Is nudging legitimate public policy?
Practical and conceptual challenges to legitimacy of nudging as a public policy

Patrik Pavlovský

ABSTRACT
There’s been a very few psycho-economical concepts in a recency, which would so noticeably caught attention of public and private sector representatives, as well as that of critical academia, than as it was done by the so called nudging. Since its conception, few countries got sufficiently motivated to set-up their own „nudge units“. It thus seems that the theoretical base and claimed powers of the nudge concept got quickly attractive enough to get translated into the actual practice of modern states. This achievement nonetheless does not by itself necessarily confirm nudge’s bulletproofness, ethical purity or its actual utility to society. There were several relevant and critical attempts to challenge this notion on all of these fronts since its conception, but there were not that many attempts to rescue nudge from these challenges.

The most vocal criticisms of nudging challenge it for the way it was translated into public policy practice (practical) and for the vagueness of its conceptual skeleton (conceptual). This paper thus gathers and reviews several critical practical and conceptual challenges that strive to discredit nudging as legitimate public policy. Challenges to nudging’s ethical dimension won’t be directly scrutinized in this paper.

KEYWORDS: nudging, public policy, legitimacy, manipulation, challenges
INTRODUCTION

There’s been a very few psycho-economical concepts in a recency, which would so noticeably caught attention of public and private sector representatives, as well as that of critical academia, than as it was done by the so called nudging. The work of Richard Thaler and Cass Sunstein that popularized this notion, also marketed it so well that only a few years after its birth, the Cabinet office of the United Kingdom founded a special advisory unit that was supposed to generate policy advices specifically tailored from the behavioral economy perspective. Since then, few other countries got likewise sufficiently motivated to set-up their own „nudge units“. It thus seems that the theoretical base and claimed powers of the nudge concept got quickly attractive enough to get translated into the actual practice of modern states. This achievement nonetheless does not by itself necessarily confirm nudge’s bulletproofness, ethical purity or its actual utility to society. There were several relevant and critical attempts to challenge this notion on all of these fronts since its conception, but there were not that many attempts to rescue nudge from these challenges.

The most vocal criticisms of nudging challenge it for the way it was translated into public policy practice (practical) and for the vagueness of its conceptual skeleton (conceptual). This paper thus gathers and reviews several critical practical and conceptual challenges that strive to discredit nudging as legitimate public policy. Challenges to nudging’s ethical dimension won’t be directly scrutinized in this paper.

The common denominator of all these is that they explicitly or de facto challenge nudging legitimacy as public policy tool. When referring to nudging’s legitimacy, I will be referring to its descriptive instead of normative account.

“If legitimacy is interpreted descriptively, it refers to people’s beliefs about political authority and, sometimes, political obligations. [...] a political regime is [descriptively] legitimate means that its participants have certain beliefs or faith in regard to it. [...] In
contrast to [...] descriptive concept, the normative concept of political legitimacy refers to some benchmark of acceptability or justification of political power or authority and—possibly—obligation” (Fabienne, 2017).

Consequently, I aim to offer reasonable arguments and strategies for answering reviewed challenges, and hence try to in effect vindicate nudging’s reputation as legitimate public policy. There seem to be 3 potential strategies for the task, only first two will be discussed in this paper:

STRATEGY #1: Reconceptualize nudging so it would denominate only uncontroversial influences;

STRATEGY #2: First define ‘guilty’ manipulation, then distinguish ‘innocent’ nudging from it;

STRATEGY #3: Find a reason why it is sometimes permissible to influence people with objectionable methods;

This paper thus presents a fusion of literature review and conceptual analysis and aspires to reasonably vindicate nudging’s reputation as public policy at least to some extent.

PRACTICAL CHALLENGES

One of the most common-sense methods for legitimizing the suitability of a new solution to a problem is to first and foremost repeatedly test it under the circumstances, in which it aspires to be an operational solution and on the basis of the performance it yields, decide whether to apply it or not. Put differently, when promoting new solutions, especially in the context of public policy, the most legitimate ones are usually those stemming from the ‘hard data’. Naturally, this method is not by far the only one and by itself sufficient for making decisions about new policies, yet it is somehow natural and reasonable to expect the hard data behind such decisions. With respect to the relatively rapid public policy upsurge of nudging, some authors questioned the sufficiency of the hard data that was
supposed to support its actual and infallible utility to society (Lowenstein et. al., 2010; Marteau et. al., 2011; Bonell et al., 2011; Lang et al., 2011; Lowenstein et. al., 2012).

„Nudging can certainly trigger behaviors that worsen our health, but can it also be used to cue behaviours that improve it? [...] While there is evidence of effectiveness for internet feedback, [...] to date, few nudging interventions have been evaluated for their effectiveness in changing behaviour in general populations and none, to our knowledge, has been evaluated for its ability to achieve sustained change of the kind needed to improve health in the long term” (Marteau et al., 2011, p. 264).

While some critical experts, following their common-sense expectations from public policy innovations, simply highlighted that the youth of nudging should be met with greater cautiousness due the still limited empirical evidence, others like Marteau and her colleagues went even further by claiming that there is no persuasive evidence that nudging can yield longly sustainable and population-wide results. Application of nudging without the hard data thus may according to these opinions turn-out either as cost-ineffective solutions or even worse, as potentially harmful if designed on top of poor evidence.

Comparatively to the policy solutions informed by the traditional economy that focus e.g. on the barriers to a trade, Lowenstein (2011; 2012) and his colleagues suggest, nudging (and the behavioral economy from which it stems) yields too often too soft and thus less efficient and more costly solutions for being considered a relevant source of policy solutions. Effectless calorie labels in restaurants or compliance rewarding system (which seem not to generate it) are cited as an example of such inefficient or expensive solutions suggested from the behavioral economy. For them, in other words, nudging is too soft form of paternalism, because in many contexts it cannot sustainably bring the desired behavioral change. Unlike e.g. sanctioning of an undesired behavior, which should directly and clearly increase the costs of this behavior and thus likely bring the behavioral change by deterring
from the undesired alternative, nudging by definition cannot generate such influential impact on the economy of a choice set and thus will at best only occasionally cause a short-term change.

Lastly, some people still ask whether the concept of nudging is indeed so innovative as the hype around it suggests. The only noticeable novelty of it ostensibly rests in its direct conceptual incorporation of libertarian ethics, all the other features were supposedly harvested from other disciplines and merely rewrapped into nicely sounding, new packaging labelled ‘nudging’.

Several facts need to be noted before answering these critical points. First and foremost, most of this practical criticism was articulated 3 to 4 years after Thaler’s and Sunstein’s book has been published, therefore only 1 or 2 years after the first, governmentally funded nudge unit has been conceived. Remorse for the lack of the hard data supporting nudge’s efficacy was thus at that time seemingly at place, for it is practically impossible to generate a surplus of scientific evidence for this or that particular application of any concept in such a short time period. It is for this reason why they were rather skeptical about the hype for nudging’s application into the public policies, since these should be taken seriously due their eminent impact on people’s lives.

Their critical voices against the hype could be actually translated into the vocabulary of behavioral economics, in particular as a warning against falling short of optimism or confirmation biases. Their common sensical arguments in other words fittingly pointed out that the rapid institutionalization of an undertested concept always bears with itself an increased likelihood of a failure(s) caused by concept’s application, for if the public decision-makers were overly optimistic or confident about the abilities of nudging or nudgers, they might have had paid a lip service to all the evidence suggesting the opposite. And it is the citizens who will pay the price. In particular, it was supposed to be the thinness
of the hard data supporting the efficacy of nudging that should have been cautiously considered as the sign of nudge’s unreliability, a sign that was according to these critics underconsidered.

Secondly, most of these practical critical voices were expressed as a commentaries of particular public policy practices rather than as the argumentative attempts for total dismissal of nudging as a concept. They were often published in newspapers rather than in traditionally academic spheres of reciprocal criticism, hence it was okay to express one’s intellectual opinion one-sidedly and only via raising doubts. On the newspaper platform, the overall subjectivity of presented position is in my opinion allowed to a greater extent than it is in the academic journals, especially because in the latter it is the community of alike experts that evaluate author’s justifications of his arguments and propositions while on the former platform, the peers are rather laic masses whose attention and hearts are to be grabbed only. This difference is important for translating the main message of these practical criticisms.

Despite the critical tone, these authors didn’t aim at the dismissal of nudging altogether, for it would be too shallow to resort to such strong position only on the basis of blame for lack of hard data immediately after a concept has been born. From a laic, this argumentative move would be understandable, but since the cited criticism came from experts, its main legacy should be understood rather as a criticism of public policy figures and practices, who and which perhaps too recklessly allowed nudging to influence people’s lives. If this is indeed true, nudging must not apologize or feel guilty for the role and trust it obtained from public officials, and thus must not answer to these worries at all. It is the public policy practitioners who are being targeted.

Surely, it is doubtless that its public policy application should have as much hard data as possible. Looking at these criticisms with the distance of time suggests that that’s exactly
for what the governmentally funded nudge units have been created for: to generate a solid hard data for all kinds of innovative solutions suggested by the insights of behavioral economy.

Nudging was also charged of softness and excessive narrowness of its solutions, a challenge that should play in favor of more traditional and harder and thus coercive forms of paternalism, because they presumably incriminate nudging of inefficacy, since apparently an occasional “good shove advances individual and social welfare considerably more” (Lowenstein et al., 2012, p.344).

While that might be in some cases practically true, it is one thing to be efficient and other to be efficient even despite greater interferences with personal freedoms. For however occasional and well-minded the ‘good shove’ might try to be, it is still a shove at expense of prima facie valuable personal freedom of the shoved person, at least within liberal democracies. And especially when the hardness is chosen primarily for the sake of policy’s utility, one has to first and foremost justify why the efficiency should be the key for adjudicating the selection between freedom-preserving tools such as nudging, and the freedom-interfering tools such as restrictions and harsh penalties.

This argument might seem plausible for a person driven by utilitarian morals, in which the benefits extracted out of higher overall amount of compliance generated by a good shove into one’s freedoms could be in themselves considered as sufficient reason for outweighing the expenses incurred by the strength of an impact of the shove. The utility obtained from more rather than less compliance is simply worth it. So of course, quadruplicating the tax on e.g. unhealthy products would resonate on their overall sales more strongly than playing with their display at the market, but it would also transform the policy impact from suggestion to dictation of the desired lifestyle since it would mean the de facto closure of a particular life-style option for significant amount of population. My point is that even if
political authority have a good reason(s) to avert some life-style, an argument has to be made as to why it is legitimate to curtail one’s freedom in the life-style choice by hard instead of soft measures.

But for doing so one has to first make an argument why it is legitimate for political authority to limit e.g. self-inflicting of harm. Lowenstein’s statement not only seems to presume such argument, it seems to also presume that the consideration of policy’s efficacy is supposed to play a role in that argument as if only relevant actions of state towards enforcement of particular behavior (in any sphere of life) were those that bring the most behavioral change. While this expectation might be true across wide ideological spectrum when we talk about particular type of policies, e.g. the crime prevention policies, it is not similarly true across similarly wide spectrum when we talk about life-styles.

Nudging moreover never aspired to entirely replace stricter forms of behavioral regulations, nor it will be able to suggest a cure for every kind of problem in a political community. There’s thus no point in arguing that nudging cannot and often is not outplayed by other forms of intervention. On the other hand, claiming that nudging cannot yield solution to complex issues or that it cannot yield solution with a long-term effect would now be unfair. The default option nudges, such as the automatic organ donation enlistment while obtaining a driver license (Wright, 2013), the environment-friendly defaults (Sunstein & Reisch, 2014), or the default payroll-based pension saving plans (Benartzi & Thaler, 2013) seem to yield very promising results in comparison to the previous settings.

Moreover, the oldest nudge unit - the UK based Behavioral Insights Team - presented among its latest achievements e.g. their ability to markedly improve academic performance at the nation-wide standardized tests (BiT, 2017, p.9), or the ability to reduce speeding via thoughtful feedback.
Nudging-enthusiast can thus in my view agree that, in order to avoid unwarranted policy consequences, it should be a standard to justify any policy solution on a basis of solid, hard data background. This is nonetheless a challenge to the standards of public policy practitioners, not to the concept of nudge. Nudge is indeed guilty of softness, it interferes with people’s behavior as if they were extremely fragile porcelain figures. It does so on purpose, for it by default values their almost ultimate authority over the course of their lives. It moreover does not and must not crown itself an answer to the problems of all kinds, nor the ultimate replacer of traditional paternalistic solutions. Like any other concept, it has its limits. It’s up to its practitioners to prove that its limits are farther than one may think.

CONCEPTUAL CHALLENGES

Nudging has one surprisingly still relevant conceptual answer to give. Was its conceptual skeleton laid out satisfactorily enough to clearly distinguish ‘nudges’ from other forms of interferences? Even if the answer is positive, does nudging indeed represents a new concept? For Bonell (et al., 2011, p.242) and his colleagues, “it isn’t clear how nudges are distinctive in any other way”. According to them, many mechanisms of nudging such as e.g. motivational interviewing, emotion triggering, peer education or structural interventions into the environment has been known to for triggering behavioral responses long before Thaler and Sunstein coined the concept.

Marteau (et al., 2011, p. 263) and her colleagues share the same doubts but simultaneously acknowledge that these doubts are not the most important conceptual issues that nudging faces. For they fittingly point that nudging’s “novelty lies in two features”: firstly, in the linkage of its effects via contrast with the rational model of human behavior used standardly in the traditional economy; and secondly, in its conceptual fusion with libertarian ethics. In other words, the innovativeness of nudging doesn’t necessarily stem
from the originality of all the ingredients used for its creation. It stems from the innovative and unique way of combining known ingredients, namely from the psychology, marketing, economy and ethics. Nudging is for social science industry what a new cocktail recipe is for bartending industry – an original product created by a creative and unique usage of the industry’s wisdom in a way not done before.

We could in fact go a step further and claim that nudging is a product designed and marketed first and foremost as a tool for public authorities who in some way care about their subjects’ freedom. Its application is naturally not limited to public sphere, yet it is mainly the public authorities (out of the pool of possible users of this tool) that may specifically appreciate its ethical dimension, since it seems obviously compatible with the legitimate mandate that democratic regimes nowadays are expected to uphold. This default compatibility moreover symbolically communicates commitment to a value of freedom, the commitment that seemingly upsurged during recent history of most political communities.

The only question that thus still begs an answer concerns the quality of nudging’s conceptual recognizability, especially when contrasted with other behavior-changing strategies. It is rather pity that Thaler and Sunstein originally devoted only minimal space to nudging’s conceptual presentation. Their characterization of nudging was adequately summarized by following paragraph:

“A nudge, as we will use the term, is any aspect of the choice architecture that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid. Nudges are not mandates” (Thaler & Sunstein, 2008, p.6).

First and rather clear condition for identifying nudge is the predictability of influence’s consequence. To count as a nudge, in other words, we must be able to predict the
influence’s consequence with sufficient certainty. This appears to be simultaneously a direct response to doubts and challenges of nudging that highlighted its potentially harmful or otherwise unwarranted effects and side-effects, for it renders influences as nudges only if they (and their effects) are well-studied and understood influences on human behavior. Random or unpredictable influence on human behavior, or an influence with unpredictable effects, is not a nudge in Thaler’s & Sunstein’s notion of the concept.

It is for this reason that one should not nudge e.g. via inducing stress, for despite the still expanding knowledge about the phenomenon and its possible effects, we cannot with sufficient certainty predict the way towards which it would influence subject’s behavior, especially if such influence would be applied to large population at once. Of course, a question of the satisfactory level of certainty about influence’s effect comes immediately to one’s mind and there appears to be no objective rule to which one may immediately and safely point. Yet this is not an issue, for this uncertainty problem is not only a problem of nudging, but of all empirical sciences as such, and finding an answer to it requires one to extend his search deep into epistemological philosophy. It is thus safe, I believe, to warrant before attempts to nudge via rather experimental and insufficiently explored and attested mechanisms.

Secondly, behavioral influences ought to count as nudging if and only if they doesn’t factually or de facto close or significantly handicap other available alternatives. Redesigning of the choice architecture in a way that other viable options are no longer available is not the kind of influence that counts as nudging. Introducing substantive rewards or punishments for some options and not for others is not the kind of influence that counts as nudging, for despite the fact that options were not closed, substantive incentives make some of them so unattractive and irrational that their presence seems de facto useless.
While the former example is clear, the de facto closure is and always will be subject of polemics, because a threshold beyond which an option becomes de facto closed has to be qualified. It nonetheless seems difficult to find a ‘bulletproof’ definition of this threshold.

The logic of option’s possible closure implicitly suggests that the availability of an option can be commonsensically evaluated on a scale between two poles: fully available and fully unavailable. Thaler’s and Sunstein’s definition stipulates that economic handicap should be the unit of this scale, while a pareto efficiency is presumably expected between the economic handicap placed upon an option and its availability. It nonetheless remains unclear against what metrics should we assess the significance of the handicap to a person concerned, since different individuals may respond differently to one level of the handicap. Should we calculate a sort of ‘tolerability index’ that would consider various economic and psychological factors that may influence one’s perception of the burden placed upon an option? Or should we stick with subjective interpretations of such perception? Should we assess the burden against the mean value of given option of concerned population? Can we uncontroversially find out option’s mean value?

Similar metrics problem haunts the second condition conceptualized into the nudge. Besides forbidding or significantly discouraging an option, a **nudge must be easy and cheap to avoid**, yet we were left with no clue as to against what to measure the easiness and the costs of avoiding a nudge.

If these conceptual deficiencies won’t be mitigated, it would be doubtful whether nudging can be indeed distinguished from other, rather unwarranted, freedom-undermining influences on human behavior such as manipulation, which would in turn subverts its credibility as a public policy. Is thus nudging anyhow different from manipulation? There are several strategies for answering alike doubts.
First, nudge enthusiast may try to reconceptualize this concept, or add more conceptual substance that would distinguish it. Secondly, one may leave the nudge concept as it is and rather demand precise definition of manipulation instead. If a reasonable concept of manipulation would be laid out, then nudges would be all the remaining influences not falling under the conceptual skeleton of manipulation. Since the concept of manipulation is commonly associated with freedom-subverting mechanisms, the non-manipulativeness build into the concept itself would ensure nudging’s moral innocence. Thirdly, perhaps one must not do first nor second, but rather come up with an argument why it is sometimes permissible to interfere with person’s freedom with objectionable (e.g. freedom-undermining) means such as manipulation. If sufficiently strong reasons would be found, then nudging, regardless of the fact whether it is being manipulative or not, would be permissible in instances when those reasons would apply. Which one is better? Following subsection aims to discuss and evaluate an attempt to use the first strategy to vindicate the reputation of nudging as legitimate public policy.

STRATEGY #1: Reconceptualization of nudging

Yashar Saghai (2013) tried to ´salvage´ nudging via first strategy. He suggested that ´influence’s control’ and ´subject’s effort´ should be the qualities sufficient for delimiting nudging from other sorts of influences. His salvation argument mirrors the logic of the original nudge definition, since he argues that the quality of nudge’s influence can be tracked on the continuum between fully non-controlling and fully controlling, while the control of an influence is being assessed against the amount of effort the subject of an influence needs to develop in order not to be steered by the influence.

“A’s influence to get B to µ is substantially noncontrolling when B could easily not µ if she did not want to µ. [...] Resistibility. A’s influence is easily resistible if B is able to effortlessly oppose the pressure to get her to µ if she does not want to µ.” (ibid., pp. 488-9)
Relatively effortless opposition is real if (1) B is capable of becoming aware of nudge’s influence over him, (2) if B is capable of canceling the effect of an influence over him, and (3) if the influence or its circumstances otherwise undermine the exercise of the beforementioned capabilities. Nudges in Saghai’s words thus need to be substantially noncontrolling to avoid controversies, while their control is to be assessed against the above conditions.

His reconceptualization of the nudge nonetheless didn’t dismantle all doubts for at least two reasons. Firstly, Saghai did not address the issue of the missing argument for a particular threshold. In fact, he instead introduced another continuums (not only that of control but also that within it – the effort needed to resist) on which similar threshold identifying the tipping point between acceptable and unacceptable level of influence is needed. It is in other words not clear beyond what level of effort needed against a nudge does it become problematic. When exactly reframed message becomes too hard to oppose? When does resisting the effects of redesigning cafeteria interior exceed the bearable level of effort?

Apart from the practical complexity of observing and measuring these phenomena, determining the limits in question seems to yield severe practical constrains on consequent usage of nudging as population-wide public policy. Different people may as matter of fact vary in the degree to which they can exercise alike capacities vis-à-vis identical kind of influence. Think of the procrastination and the varying abilities to tackle it among the people you know. Some of your close ones may not suffer from this tendency almost at all because it is relatively easy for them to exercise their will to accomplish their tasks. Others, however, may seem like pathological procrastinators precisely because the very same exercise is much harder for them. In a way, thus, some of these people are naturally capable of resisting the procrastination tendency relatively effortlessly, others need a lot of
effort for achieving the same. Now how do we apply a nudge capitalizing on the procrastination tendencies, such as e.g. some default nudges may, as a population-wide policy? It is highly doubtful that we would be able to single out a population with no, seemingly pathological procrastinators what so ever. Applying it despite the presence of these procrastinators would on the other hand render the nudge freedom-undermining vis-à-vis them.

Secondly, Saghai speaks of nudging’s influences as if they were always a form of pressure from underneath of which one struggles to exit (Wilkinson, 2013). His substantial non-control criterion in other words conditions nudging not to be hard to exit. But as Wilkinson (ibid., p. 486) fittingly puts it, “many influences involve not effort but failures to notice, as with anchoring, defaults, framing and so”. This points to the fact that it is one thing to actually have capacity to become aware or to cancel nudge’s influence, yet it is completely another thing to use it. Failing to become aware or to cancel nudge in other words must not have anything to do with the capacity for so doing.

Moreover, Saghai’s reconceptualization stipulates that nudge becomes problematic if its subject cannot exit its steering influence when the aimed decision is not in accordance with nudgee’s preferences. But what if steering influence includes an attempt to change subject’s preferences, e.g. by persuasion or induction, in order to (at least momentarily) suspend what otherwise could have been a significant barrier to its effectiveness? Or what about nudges that attempt to steer people towards what they actually want yet usually fail to do, as paternalism does? Does the non-control condition do any work in distinguishing non-manipulative nudges from the manipulative ones, or is it after all the consistency with subject’s preferences that distinguishes them?

It seems therefore that the first strategy for vindication of nudging’s credibility as public policy raises more questions and practical problems than answers, at least in Saghai’s way
of using it. Perhaps due fear of this, the authors of the nudge themselves inclined to the second and third strategy for vindication of nudging’s credibility as public policy.

“[… ] Imagine a graphic warning about opioid addiction, or about the use of cell phones while driving, that would create immediate fear of revulsion, or intensely engage people’s emotions; it might be objected that nudges of this kind count as a form of manipulation. To know whether they do, we need a definition of manipulation” (Sunstein, 2017, p.6) [emphasis added by author].

Although this second strategy looks simpler and easier to accommodate, the concept of manipulation is similarly disputed concept to that of nudging, which inevitably complicates salvaging the nudging by a reasonable notion of it. Sunstein (ibid.) moreover used very few words to describe his notion of manipulation. He characterizes it as an activity that tries to disrupt “people’s capacity for rational deliberation”. End of story.

This notion of manipulation exhibits an interesting similarity with Saghai’s reconceptualized notion of nudging in a sense that both notions conceptually identify people’s core capacities for determining the course of one’s life. While Sunstein refers to the capacity for rational deliberation, a trait that can also be translated as a natural capacity for formation, revision and rational pursuance of authentically own conception of a good life¹, Saghai (2013b) (re)frames the core traits in terms of goal-conflicts recognition and resolution capabilities.

Whether you start with the reconceptualization of nudging or whether you start from manipulation, therefore, both strategies seem to aim at similar goal: singling out the traits with operations of which should not be anyhow interfered with, because of their detrimental, path-setting purpose for life of every human being.

¹ Ref. To John Rawls’s notion of it
The main reason for identifying these traits seems to be not that hard to find: subverting them would amount to a moral offence or disrespect to a worth of free an equal human being, since the capacities are framed as vital for determining the course of one´s life. One may thus say that manipulation in a way makes the manipulated agent object rather than subject. Freedom is especially important aspect in this, for it brings attention to an agency. Nudging and manipulation are both intuitively perceived as external influences that strive to change a behavior thus the path of one´s agency, but for those concerned about these kinds of influences, agency´s path is seemingly prima facie driven by impulses from within of the agents, or at least by inwardly justified impulses perhaps originating elsewhere. This might be good starting point for the discussion of the third strategy but let’s not rush.

Since a manipulation is contested because it ostensibly by definition aims to ´trick´ or ´substitute´ such inward agenda-setters of agency, some nudges (e.g. some so called opaque nudges) receive negative attitudes because they imitate the functioning of manipulation – they work because their operation is being hardly or often never noticed. It is precisely because of this imitating operationalization of some nudges that nudging is, especially when compared with manipulation, being suspected of wronging person. But are such nudges indeed manipulative or are they not? How do we actually recognize manipulation?

STRATEGY #2: Starting from the manipulation

In this sub-section, i aim to discuss whether defining the concept of manipulation prior to that of nudging helps us conclusively vindicate the reputation of nudging as legitimate public policy strategy. It would do so if the definition would provide us with enough clues to sufficiently clearly distinguish ´innocent´ nudging from ´guilty´ manipulation. As a stepping stone for my discussion, I’ll use similarly motivated enquiry of this topic by T.M. Wilkinson
(2013), mainly because it itself discusses the matter very well yet at times misrepresents manipulation’s true character.

**How not to start**

Wilkinson (ibid., pp. 345 -7) claimed that the character of manipulation could be described by 4 features: (a) pervasive method of influence on individual’s decision-making processes; (b) intentionality; (c) successful alteration of behavior; and (d) prima facie wrongness. Wilkinson nonetheless concludes that he’s unable to conclusively draw a line delimiting wronging manipulation from non-wronging nudges solely on the basis of the concept that he proposes. This is so because Wilkinson’s notion of manipulation incorporates a few obsolete features.

Let’s start with the last feature. The inherently wrong forms of manipulation represent only one of the possible subclasses of manipulation that could be naturally observed in our lives, and thus that definition of manipulations should not contain this feature because it introduces a confusing frame into the discussion.

Think of rearranging your apartment, that picture of you from that vacation sure looks better on the wall of your hallway than on the shelf where it was until recently. By relocating it from the shelf to the hallway you’ve manipulated with picture’s emplacement.

Think of sending money from a saving account to a regular account in order to purchase something. Think of bending paper plane’s wings a bit more after first short test flight.

Think of playing in a puppet theatre or blowing into the sails of a toy yacht to manipulate with its course. Aren’t these all intuitive scenarios in which we can talk of manipulation? Moreover, aren’t these all morally unproblematic scenarios of manipulation? They may and may not be morally problematic. We would most likely judge against an effect their consequences may bring about into individual’s life. The point is that while we can’t say
whether these are wrong or okay actions at the moment, we can say that they are cases of manipulation.

Both physical and psychological manipulations are forms of manipulation. An adequate definition of manipulation cannot deny that either cases are forms of manipulation. Both types of manipulation may but must not necessarily be morally relevant. But if this is so, it cannot be true that manipulations are by definition prima facie objectionable because as it is intuitively apparent from the above examples, some manipulations are intuitively non-problematic at the first sight. This intuition seems to be often stimulated by nudging as well.

Incorporating the (d) in the definition of manipulation is thus too hasty and unduly diverting our attention only towards some manipulations, instead of looking for reasons why a manipulation is objectionable act in a first place. The (d) condition frames the starting point as if we were obliged to first find reasons why it is not objectionable to act manipulatively.

Now remember that this is important for vindication of nudging’s reputation because nudging is being charged of being a form of manipulation. If manipulation should not be considered automatically prima facie wrong, neither should nudging, if it is form of manipulation. Starting from the non-moralized conception of manipulation thus shifts the discussion from ‘why we shouldn’t consider manipulation wrong’ to ‘when we should consider manipulation wrong’, which in turn gives some leeway to nudging as well.

Wilkinson’s inability to conclusively distinguishing nudges from manipulation is at least in part a consequence of incorporating the (d) conditions into the definition of manipulation. The (d) condition primes him to look for wronging influences, but once he’s confronted with influences on one’s autonomous agency that at the first sight act manipulatively but trigger our intuitions about their moral innocence, e.g. due their apparent ‘mildness’ in
comparison to other manipulative influences that seem comparatively coarser, he has to admit inconclusiveness of his endeavor.

And the reason is simple. Since Wilkinson is starting from the presumption that all manipulations are prima facie objectionable, he considers nudges to be different in character from manipulations because, as he himself puts it, there are well-known examples of nudges that are both, apparently manipulative (hence prima facie objectionable) and apparently non-manipulative (vice versa):

“Since lying and subliminal advertising are not excluded by the definition of a nudge, we can say that some nudges could be manipulative. Thaler and Sunstein must agree; [but] some cases are obviously not manipulative” (ibid., p.347)

Since in his view manipulations cannot be prima facie unobjectionable, nudges cannot be equivalents of manipulations, for our intuition often suggests that they can be prima facie unobjectionable. It is nonetheless okay to talk about nudges and nudging entirely as about form of manipulation. If it is reasonable to discount Wilkinson´s (d) condition, we can safely consider nudges to be kosher even if they are manipulations for the burden of proof is on the side of those who want to condemn manipulation as inherently problematic.

In short, the tendency to hastily condemn manipulation as prima facie wrong diverts our attention from the real extent and naturalness of the concept of manipulation. Prima facie unobjectionable forms of manipulation do exist, and it is absolutely alright.

My little excursion to the above examples of at the first sight unobjectionable cases of manipulation has in my view similar purpose as the Thaler’s and Sunstein’s original claim about the inevitability of the choice architecting. Simply said, they claim that whether we want it or not, our environment influences the way we perceive our options. By opening the discussion around the importance of choice architecture, they on the one hand expose the tendency of academia to focus the discussion of manipulation predominantly on the
psycho-cognitive dimension of it and thus only on the ways an influence approaches our inner cognitive and decisional ‘strings’ or ‘levers’, often disregarding its origin or source. Behavioral economists draw attention to our environment, to its physical form and to the psycho-cognitive dimension of physical manipulation with our environment.

But since physical manipulation with our environment may be morally unproblematic and since the design of the environment inevitably affects the way we perceive our options, the choice architecture statement on the other hand also acts as an ice-breaker and carries out an implicit claim that we should be conscious about our seemingly common tendency to hastily condemn any manipulation as guilty of wronging us. Again, prima facie unobjectionable forms of manipulation do exist, and it is absolutely alright. The definition of manipulation should be encompassing both, prima facie objectionable and unobjectionable manipulations. I’ll call a definition that does so the naturalistic notion of manipulation.

Their statement nevertheless implies wide notion of manipulations in a sense that it makes manipulation out of too many things, even contra-intuitively. Yes, it may appear so as if suddenly too many things could be seen as manipulative in a way. But that should not be concern at all. We do not need to be concerned whether particular alignment of trees could be framed as manipulative or whether friendships are in this light forms of (and often unconscious) manipulative relationships. We need to be concerned only with the ‘bad manipulation’ and thus with features or reasons distinguishing this subclass of manipulation. Wilkinson’s (d) condition seems to be unqualified for this job.

So how does the other parts of his manipulation fair in helping us distinguish ‘guilty’ manipulations from ‘innocent’ nudges? What difference e.g. the intentionality, another of Wilkinson’s ostensibly defining feature of manipulation, makes in defining manipulation? As Wilkinson (ibid., p.345) himself puts it, the intention should be somewhat relevant for
distinguishing ´guilty´ manipulation from ´innocent´ nudge: ´[...] Manipulation involves an intentional actor. This intention condition is of great importance in explaining why manipulation is wrong.

Intentionality´s importance seems to nonetheless end at the point of narrowing down the pool of possible cases of ´guilty´ manipulations by expunging the clearly innocent cases. As the firstly presented intuition about the manipulation suggests, even tornados could be (strictly speaking) considered manipulative. Yet we cannot ascribe intention to tornados, as we cannot do so to any recurring natural phenomena, because they don´t posses the ability to form an intention. Doing wrong seems to be an exclusive domain of intentional and moral actors. That’s why when talking about repositioning of a picture to the hallway (by a person, not tornado), we cannot exclude the possibility of action´s manipulativeness because the change could have occured as part of an intentional action by a moral actor. Simply said, it could but must not have been manipulation.

Once the upholding of intentionality condition indicates that ´guilty´ manipulation is possible because it has been done by a moral actor capable of forming intention, it can’t be of no more help. Recall that the masterquestion of my endeavor demands conceptual detachement of at the first sight ´innocent´ nudging from a ´guilty´ manipulation. Both are likewise intentional actions trying to steer people towards particular direction, yet none none is less intentional than the other. What helps to distinguish the moral significance of an utmostly mild nudge (such as e.g. asking a person whether he ones one of the predefined mixtures in his sandwich or whether he wants to build his own combination) from a coarse manipulation (e.g. flooding someone with red herrings to divert him from particular decision) is thus not the intention, because both are intentional actions. It is the quality of their methodology – in this case apparently their intensity (mildness vs. coarseness) - with which they are both intentionally applied over a target. Intention seems
thus to be of no use for answering the ‘masterquestion’ about the differences between ‘guilty’ manipulations and ‘innocent’ nudges.

There seems to be another reason why to believe that Wilkinson overstates its importance when stipulating that the ‘[..] intention condition is of great importance in explaining why manipulation is wrong’ (ibid., p.345). Consider also following two scenarios of sincerely paternalistic manipulations. Motivated by a concern for friend’s well-being and out of paternalistic reasons, Aaron sincerely wants to prevent his friend Bob from harming Chad, for if Boris would have done his deed, he could be facing imprisonment. So Aaron decides to use some underhand method of manipulation as his tool and his resourcefulness allows him to succeed in manipulating his friend Boris. Both the underhand methods and the intention would occur, alongside the success. Wilkinson would thus, in my view, conclude that Aaron wronged Boris despite his paternalistic motivation, partly because of his intention.

But then there’s the second scenario, which is almost identical except one thing: Aaron picks his tool for paternalizing his friend at random, almost alike to the times we fumble in a drawer for a spoon without looking. He misguides Boris by saying Chad’s out of town (despite the fact that he’s not) – the first thing that came to his mind when Boris asked him about Chad. He simply wants to help his friend, yet he has no particular intention to use particular mean. In this scenario, it seems intuitively harder to justify the presence of an intention to manipulate for it seems as if Aaron has no intention what so ever about what tool to use for his cause. In light of Wilkinson’s characterization of manipulation, this would make sufficient difference to consider the scenario 2 as a case of non-problematic use of manipulation. Yet it is quite a big difference and it seems intuitive to still consider the use of manipulation in the second scenario somehow dubious. At minimum, a specification seems

---

2 “[..] concerns about manipulation apply to non-paternalistic nudges as well as paternalistic ones” (Wilkinson, p.346). Lets moreover presume, for the sake of this argument, that the amount of diverted harm would not justify the damage done to autonomy.
to be needed as to what kind of intention needs to be present, or more precisely, how exactly should be the intentionality condition understood. Does it needs to be directly an intention to use manipulation? Or does it suffice when manipulation is part of an intentional action while not necessarily accompanied by specific intention for manipulation’s particular use?

Last but not least, if my point from the beginning of this section holds and thus manipulation should not be considered automatically prima facie wrong, having an intention to manipulate does not automatically mean having an intention to wrong somebody. Hence the intentionality condition cannot help us distinguish the innocent nudges from the guilty manipulations either.

The only time I can imagine the intention’s helpfulness in distinguishing ‘innocent’ nudges from ‘guilty’ manipulation is in a scenario, where we already know the difference between these two techniques a priori. If we would consider intention to be inevitably connected with some actions, i.e. as intention to something, only then would it be helpful identifier for we would either have intention to nudge or intention to manipulate. We would nonetheless need to know a priori what’s the difference between these two actions, if any.

This subsection showed how not to start vindicating nudging’s reputation by defining the manipulation first. It was argued that not intention nor prima facie wrongness provide us with correct cues for distinguishing ‘innocent’ nudging from ‘guilty’ manipulation. The former at max narrows down the pool of influences on our decision-making that could be possibly wrong, the latter unreasonably narrows down the notion of manipulation only to one subclass of manipulations. The next subsection will focus on what might be a useful starting point for indicated endeavor, and why.
Where to start

So as with the intuition from the beginning that manipulation is actually prima facie neutral and more natural concept, there are other intuitions about the nature of manipulation that may help us better portray manipulation’s character in contrast with Wilkinson’s try. If reasonable, they might be a safe start for vindicating nudging’s reputation by defining manipulation first.

It was already pointed out that manipulation seems to be a scalable phenomenon, thus e.g. something that can be of mild but also of very strong intensity. This intuition is correlative to the common tendency to denominate manipulation as a form of influence.

Moreover, given the fact that most of the academic discussion focuses on cases of manipulation where A manipulates B, it seems natural to consider it a relational phenomenon. This encompasses two things. Firstly, there’s a relationship between an initiator and a victim or recipient of manipulation. Secondly, alongside the initiator and recipient of manipulation, there must be a source and a target of the manipulative influence, both independent in their nature from the persons in the relationship. There simply cannot be a manipulation if there’s not something or someone that has been manipulated, as well as someone or something that manipulated.

These intuitions about manipulation are present in Wilkinson’s notion of the manipulation as well. Its first defining feature, the pervasive influence on the way an individual see his options, is a nice demonstration of the difference between the target and the subject (recipient, or victim) of manipulation. It is an individual who’s in effect subject (hence victim) of manipulation, yet it is his decision-making capacities (the channels through which one sees his options) that manipulation manipulate with.

Now consider e.g. a political advertisement that tries to stimulate strong negative emotion and associate it with the consequences of voting for political opposition. Warning of
disease from migrants, or of imminent violence after refugees enter country’s territory, or suggestion of a betrayal and endangering of close ones in case of voting for opposition, all supported by very suggestive graphics. The advertiser is the initiator, while the emotional content is the source of manipulation. Voters are recipients and potential victims of this campaign, and their emotional levers through which they partly perceive the value of their options are manipulation’s target.

Last but not least, in order to finish the portrayal of manipulation, the possible first step in distinguishing ‘innocent’ nudge from ‘guilty’ manipulation, we need to understand the quality of this relational phenomenon in both descriptive and normative sense. In other words, if its defining feature is a pervasive method of influence on the ways we see our options, we need to know what the defining quality of pervasion is (descriptively), as well as why and when this quality deserves to be ethically condemned (normatively). In Wilkinson’s terminology, I opt for what he called the general approach in defining the pervasiveness of manipulative methods, i.e. for finding a general rule against which we will assess all the cases.

Defining manipulation as prima facie bad presupposed some normative rule, yet this section showed that such presupposition is too hasty and unreasonable. Whatever influence operated similarly to the manipulation was so far automatically likewise wrong. Introducing the concept of nudging changed this.

Recall that a nudge is any aspect of choice architecture (i.e. of context or environment in which we decide) that successfully influences individual’s behavior, given certain constraints. Together with what was said so far, it has become apparent that nudging operates similarly to manipulation. A choice architect (initiator) deliberately influence an aspect of someone’s choice architecture (source) so it successfully alters individual’s (recipient or victim) behavior (target), which is of course presumed to be a result of
deliberate decision-making (target) in given context. But since some nudges highlighted the fact that there are great variations in intensity with which a source of at the first sight manipulative influence may affect its target, the question became again relevant. The discussion of this normative rule won’t be addressed here. It will nonetheless offer a possible descriptive, general rule.

In particular, manipulative relationship could be considered a relationship of power, hence a relationship where A applies or demonstrates in some way its force or power on, against or over B.

This descriptive view of manipulation aligns nicely with both, the cases of physical and psychological manipulation from the literature. For the former, recall the examples from the beginning of this section, the example of applying power to move the strings controlling a puppet in a theatre, the example of applying force via gentle blow to the sails of a toy yacht in order to change its course. For the latter, Robert Noggle’s (1996) instructive account of manipulation as a control via ‘adjusting one’s psychological levers’ out of their ideal setup seems to be translatable to a power-relationship between the initiator, the source and the target as well. One has to convey some form of force over other’s levers (by which he meant primarily 3 things: desires, beliefs and emotions) to adjust them. Breiker’s (2004) metaphor about manipulation as an act of ‘pulling one’s strings’ similarly intuitively aligns with this frame.

How to continue from here?
Framing manipulation in terms of power application directly point to the possible reasons due which it could and so commonly is condemned, hence to the normative clues that may ultimately help us distinguish ‘guilty’ subclass of manipulation from the non-guilty one. Here are some suggestions as to where the discussion of the missing normative general rule might head.
Manipulation characterized as an application of power over (on, or against) a target implies taking a temporal control over a target, or at least over some of its crucial parts (the levers or strings). Seeing the target from the Kantian perspective, i.e. as an agent who’s to be seen as end in itself, where having the sole control over the levers or strings play central role for such position³, manipulation would be without a doubt wronging its victim.

This is where Wilkinson’s notion of manipulation is right, the power control can be proven only if the alteration of behavior by manipulation is successful, thus, when manipulator indeed sets a course for its target. Otherwise, the control (or the power) would be hardly visible. Just imagine a puppet player who tries but fails to move the body parts of his puppet despite having access to its strings.

When manipulating, a manipulator treats its target as a mere tool for furthering own interests. But even if manipulator would sincerely manipulate to further target’s interests, it wouldn’t change the fact that the target would be the puppet under manipulator’s control and that there’s something wrong about this, especially when the manipulator ought to by default enjoy equal rights and privileges as its target. The wrongness felt in these cases is perhaps rooted in the fact that the actions of the manipulated actor cease to manifest his free will. It appears like we have arrived to the one of possible starting points of manipulation’s wrongness.

Another reason why manipulation (as power relationship) could be condemned manifests itself in the stipulation that the manipulator treats its target as a mere tool without special mandate for such action. Such treatment may be charged of being disrespectful to a moral agent, for it approaches him as if he was not able to form, or to revise or to pursue own conception of the good, the skill that especially in liberal political philosophy (Rawls, 1995; ³ Unless special reasons allowing such control appear, as e.g. from a voluntary consent of such person.)
Quong, 2013) equally co-defines the status of every moral agent. Just imagine two alike adults and one treating the other as a child to picture the point of this charge to manipulation.

It is however not clear whether these two interpretations of manipulation overcome the continuum problem identified during the analysis of the strategy #1. Manipulation as a relationship of power provides no explicit clue for identifying the border after which an innocent nudge or manipulation becomes a guilty one, hence when exactly the interference becomes too coarse. The normative side of the story does specify why the power-relationship might be a problem, but again, it doesn’t describe when exactly (at what intensity) it becomes wronging.

What the power-relationship notion of manipulation does is that it comprehensively warns us when a manipulation can possibly occur. In Bachrach’s and Baratz’s (1962) view, power-relationship manifests itself in decision-making processes not only when A makes observably B to decide in a way B wouldn’t otherwise do as the Robert Dahl initially suggested, but also when A makes B observably not to decide at all. In Lukes’s view (2005), moreover, power may be manifested also in a latent and unobservable conflicts, thus even as A’s ability to prevent B from even considering to decide in particular way in a first place, thus as a sort of agenda-setting power.

The power interpretation is very explicit and straightforward about this: it is bad because it manifests someone’s power on, against or over us, which either offends our freedom or disrespects us. The problem as to how to measure it and where to set the borderline of innocence nevertheless still remains.

---

4 For John Rawls (1995, p.19), citizens are in part equal in a sense that they all could be equally rendered as having fundamental interest in developing their two moral powers, which in turn qualify them as eligible moral actors for a communal life.
All the considerations discussed in this section lead me to the conclusion that morally relevant manipulation is best conceptually portrayed as an intentional power-relationship that manifests itself among various moral actors to the extent that manipulation’s victim is either successfully made to decide in a particular way, or to the extent that its target is successfully prevented from deciding, or to the extent that manipulation’s target is successfully prevented from actually considering particular options for a decision, while all these scenarios are being contrasted with an ex ante prediction based on a scenario where no intentional power-relationship would take place.

Stated otherwise, manipulation occurs to the extent that an option is successfully translated into, not translated into, or not even considered for a decision; as a result of intentional power-application of moral actor(s) on, over or against other moral actor(s), for identification of which we need an ex ante prediction about the ways the decision could have gone were there no intentional power-relationship.

Starting by the definition of manipulation for vindicating nudging’s reputation as legitimate public policy brings us as far as to the realization that nudging and manipulation might belong to same conceptual family. We can technically speak of both as of manipulations, yet since it is reasonable to define the family of manipulation as encompassing both morally objectionable and unobjectionable manipulations, we have no reason to automatically morally condemn nudging operating manipulatively. One has to make separate argument as to when the manipulation and thus nudging becomes morally objectionable. Nudging may become morally objectionable either when we it becomes disrespectful to victim, or when it makes its victim unfree, in either case it is because of too much power a manipulator has on, over or against the victim, given that manipulation is scalable phenomenon. As with the strategy 1, neither the strategy 2 ultimately helps
distinguish ´innocent´ nudges from ´guilty´ manipulation and hence once and for all vindicate nudging´s reputation as legitimate public policy.

CONCLUSION

Nudging represents unique combination of knowledge from many well-established scientific disciplines. It is unique especially due its clear association with particular family of ethical doctrines that value freedom, as well as due its roots in newly formed science that unceasingly uses the economical model of rational agent as a rebound bridge for its inferences.

If nudging aspires to be used legitimate public policy, it is reasonable to use it on top of solid amount of quality hard data. Especially in modern democracies, policy-makers might appreciate nudging for its softness with respect to the amount of interference it brings about to one´s personal freedom.

Considering Thaler´s and Sunstein´s (2008, p.6) original definition of the nudging, some obvious conceptual overlaps between conceptual skeletons of nudging and manipulation exist. If nudging should be understood as intentional designing of choice architecture, thus as mindful playing with any aspect that “alter´s people behavior in predictable way”, where mindfulness represents all the constrains Thaler and Sunstein demanded as part of their libertarian-paternalistic project (e.g. the easiness and cheapness to avoid the targeted behavior, but also transparency and publicity, the conditions not discussed in this paper), and if the naturalistic notion of manipulation presented is credible, then nudging is nothing less than mindful form of manipulation connected with particular ethical doctrine. It is moreover unreasonable to consider it prima facie wrong.

It appears as if most of the fuzziness surrounding the concept of nudging boils down to its ethical dimension and to implications of this dimension. Identifying the precise borderline
after which nudging exceeds its mild and unproblematic level of manipulativeness seemingly very much depends on the particular ethical notion of agency to which a thinker resorts. The third strategy for vindicating the legitimacy of nudging as a public policy may thus yield the most conclusive answers to the debate discussed in this paper, because it mitigates the need for identifying of the precise borderline. The complexity of this strategy nonetheless begs for completely separate analysis that exceeds the limits of this paper.

Although this paper didn’t conclusively answer all the puzzles surrounding the concept of nudging, it shed some new light on its conceptual nature vis-à-vis its conceptual cousin, manipulation. It moreover suggested by which routes the ultimate vindication of nudging’s reputation as legitimate public policy may try to proceed.
Bibliography


