Title
On the moral appeal of nudges

Permalink
https://escholarship.org/uc/item/6qb5412t

Author
Manoranjan, Vishnu

Publication Date
2017
On the moral appeal of nudges

Vishnu Manoranjan

Many nudges are representative of manipulation\(^1\). Consider the basic conception of *nudge* put forth by Thaler and Sunstein (2008): actor *A* nudges actor *B* when *A* makes *B* more likely to perform action *p* by activating *B*’s automatic cognitive machinery, while preserving *B*’s freedom of choice\(^2\). Here, *A* attempts to influence *B* not through straightforward presentation of facts but through activation of automatic machinery. *A* purposefully leaves *B* somewhat in the dark. *A*’s nudge is manipulative. Other authors (Blumenthal-Barby 2016; Bovens 2009; Halpern 2016; Saghai 2013) have focused their attention on whether nudges—by virtue of being manipulative—pose a threat to autonomy\(^3\). Here, I seek to shift the focus to foreseeability—a concept borrowed from Sarah Buss (2005). Following Buss, I take it that foreseeability and not autonomy is the victim of manipulation. Here I will attempt to get at the cases in which nudges pose a threat to foreseeability and also attempt to develop rough criteria for when foreseeability impairment is justified.

Buss (2005) has offered a compelling account of why manipulation does not pose a threat to autonomy simpliciter. I do not wish to recapitulate the full breadth of Buss’s account here but I think some words are warranted\(^4\). Buss’s account takes aim primarily at two common arguments for why manipulation undermines autonomy. The first argument goes as follows: manipulation prevents the manipulated party from being a source of reasons “all the way down.” When someone is manipulated, the argument goes, the grounds for her decisions are placed in her manipulator’s will. The second argument states that to manipulate someone is to treat her in a way that she would not autonomously endorse.

To address the first argument, Buss points out that nonrational influence is ubiquitous. If we say that a person is not autonomous when her actions are at least partially driven by external influences, then “true rational agency is not a conceptual possibility” (Buss 2005). Past and present states of affairs over which we exercise no control surely move us in ways we do not recognize or attend to. And whether a person makes her decision via automatic cognitive machinery or measured deliberation, she is subject to “countless nonrational influences of which she is unaware” (Buss 2005). In response, one might try a different tack and say that an autonomous agent must act in a manner that she herself views as justified. However, deceit and manipulation—as pointed out by Buss—do not run afoul of this conception of autonomy either. An agent who is manipulated still makes her own determination as to how to act; “She govern[s] herself in light of what she … know[s]” (Buss 2005). To object here that a manipulated agent

---

\(^1\) Following Buss (2005), I use the terms manipulation and nonrational influence interchangeably here.

\(^2\) The common example: A cafeteria manager puts healthy food at eye level and unhealthy food in a less visible location; the customer is more likely to choose a healthy option (Thaler and Sunstein 2008).

\(^3\) The worry here (for proponents of nudges) is whether nudging can justifiably be used as Thaler and Sunstein (2008) have suggested: to move people towards “good” outcomes (e.g. healthy choices), without infringing on other values (e.g. autonomy).

\(^4\) For the full, well-developed account, see Buss (2005).
would have acted differently and more autonomously given better information is to implicitly endorse a conception of autonomy that requires an agent to be completely informed in order to act autonomously. Since no person meets this standard and we still take everyday persons to be (mostly) autonomous, this conception is inconsistent with our commonplace notion of autonomy.

Buss responds to the second argument—that to manipulate someone is to treat her in a way that she would not autonomously endorse—by pointing out that an agent cannot endorse an influence that she is unaware of and a lack of awareness is just a feature of any decision-making process. It is impossible for a person—at the time of deciding—to consider and bring to mind all the influences that impact her deliberations. As Buss (2005) states, “we cannot possibly survey all the facts that are potentially relevant to our decisions.” Nonetheless, we do not regard this ignorance as inconsistent with autonomy. Thus, manipulation, which leaves the manipulated party unaware of and thereby unable to endorse particular influences, cannot be autonomy undermining.

Given this forceful and, in my view, convincing account of why manipulation does not pose a threat to autonomy, the reader is left to ask, what makes manipulation objectionable? Buss makes some motions towards the answer. She points out that when we talk of manipulation (and deception), we often have in mind two desirable (and distinct) conditions: “(1) having the capacity to govern our own actions … and … (2) knowing the (likely) effects of our actions, as well as the value of both these effects and the actions themselves.” (Buss 2005). This second condition I call foreseeability and this is what manipulation impairs. When an agent is manipulated, certain information relevant to her decisions is hidden or made less apparent; she is thereby less equipped to predict or foresee the consequences of her actions. Though she governs her own actions, she is left surprised by their results. This is a distressing spot to be in. And it is an aversion to this situation, I think, that commonly underlies our objection to manipulation.

Many of the casual chains of daily life are ones that we readily predict. When we feel thirst, we reach for a glass of water, knowing that drinking water will extinguish our thirst. Here our foreseeability is functionally intact; we move purposefully, assured of our governance over down-chain effects. There are yet other cases in which our foreseeability is impaired. This impairment ranges from minor, where we are sometimes surprised by an effect of our action, to major, where we take actions blindly with no notion of what might obtain. To the extent that pertinent causal chains are obscure to us, we do not foresee. To lack foreseeability is to be left in the dark and to no longer have authorship over down-chain effects. I take it that—although this lack of control is inherently distressing—what really raises our objection to foreseeability impairment is that it leaves us unable to beget those effects that we want. Thus, in my view,

---

5 This conception of foreseeability is not dissimilar to one advanced by David Lewis (1986). Lewis (1986)—talking of what it means to kill rather than simply to cause a death—says, “If a chain is insensitive enough that you can predict it, then it is insensitive enough that you can kill by it.” I say, if an agent can predict a causal chain, then she can author down-chain effects.

6 Buss (2005) brings up another reason why we regard manipulation as unpalatable: the manipulator treats that manipulated as “an (autonomous) object in his world… rather than someone with whom he shares the world.” I agree with Buss’s point but, nonetheless, the focus here is foreseeability and when we might justify impairing it.
foresightability has *instrumental value*; it is valuable insofar as it aids us in bringing about valuable outcomes.

I take it that many nudges are representative of manipulation and thereby threaten foresightability. These sorts of nudges will be those discussed here on. Let us remind ourselves of how such nudges⁷ work. Say that actor $A$ nudges actor $B$. To nudge $B$, $A$ arranges $B$’s external circumstances in such a way that (through automatic cognitive machinery) $B$ is weighted towards taking a particular action. $A$ tries to get $B$ to decide to do what $A$ wants her to do. In so doing, $A$ obscures some features of the decision-making process and leaves $B$ less in touch with reality. $A$ does not present material bearing on $B$’s decision straightforwardly. Thus, $A$ threatens $B$’s foresightability.

For proponents of nudges, the questions of interest are as follows: (1) do nudges impair foresightability and (2) if so, when is foresightability impairment justified?

Let us deal with the first question first. From the foregoing depiction of nudging, it appears that nudges—by virtue of obscuring causally relevant features—do impair foresightability. I take it, however, that there are special cases in which nudging does not impair foresightability. Consider the case in which $B$ is bereft of foresightability, with no predictive powers to speak of. In this case, $A$ nudging $B$ does not impair $B$’s foresightability (there is just no foresightability to impair). If $B$ were to act on her own, without being nuded, her movements would be random, unmoored. Post-nudging, $B$ is no more in the dark than prior.

Consider another case: $B$’s predictive powers are intact but she lacks insight into how she will feel about the possible down-chain effects of her actions. That is, she is unable to assign value to these effects. Here, again, the down-chain results of $B$’s actions remain mysterious to her. And here, again, her foresightability is lacking. Any action $B$ takes will be arbitrary and not based on notions of consequence ($B$ has no such notions). So we conclude here again that nudging does not impair $B$’s (absent) foresightability.

Take another case: say no matter what $B$ chooses to do, $p$ or $q$ will occur, each with equal likelihood. Here, the causal chain that renders future outcomes is insensitive to $B$’s movements. $B$’s foresightability is thus stripped of instrumental value; it cannot be employed to bring about or increase the likelihood of a favored outcome; it is functionally inert. Nudges here are inconsequential and do not functionally impair foresightability. Neither uninfluenced action nor nudged action will affect which outcome obtains.

---

⁷ When I talk of nudges, here, I have in mind doctor-patient interactions; more specifically, I am thinking of those doctor-patient interactions in which a concern for patient wellbeing is taken as granted. This is for two reasons. First, these are just the types of interactions that interest me where nudging is concerned. And second, it allows me to *prima facie* sidestep another worry about manipulation—-that it leaves the manipulated party treated merely as a *tool*—because, at a minimum, a concern for another’s wellbeing is a component of treating her as an end (Buss 2005). (Of course, a full account of treating an agent as an end likely requires more to be said, but I will set that job aside in this paper).
Let us retrace our steps. The foregoing cases are representative of situations in which nudging cannot be objected to on the basis of foreseeability impairment. We are left with the perhaps unexciting conclusion that nudges do not run afoul of foreseeability just in case: (1) the nudged party lacks antecedent foreseeability or (2) antecedent foreseeability lacks instrumental value.

What about cases in which an agent’s foreseeability is neither completely lacking nor stripped of instrumental value? Can further impairment be justified? We have arrived at the second question of interest.

From here, I will try to locate criteria that justify nudging in light of its consequent foreseeability impairment. I do not mean to say that these criteria are necessary to justify foreseeability impairment. I do think, however, that these criteria will help us spell out the easier cases and help identify the harder ones.

Consider the following case: Patient $P$ is visiting the office of her doctor $D$. $P$ has a stomach ache and is contemplating taking an herbal tonic which she believes will cure her. $D$, however, knows this tonic to be poisonous. She also knows that $P$ has a number of misapprehensions about digestive health and that disabusing $P$ of these misconceptions is not a practical possibility. In the end, $D$ nudges $P$. She deliberately presents treatment options in a manner—through vocal inflection and posture—that she knows will weight $P$ heavily towards a safe and efficacious antacid.

I take this to be a case in which nudging is manifestly justified. What features might we point to to build our justification? First, $P$’s antecedent foreseeability is significantly impaired. It is so impaired that she wrongly predicts which effect her action will bring about. Prior to nudging, $P$ is in poor position to beget her desired outcome. $P$’s foreseeability has little value. Thus, any additional diminution in foreseeability is minimally destructive.

Second, $P$’s foreseeability is resistant to repair. That is, $P$’s foreseeability cannot be readily bettered. Her mind is crowded with misconceptions not easily swept away. If $D$ were capable of readily repairing $P$’s foreseeability through some action $R$, then nudging—rather than $R$-ing—would be harder to justify. I imagine $R$-ing as a minimal-cost process of clarification—after which $P$ would be expected to arrive at a reasonable outcome, un-nudged and by virtue of her own improved contact with reality$^9$. Conceptually, $R$-ing is preferable to nudging because it fosters a suitable outcome without impairing foreseeability.

Third, the outcome pursued by $D$—by nudging $P$—is of high value. $D$ nudges $P$ to preserve $P$’s wellbeing, to prevent substantial bodily harm, to prevent the manifestly bad outcome

---

$^8$ This process of criteria location is not dissimilar to the scheme employed by Douglas Husak (2012) as he attempts to find justification for paternalism.

$^9$ $R$-ing as I envision it brings an agent closer in contact with reality, distinguishing it from nudging. Of course, there may be cases in which a physician thinks that even if a patient were to have intact foreseeability she would still make a poor—by the physician’s lights—decision. In such cases, the cost of foreseeability impairment must be balanced against that of allowing said patient to make an un-nudged—and possibly detrimental—decision.
inadvertently pursued by $P$. When the end promoted by nudging is of high value it more easily outweighs the value-defeating effect of foreseeability impairment.

*Fourth*, the nudge applied by $D$ is non-faulty; that is, it is likely to bring about its anticipated end. We could not justify a nudge borne out of incompetence. Nor could we justify a nudge half-heartedly applied. The nudging party—whose foreseeability supplants that of the nudged—has a duty not only to promote a *good* end but also to render a robust nudge, one that makes said end likely to occur.

So the criteria I put forth for justifying nudges are as follows: (1) the would-be-nudged already possesses significantly impaired foreseeability and (2) said foreseeability is not readily bettered and (3) the outcome pursued via nudging is of high value and (4) the nudge applied is non-faulty. These criteria, I think, capture the easily justifiable cases. And to the extent that a case deviates from these criteria, it will be harder to justify.

One more point—up until now, I have been talking of foreseeability impairment as if it is confined to just the decision domain in which nudging takes place. But it is not hard to imagine that once in possession of a distorted reality, an agent might find her misapprehensions difficult to shake. She might in turn make numerous decisions on the basis of these nudge-induced misapprehensions and thus, her foreseeability impairment might be broadly destructive. Practically, I think it is difficult to gauge how sticky and far-reaching a particular impairment in foreseeability will be (i.e. how long an agent will retain a distorted viewpoint and how many decisions she will make in light of it). All this is to say that weighing out the negative effects of foreseeability impairment presents a practical challenge.

Let us now walk through the practical upshot of the criteria previously highlighted. Say $D$ is faced with another patient $Q$. $D$ is a good doctor, and is concerned about $Q$’s wellbeing. She thinks it best if $Q$ were to take some drug $y$ for high blood pressure. And she is thinking about whether to nudge $Q$. How might this go? $D$ might first reasonably ask herself whether she need nudge at all. She might remind herself that nudging impairs foreseeability and that this impairment might be far-reaching. Thus $D$ might think it better to present information straightforwardly, avoiding foreseeability impairment if possible (see criterion 2). If nudging need be considered, $D$ will be left to weigh the costs of foreseeability impairment against the benefits of getting $Q$ to take $y$ (see criteria 1 and 3), a challenging task (given the difficulty in gauging the extent to which a particular nudge impairs foreseeability). And then if a nudge is to be applied, it must be done so with finesse (see criterion 4)—just enough thrust to bring about the intended outcome but not so much as to leave $Q$ woefully and enduringly misguided. Where foreseeability impairment is concerned, nudging is a delicate business.

A few notes about the limitations of this paper are in order. As alluded to above, the criteria set out here are meant to form a rough sketch of those situations in which nudges are justified despite consequent foreseeability impairment. Perhaps only the easy cases are captured. So an analysis of tough cases might be in order. Additionally, I concern myself here only with the

---

10 I can think of two major challenges that might arise in tough cases: the practical challenge of weighing out the negative effects of a particular impairment as well as the difficulty of weighing out the upside of a certain nudge (particularly if the nudging party possesses limited foreseeability).
instrumental value of foreseeability. If foreseeability has intrinsic value, then an account of this sort of value is needed.

What makes nudging justifiable is still a live issue. But I think much of the focus has been misplaced. The concept of foreseeability deserves greater attention.
References


