by these coercive powers. However, any manipulative act by a private firm or an individual, if legitimate, is also backed by the coercive power of the government. That is what entails its legitimacy.

The concern should always be: should this act of manipulation be legitimate? Thus, concern about government-run cafeterias arranging food to encourage healthy eating should also elicit concern if privately run cafeterias arrange the food to make high-value-added but unhealthy choices more salient. Similarly, concern about public-sector announcements cautioning citizens about the risks of smoking,

alcohol or diets rich in salt, sugar and fat should equally elicit concern about private-sector advertising encouraging the consumption of these products through similar techniques. Of course, this condition does not imply that these situations are equivalent from an all-things-considered standpoint. Welfare-enhancing nudges have moral advantages over welfare-decreasing ones. Our condition merely articulates that the criteria for identifying manipulation and for judging whether a particular case of manipulation is morally justified should be applied consistently across other domains.

Some people hold an asymmetry thesis. Mark White, for example, argues that private-sector manipulation is morally better than government manipulation, because it is (1) non-coercive; (2) not paternalistic; and (3) expected and guarded against by consumers (White 2013). His claim that manipulation in markets is non-coercive is based on competition: 'If Jennifer gets sick of her coffee shop's manipulative promotions, she can try to find another coffee shop with practices she likes better' (White 2013: 107). It is not paternalistic, because businesses have no intention of benefiting consumers unless it results in increased profits. Finally, and most importantly for White, consumers are already well equipped to deal with manipulation in markets, allowing them to minimize potential damage, whereas citizens' assumptions about government benevolence make them exceedingly vulnerable: 'the core difference between profit-motivated manipulation by businesses and paternalistic manipulation by government: we expect businesses to do it, *but we expect more from our government*' (White 2013).

All three arguments are problematic. As to the first, one can only avoid nudges if one is aware of them. Even then, one can only exit, as in the Jennifer example, if there is a non-manipulatory alternative – which is often implausible; consider, for example, how cars are uniformly marketed. It is not clear why nudges that are to the target's benefit are morally wrong, but those that benefit the manipulators, even when known to harm consumers, are not wrong. Calling the former paternalistic does not make it worse than the latter. Furthermore, surveys do not suggest that the government is more trusted than private business.

Most importantly for our purposes, a classification of a given public policy as manipulative depends on the intentions of the putative manipulator. While it is true that the degree of success in manipulating others often depends on the degree of vigilance, experience and disposition of the targets, we do not classify an act as non-manipulative merely because its targets have become better at resisting. This is not to say that developing strategies for resisting manipulation is not worthwhile and should even be of interest to normative theorists. However, this relates to a different type of enquiry from ours.

Overall, situational nudges involve some degree of manipulation, particularly if the nudge is not explicitly disclosed. However, this does not mean the situational nudge is *more* manipulative than any potential policy alternatives. After all, the status quo might already be manipulative and new nudges could reduce manipulation overall. For example, government could provide eye-level slots in supermarkets for healthy food products, replacing current commercial slotting practices.²⁰

²⁰There is another important consideration about justified manipulation we do not discuss here, as our main concern is with public policy. Sometimes the weak have few resources other than attempting to manipulate the powerful without the latter's knowledge. In an all-things-considered assessment, such manipulation might be considered justified. By our account, it would still be manipulation.

5. Conclusion

Our account of manipulation improves upon previous models by avoiding excessive idealization without unduly privileging the status quo. Previous models either (1) rely on unrealistic accounts of deliberation and rationality, rendering regular interpersonal and communicative practices manipulative, or entail that people cannot be manipulated since they do not fulfil the strong autonomy conditions; or (2) ignore how much our lives are already manipulated, particularly by corporations alongside their advertising and marketing intermediaries. Our conditions acknowledge boundedly rational humans who are often persuaded by inexperienced testimony or shoddy arguments without thereby being manipulated. Yet we can also analyse the historical process of manipulation. Passing on misinformation about vaccines and autism, or conspiracy theories regarding child abuse among US elites, or inaccurate health claims by manufacturers of sugar-sweetened beverages is not in itself manipulative if people truthfully convey and believe the stories. However, the originators of the misinformation may still have manipulated everyone in the chain (Dowding 2016). Our account identifies such manipulation, assigning moral responsibility to the correct agents.

Second, we address a prominent gap in the literature on the ethics and politics of nudging by focusing on the intentions of those who engage in manipulation fulfilling Engelen and Nys's (2020: 15) recent plea for 'an analysis of manipulation that shifts the focus from the autonomy-denying aspect it supposedly has on victims of manipulation, to the specific role and intentions of the manipulators'. Our symmetry condition allows for more sophisticated comparative assessments of public- versus private-sector practices. There might be a role for government manipulation to counter private-sector manipulation, something Schmidt (2017) on republican freedom grounds argues. Our hypothetical disclosure test should be of particular interest to republicans in this regard.

Third, our account can explain the intuitions behind other previous accounts without falling into the traps set by counterexamples. It correctly locates the covertness we associate with manipulation as hiding or partly hiding the intentions of the agent because they fail the SR and/or the TH conditions. It does not entail that the act itself is covert. Further, it demonstrates why some think manipulation involves perverting our reasoning process without relying upon conscious deliberative or idealized rationality. Levy (2019) defends 'behind the back' processes as constituting external reasons and being rational because often our unconscious heuristic reasoning processes are superior to our conscious ones. The social and evolutionarily developed processes are optimal. However, the most egregious manipulation – such as that which leads to addictive gambling, or over-eating manufactured foods – perverts the evolutionarily developed neural reward systems, leading to suboptimal and dangerous decision-making (Ross *et al.* 2008; Dowding 2020). Internet algorithms lead us down paths much further than we would choose at the beginning of the journey, because they appeal to deep-rooted dispositions (Klenk 2022).

Finally, our account provides a means by which to gauge how manipulative practices are, in order to assess their justification as public policy. Situational influence is manipulative, but our hypothetical disclosure process can lead us to

make judgements about the degree of manipulation that can assist in the normative assessment. We have applied our account of manipulation to public policy and politics, focusing on nudge, the subject of much recent literature on manipulation. Previous literature about nudge and manipulation is somewhat insular and piecemeal. Building on Dowding's conditions, we bring these issues towards broader debates about regarding the normative status of different ways of influencing people (for example, Grant 2006, 2011). Our account, it should be noted, provides a method by which to judge whether nudge, government regulations, private-sector advertising practices, newsroom or new media processes, standard rhetoric or political propaganda are manipulative or not. We have not tried to provide a blanket claim that any of these processes are or are not manipulative and we have not sought to justify nor critique specific public practices such as nudge – though we have commented on some examples to illustrate our argument, and it should be clear that we do not consider many nudges to be worryingly manipulative. Our main aim, however, is to provide a tool by which to judge any actual practice, both in itself and by any comparative process with a similar aim.

Acknowledgements. Previous versions of this paper have been presented at the Australian National University School of Politics and IR seminar in March 2021, at the Politics, Philosophy, and Economics (PPE) Society Conference in New Orleans in February 2022, at the UNC Political Theory Colloquium in March 2022, and at the GOODPOL closing workshop Centre for Advanced Studies, Oslo in May 2022. We would like to thank participants for their comments, notably Susan Bickford, William Bosworth, Daniel Casey, Annalisa Costella, Jakob Elster, Anne Gelling, Benjamin Goldsmith, Yi-Hsuan Huang, Begum Icellier, Michael Kumove, Luise Papcke, Alexandru Marcoci, Mehta Majumdar, Charlie Miller, Lars Moen, Sam Schmitt, Jeff Spinner-Halev, Katie Steele, Jana von Stein, Daniel J. Stephens, Shang Long Yeo, John Uhr, Matthew Young, and Michael Zekulin, as well as the anonymous reviewers.

References

- Ackerman F.** 1995. The concept of manipulateness. *Philosophical Perspectives* 9, 335–340.
- Adragna A.** 2014. Many Republicans privately support action on climate, despite public statements. *Bloomberg Law*. <https://news.bloomberglaw.com/environment-and-energy/many-republicans-privately-support-action-on-climate-despite-public-statements>.
- Barnhill A.** 2014. What is manipulation? In *Manipulation: Theory and Practice*, eds. C. Coons and M. Weber, 51–72. New York, NY: Oxford University Press.
- Baron M.** 2003. Manipulateness. *Proceedings and Addresses of the American Philosophical Association* 77, 37–54.
- Baron M.** 2014. The *mens rea* and moral status of manipulation. In *Manipulation: Theory and Practice*, eds. C. Coons and M. Weber, 98–109. New York, NY: Oxford University Press.
- Bauder D.** 2021. Fox's vaccine criticism focuses attention on its own policy. *AP News*. <https://apnews.com/article/joe-biden-business-health-arts-and-entertainment-fox-corp-26096a8781c7c7f1d6c0ddf98a5fe6d>.
- Blumenthal-Barby J.S. and H. Burroughs** 2012. Seeking better health care outcomes: the ethics of using the 'nudge.' *American Journal of Bioethics* 12(2), 1–10.
- Bovens L.** 2009. The ethics of nudge. In *Preference Change: Approaches from Philosophy, Economics and Psychology*, ed. T. Grüne-Yanoff and S.O. Hansson, 207–220. Berlin: Springer.
- Calzon B.** 2021. Misleading statistics examples: discover the potential for misuse of statistics and data in the digital age. *The Datapine Blog: News, Insights and Advice for Getting Your Data in Shape*. <https://www.datapine.com/blog/misleading-statistics-and-data/>.
- Commonwealth of Australia** 2020. Better choices: enhancing informed decision-making for online wagering consumers. *Department of the Prime Minister and Cabinet*. https://behaviouraleconomics.pmc.gov.au/sites/default/files/projects/better-choices-online-wagering-report_0.pdf.

- Davenport C. and E. Lipton 2017. How G.O.P. leaders came to view climate change as fake science. *New York Times*. <https://www.nytimes.com/2017/06/03/us/politics/republican-leaders-climate-change.html>.
- Dowding K. 2016. Power and persuasion. *Political Studies* **64**, 4–18.
- Dowding K. 2018. Emotional appeals in politics and deliberation. *Critical Review of International Social and Political Philosophy* **21**, 242–260.
- Dowding K. 2020. *It's the Government, Stupid: How Governments Blame Citizens for Their Own Policies*. Bristol: Bristol University Press.
- Dowding K. and M. van Hees 2008. In praise of manipulation. *British Journal of Political Science* **38**, 1–16.
- Drennan A. and S. Hardin 2021. Climate deniers in the 117th Congress. *Center for American Progress*. <https://www.americanprogress.org/article/climate-deniers-117th-congress/>.
- Ecarma C. 2021. Fox News' anti-vax mandate messaging is out of step with its own strict policies. *Vanity Fair*. <https://www.vanityfair.com/news/2021/10/fox-news-anti-vax-messaging-policies>.
- Engelen B. and T. Nys 2020. Nudging and autonomy: analyzing and alleviating the worries. *Review of Philosophy and Psychology* **11**, 137–156.
- Goodin R.E. 1980. *Manipulatory Politics*. New Haven, CT: Yale University Press.
- Gorin M. 2014a. Do manipulators always threaten rationality? *American Philosophical Quarterly* **51**, 241–252.
- Gorin M. 2014b. Towards a Theory of Interpersonal Manipulation. In *Manipulation: Theory and Practice*, eds. C. Coons and M. Weber, 73–97. New York, NY: Oxford University Press.
- Government of Singapore 2013. What is HOTA all about? Under the Human Organ Transplant Act (HOTA), four organs, namely the kidneys, liver, heart and corneas, can be recovered in the event of death for transplantation. *Gov. Sg.* <https://www.liveon.gov.sg/>.
- Grant R.W. 2006. Ethics and incentives: a political approach. *American Political Science Review* **100**, 29–39.
- Grant R.W. 2011. *Strings Attached: Untangling the Ethics of Incentives*. Princeton, NJ: Princeton University Press.
- Grüne-Yanoff T. 2012. Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* **38**, 635–645.
- Halpern D. 2015. *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. Harmondsworth: Penguin.
- Hanna J. 2015. Libertarian paternalism, manipulation, and the shaping of preferences. *Social Theory and Practice* **41**, 618–643.
- Hausman D.M. and B. Welch 2010. To nudge or not to nudge? *Journal of Political Philosophy* **18**, 123–136.
- Hemmeter J., E. Safran and N. Wilson 2020. Communicating program eligibility: A Supplemental Security Income (SSI) field experiment. GASOES. [https://oes.gsa.gov/assets/publications/1723%20-%20Hemmeter%20et%20al%20\(2021\)%20-%20Communicating%20Program%20Eligibility%20A%20Supplemental%20Security%20Income%20\(SSl\)%20Field%20Experiment.pdf](https://oes.gsa.gov/assets/publications/1723%20-%20Hemmeter%20et%20al%20(2021)%20-%20Communicating%20Program%20Eligibility%20A%20Supplemental%20Security%20Income%20(SSl)%20Field%20Experiment.pdf).
- Klenk M. 2022. (Online) manipulation: sometimes hidden, always careless. *Review of Social Economy* **80**, 85–105.
- Krstić V. and C. Saville 2019. Deception (under uncertainty) as a kind of manipulation. *Australasian Journal of Philosophy* **97**, 830–835.
- Levy N. 2019. Nudge, nudge, wink, wink: nudging is giving reasons. *Ergo* **6**, 281–302.
- Mills C. 1995. Politics and manipulation. *Social Theory and Practice* **21**, 97–112.
- Negin E. 2021. Despite cutbacks, ExxonMobil continues to fund climate science denial. *DownToEarth*. <https://www.downtoearth.org.in/blog/climate-change/despite-cutbacks-exxonmobil-continues-to-fund-climate-science-denial-79902>.
- Noggle R. 1996. Manipulative actions: a conceptual and moral analysis. *American Philosophical Quarterly* **33**, 43–55.
- Noggle R. 2020. Pressure, trickery, and a unified account of manipulation. *American Philosophical Quarterly* **57**, 241–252.
- Noggle R. 2021. Manipulation in politics. In *Oxford Research Encyclopedia of Politics*, ed. W.R. Thompson. Oxford: Oxford University Press.
- Noggle R. 2022. The ethics of manipulation. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. <https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/>.
- Nolan J.M., P.W. Schultz, R.B. Cialdini, N.J. Goldstein and V. Griskevicius 2008. Normative social influence is underdetected. *Personality and Social Psychology Bulletin* **34**, 913–923.

- Office of Evaluation Sciences** 2018. Increasing SSI uptake among a potentially eligible population. GASOES. <https://oes.gsa.gov/projects/increasing-ssi-uptake/>.
- Pettit P.** 2012. *On the People's Terms: A Republic Theory and Model of Democracy*. Cambridge: Cambridge University Press.
- Plott C.R. and M. Levine** 1978. A Model of Agenda Influence on Committee Decisions. *American Economic Review* **68**, 146–160.
- Rawls J.** 2005. *Political Liberalism: Expanded Edition*. New York, NY: Columbia University Press.
- Riker W.H.** 1986. *The Art of Political Manipulation*. New Haven, CT: Yale University Press.
- Ross D., C. Sharp, R.E. Vuchinich and D. Spurrett** 2008. *Midbrain Mutiny: The Picoeconomics and Neuroeconomics of Disordered Gambling*. Cambridge, MA: MIT Press.
- Schmidt A.T.** 2017. The power to nudge. *American Political Science Review* **111**, 404–417.
- Schmidt A.T. and B. Engelen** 2020. The Ethics of Nudging: An Overview. *Philosophy Compass* **15**, e12658.
- Sunstein C.R.** 2016. Fifty shades of manipulation. *Journal of Marketing Behavior* **1**, 214–244.
- Thaler R.R. and C.R. Sunstein** 2008. *Nudge: Improving Decisions about Health, Wealth and Happiness*. New Haven, CT: Yale University Press.
- Tufte E.R.** 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
- Waldron J.** 2014. It's all for your own good. *New York Review of Books*, October 9. <http://www.nybooks.com/articles/archives/2014/oct/09/cass-sunstein-its-all-your-own-good/>.
- White M.D.** 2013. *The Manipulation of Choice: Ethics and Libertarian Paternalism*. New York, NY: Palgrave Macmillan.
- Whitfield G.** 2021. Two puzzles for shared-reason accounts of persuasion. *Journal of Political Power* **14**, 324–339.
- Whitfield G.** 2022. On the concept of political manipulation. *European Journal of Political Theory* **21**, 783–807.
- Wilkinson T.M.** 2013. Nudging and manipulation. *Political Studies*, **61**, 341–55.
- Wood A.W.** 2014. Coercion, manipulation, exploitation. In *Manipulation: Theory and Practice*, eds. C. Coons and M. Weber, 17–50. New York, NY: Oxford University Press.

Keith Dowding is Distinguished Professor of Political Science and Political Philosophy at the Australian National University, Canberra, Australia, and was previously Professor of Political Science at the London School of Economics and taught at Brunel University and the University of Oxford. He received his DPhil from the University of Oxford. He has published widely in political science, political philosophy, philosophy of social science, public administration and public policy as well as urban economics. His latest book *Its the Government, Stupid*, combines political philosophy and public policy. He edited the *Journal of Theoretical Politics* for many years. URL: <https://researchers.anu.edu.au/researchers/dowding-km>

Alexandra Oprea is Assistant Professor of Philosophy at the University at Buffalo in Amherst, New York, USA, and held previous positions in Political Science at UNC Chapel Hill and the Australian National University. She works in the interdisciplinary field of PPE (politics, philosophy and economics). Her work has appeared or is forthcoming in top philosophy journals, as well three of the top five political science journals such as *American Political Science Review*, *Journal of Politics* and *British Journal of Political Science*. Email: aoprea@buffalo.edu. URL: <https://alexandraoprea.com>