Nudge and the Manipulation of Choice

A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy

Pelle Guldborg Hansen and Andreas Maaløe Jespersen***

In Nudge (2008) Richard Thaler and Cass Sunstein suggested that public policy-makers arrange decision-making contexts in ways to promote behaviour change in the interest of individual citizens as well as that of society. However, in the public sphere and Academia alike widespread discussions have appeared concerning the public acceptability of nudge-based behavioural policy. Thaler and Sunstein's own position is that the anti-nudge position is a literal non-starter, because citizens are always influenced by the decision making context anyway, and nudging is liberty preserving and acceptable if guided by Libertarian Paternalism and Rawls' publicity principle. A persistent and central tenet in the criticism disputing the acceptability of the approach is that nudging works by manipulating citizens' choices. In this paper, we argue that both lines of argumentation are seriously flawed. We show how the anti-nudge position is not a literal non-starter due to the responsibilities that accrue on policy-makers by the intentional intervention in citizens' life, how nudging is not essentially liberty preserving and why the approach is not necessarily acceptable even if satisfying Rawls' publicity principle. We then use the psychological dual process theory underlying the approach as well as an epistemic transparency criterion identified by Thaler and Sunstein themselves to show that nudging is not necessarily about "manipulation", nor necessarily about influencing "choice". The result is a framework identifying four types of nudges that may be used to provide a central component for more nuanced normative considerations as well as a basis for policy recommendations.

I. Introduction

In the last three decades, advances in behavioural economics and psychology have revealed how our decision-making and behaviour are systematically biased by the interplay of psychological with what ought to be, from the perspective of rationality, irrelevant features of the decision-making context. In general, these behavioural insights teach us how decision-making contexts may systematically lead us to fail in acting on our well-informed intentions or achieve our preferred ends. In the area of public policy-making, particularly, such advances teach us how neglecting these insights may be responsible for failures of public policy to reach intended effects, and why paying more attention to them seems likely to provide a key to dealing effectively with important societal challenges such as global-warming, obesity epidemics, and poor economic decision-making.

The seminal book that brought the idea to a broader audience was *Nudge: Improving Decisions*

* Director of The Initiative for Science, Society & Policy (ISSP); Institute for Marketing & Management, University of Southern Denmark.

** The Initiative for Science, Society & Policy (ISSP).

The authors would like to thank two anonymous reviewers at EJRR as well as Prof. Robert Sugden (UEA), Prof. Bent Greve (RUC) and John Parkinson (Bangor University) and colleagues from the Institute for Marketing and Management (SDU) who all have provided helpful comments and question.

About Health, Wealth and Happiness,[1] written by behavioural economist Richard Thaler and law scholar Cass R. Sunstein. In their book, Thaler and Sunstein suggested that public policy-makers and other choice architects arrange decision-making contexts in ways to promote behaviour that is in our own, as well as society's general interests. The proclaimed advantage in doing this is that public policy-makers might influence – in a cheap and effective way – our everyday choices and behaviours without recourse to injunctions or fiddling with incentives. That is, "nudging" seems to offer policy-makers an effective way to influence citizens' behaviour without restricting freedom of choice, imposing new taxations, or tax-reliefs. Thaler and Sunstein have coined the seemingly oxymoron term, "Libertarian Paternalism",[2] which characterizes the attractive regulation paradigm that arises out of the "nudge approach" to behavioural change in public policy-making, when enacted to serve the interests of the citizens as these are judged by themselves.

Four years later, the nudge approach has achieved widespread recognition in two of the largest Western democracies. Until recently, Sunstein was an advisor on regulatory affairs for US President Barack Obama. Thaler is an advisor for UK Prime Minister David Cameron's Behavioural Insights Team (BIT), popularly referred to as the "Nudge-unit." These efforts have led to the production of a series of reports and discussion papers aimed at public policy-makers. The UK Institute of Government has published the MINDSPACE report in 2010,[3] drawing heavily on the nudge approach; through the UK Cabinet Office, BIT has published the first, "Behavioural Insights Team Annual Update 2010–2011";[4] the Centre for Strategic Analysis for the Prime Minister of France has published a report on "Green Nudges";[5] and the European Commission has published the report, "Nudging lifestyles for better health outcomes".[6]

The impact extends beyond mere speculations. In the US, the nudge approach has been used to design the 401(K) pension scheme,[7] suggested as a tax refund system[8] and recently, to suggest a controversial ban on super-size sodas in New York City.[9] In the UK, large-scale nudge experiments have been carried out by BIT to improve compliance in tax reporting[10] and to lower alcohol consumption among youth.[11] Also, both in the US and the UK, nudging has inspired the implementation of prompted choice in registering for organ donation. These attempts at "nudging" in public policy-making have been of varying success, but overall seem promising.[12]

However, this new undertaking in behavioural change has not always been met with enthusiasm. Both academics and public commentators have leveled harsh criticism – political, practical, and ethical – against the approach. In the UK, the libertarian blog Spiked featuring a series of liberal academics has even gone as far as declaring "war on nudge".[13]

---

1 Richard Thaler and Cass Sunstein, *Nudge – Improving Decisions about Health, Wealth and Happiness* (New Haven, CT: Yale University Press 2008)

2 Richard H. Thaler and Cass R. Sunstein, "Libertarian Paternalism is not an oxymoron", 70 *The University of Chicago Law Review* (2003), pp. 1159–1202.

3 Paul Dolan, Michael Hallsworth, David Halpern et al., "MINDSPACE – Influencing behaviour through public policy. Institute for Government", 2010, available on the Internet at: <http://www.instituteforgovernment.org.uk/publications/mindspace> (last accessed on 09 January 2013)

4 Cabinet Office Behvioural Insights Team. Behavioural Insights Team Annual Update 2010–2011, 2011, available on the Internet at: <http://www.cabinetoffice.gov.uk/resource-library/behavioural-insight-team-annual-update> (last accessed on 09 January 2013).

5 Oliver Oullier and Sarah Sauneron, "'Green Nudges' new incentives for ecological behaviour", 2011, available on the Internet at: <http://www.strategie.gouv.fr/en/content/policy-brief-216-nudges-green-new-incentives-green-Behaviour-march-2011> (last accessed on 09 January 2013).

6 Brigitte Piniewski, Cristiano Codagnone and David Osimo, "Nudging lifestyles for better healtlh outcomes: Crowdsourced data and persuasive technologies for behavioural change", 2011, available on the Internet at: <http://ipts.jrc.ec.europa.eu/publications/pub.cfm?id=4219> (last accessed on 09 January 2013)

7 Edmund L. Andrews, "Obama Outlines Retirement Initiatives 2009", available on the Internet at: <http://www.nytimes.com/2009/09/06/us/politics/06address.html?_r=1> (last accessed on 09 January 2013).

8 James Surowiecki, "A Smarter Stimulus", 2009, available on the Internet at: <http://www.newyorker.com/talk/financial/2009/01/26/090126ta_talk_surowiecki> (last accessed on 09 January 2013).

9 Michael M. Grynbaum, "New York Plans to Ban Sale of Big Sizes of Sugary Drinks", 2012, available on the Internet at: <http://www.nytimes.com/2012/05/31/nyregion/bloomberg-plans-a-ban-on-large-sugared-drinks.html?pagewanted=all> (last accessed on 09 January 2013).

10 Cabinet Office – Behavioural Insights Team, "Applying behavioural insights to reduce fraud, error and debt", 2011 available on the Internet at: <http://www.cabinetoffice.gov.uk/resource-library/behavioural-insights-team-paper-fraud-error-and-debt> (last accessed 09 January 2013).

11 Cabinet Office – Behavioural Insights Team, "Applying behavioural insights to health", 2011, available on the Internet at "http://www.cabinetoffice.gov.uk/resource-library/applying-behavioural-insight-health" (last accessed 09 January 2013).

12 Nudge theory trials 'are working' say officials, 2012, available on the Internet at: <http://www.bbc.co.uk/news/uk-politics-16943729> (last accessed on 09 January 2013).

13 Brendan O´Neill, "A message to the illiberal Nudge Industry: Push off", 2010, available at: <http://www.spiked-online.com/site/article/9840/> (last accessed on 09 January 2013).

A persistent and central tenet in the political and normative criticism has been the claim that nudging works by "manipulating people's choices".[14] [15]

This claim is fundamental to various other criticisms of the approach. For instance, arguments such as: that Libertarian Paternalism is an oxymoron since the nudge doctrine is merely paternalism in disguise;[16, 17] that nudging is a return to present-day behaviourism;[18] "that the psychological mechanisms that are exploited [...] work best in the dark";[19] that effects of nudges are likely to disappear if nudges become transparent;[20] that nudging can encourage abuse of power by technocrats;[21, 22] and finally, that nudging impairs our autonomy and our ability to make moral choices for ourselves.[23]

The claim that nudging works by manipulating people's choices, then, is not only a problem because it seems to make the approach incompatible with public policy-making in a modern democracy. Indeed, state manipulation with the choices of citizens appears to be at odds with the democratic ideals of free exercise of choice, deliberation, and public dialogue. It is also a problem because it provides a fundamental premise for an array of other criticisms that challenge the legitimacy and efficacy of adopting the nudge approach in public policy-making.

Thaler and Sunstein seem to admit as much: nudging is a manipulation of choices.[24] But they generally dismiss the above criticisms with a three-pronged defense. First, they claim that our choices are always being influenced by the context of choice, whether we like it or not, making the anti-nudge position a literal non-starter.[25] This is backed by a second claim, which is that because nudges work without limiting the original set of choices, or without fiddling with existing incentives, citizens remain free to choose otherwise. When even these two claims do not put our worries to rest, the final claim is that, if guided by libertarian paternalism and a Rawlsian publicity principle, the relevant political and normative concerns are met.[26]

The position of Thaler and Sunstein can thus be summarized quite simply: because our choices are always influenced by the decision-making context, and because such influence is often manipulated by far more intrusive or subtle measures – taxation, regulation, marketing, etc. – nudging is an admissible approach to behaviour change in public policy-making. That is, as long as the ends nudged toward are consistent with general preferences of citizens, and the means chosen are publicly defensible along lines of Rawls' Publicity Principle.

In this paper, we argue that this line of defence is seriously flawed. We argue that while it is true that our choices and, more generally, our behaviour are always influenced by the decision-making context, the intentional intervention aimed at affecting behaviour change ascribes certain responsibilities to the public policy-maker, which are not addressed satisfactorily by Thaler and Sunstein. Further, we argue that these responsibilities cannot be dismissed by pointing out that nudges are liberty preserving. While it is true in principle that citizens are free to choose otherwise, one can hardly appeal to this in a practical context because the nudge approach to behavioural change is applied exactly in contexts where we tend to fall short of such principles.

This seems to leave us open to the critics' claim that "nudge" is a public policy approach based on the manipulation of citizens' choices. Against this we argue that the characterization of nudging as mere manipulation of choice is too simplistic. While both classical economic theory and behavioural economics describe behaviour as the result of choices, the psychological dual process theory that underpins behavioural economics, used by Thaler and Sunstein,[27] distinguishes between automatic processes/behav-

14 Luc Bovens, "The Ethics of Nudge", in Till Grüne-Yanoff and Sven O. Hansson (eds) *Preference Change: Approaches from Philosophy, Economics and Psychology* (Berlin and New York: Springer, Theory and Decision Library A, 2008).

15 Signild Vallgårda, "Nudge a new and better way to improve health?", 104(2) *Health Policy* (2012), pp. 200 *et sqq*.

16 Gregory Mitchell, "Libertarian Paternalism is an oxymoron", 99 (3) *Northwestern University Law Review* (2004).

17 Riccardo Rebonato, *Taking Liberties – A Critical Examination of Libertarian Paternalism* (New York: Palgrave Macmillan, 2012).

18 Adam Burgess, "'Nudging' Healthy Lifestyles: The UK Experiments with the Behavioural Alternative to Regulation and the Market", 3(1) *European Journal of Risk Regulation* (2012), pp. 3–16.

19 Bovens, *The Ethics of nudge*, *supra* note 14, at p. 4.

20 *Ibid*

21 Henry Farrell and Cosma Shalizi, "'Nudge' policies are another name for coercion", *New Scientist*, Issue 2837. (2011).

22 Rebonato, *Taking Liberties*, *supra* note 17, at p. 4.

23 Frank Furedi, "Defending moral autonomy against an army of nudgers", 2011, available on the Internet at: <http://www.spiked-online.com/index.php/site/article/10102/> (last accessed on 09 January 2013).

24 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 82 and p. 239.

25 *Ibid*, pp. 10–11.

26 *Ibid*, pp. 244–245.

27 *Ibid*, pp. 17–101

iours on the one hand, and deliberate choices on the other. However, nudging always influence the former, but only sometimes affects the latter. The conceptual implication of this is that nudging only sometimes targets choices. What remains then is the accusation that nudging works by "manipulation."

Turning to this, we argue that Thaler and Sunstein's appeal to Rawls' Publicity Principle is insufficient; as a safeguard against non-legitimate state manipulation of people's choices, it is severely lacking. Instead, we introduce an epistemic distinction between transparent and non-transparent nudges, which serves as a basis for distinguishing the manipulative use of nudges from other types of uses. The result is a conceptual framework for describing the character of four broad types of nudges. These may provide a central component for more nuanced ethical considerations and a basis for various policy recommendations. It is our hope that this framework may clear up some of the confusion that surrounds the ethical discussion of the nudge approach to behavioural change, and better inform its adoption in public policy-making.

II. The Nudge approach in behavioural change policies

The basic premise of Thaler and Sunstein's *Nudge* is that human decision-making and behaviour – in contrast to the decision-making of perfectly rational agents inhabiting the models of standard economics – is often influenced in systematic ways by subtle, seemingly insignificant changes in the decision-making context. That this is true is a well-established fact of the 'biases and heuristics program,' pioneered by the late Amos Tversky and Nobel Laureate Daniel Kahnemann, as well as of social and cognitive psychology.[28,29] In fact, it was this program, which gave rise to the discipline of behavioural economics in the first place, of which Thaler is often considered to be one of the founders.

The contribution of Thaler and Sunstein's *Nudge*, however, is not that of conveying novel scientific insights or results about previously unknown biases and heuristics (something that Thaler has championed in his academic publications.[30,31] Instead, it is the notion of "nudge" itself, and the suggestion of this as a viable approach in public policy-making to influence citizens' behaviour while avoiding the problems and pitfalls of traditional regulatory approaches.

In their book, "traditional approaches" comprise standard regulatory approaches, such as the provision of information (e.g. by general information campaigns), direct regulation (bans and injunctions), and indirect regulation, understood as the design and manipulation of economic incentives.[32] According to Thaler and Sunstein, such approaches may not only have their own drawbacks. It is also increasingly evident that these traditional approaches often fail to elicit effective behaviour change. This latter feature is particularly problematic today where we face some of the greatest social and global challenges of history: population wide obesity epidemics, global climate-change, and the costs of aging populations.[33]

According to Thaler and Sunstein, part of the explanation for this problem is that the traditional public policy approaches to behaviour change share a theoretical basis in the intellectually and politically celebrated idea that humans are generally capable of acting rationally. That is, the idea that we generally can act optimally, according to our reflected preferences, as long as we are given true information, the right incentives, and reasonable rules to guide us.[34] This means that in public policy-making, humans are traditionally conceived of as strikingly similar to the perfectly rational Econs inhabiting the universe of standard economics.[35] While this is both a great ideal to aspire to, and conducive to our self-image, Thaler and Sunstein assert that using this ideal as a basis for real world public policy-making often results in failure. Our ideals bear little resemblance to what behavioural economics have revealed about our actual everyday decision-making and behaviour: viz., that all too often we choose and behave in ways that are bad for us, our loved ones, and society at large even when the pre-conditions for rational decision-making are present. At fault are subtle changes and

28 Daniel Kahneman and Amos Tversky, "The framing of decisions and the psychology of choice", 211 *Science* (1981) pp. 453–458.

29 Keith Stanovich, *Rationality and the Reflective Mind (Oxford:* Oxford University Press, 2010).

30 Richard H. Thaler, "Mental accounting and consumer choice", 4 *Marketing Science (*1985), pp. 199–214.

31 Richard H. Thaler, "Mental accounting matters", 12(3) *Journal of Behavioural Decision Making* (1999), pp. 183–206.

32 Thaler and Sunstein, *Nudge, supra* note 1, at p. 13–14, p. 243.

33 Piniewski, Codagnone, and Osimo, "Nudging lifestyles for better healtlh outcomes", *supra* note 6, at p. 3.

34 Thaler and Sunstein, *Nudge, supra* note 1, at p. 17.

35 *Ibid*, pp. 6–17

elements in the behavioural and decision-making context, even when the pre-conditions for rational decision-making are present.

Now, Thaler and Sunstein suggest that if subtle changes in the behavioural and decision-making context lead us astray from our own best interest, the insights into how and why this happens may also be used to gently 'nudge' us in the direction of what is good for us – our health, wealth, and happiness.[36] In particular, since this may often be done without further costs, recourse to traditional policy-measures, or provoking conflict with existing political ideologies, such an approach seems highly attractive.

1. Nudge – by definition

To properly assess Thaler and Sunstein's suggestion, it is crucial to determine what exactly is meant by a 'nudge' as well as to ensure that the core notion is viable for the present discussion.

Thaler and Sunstein define a nudge as:

> "... any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives".[37]

Here, the notion of 'choice architecture' is the equivalent of what we have described as the behavioural and decision-making context. Thus, a nudge is any aspect of this context that leads behaviour astray from the predictions of standard economics.

Yet, as pointed out in a later paper by Hausman and Welch,[38] rational agents are not only responsive to economic incentives. For instance, the payoff function of a rational agent is determined by the prospect of pain as well as penalties. Thus, if taken at face value, the definition would render a 10.000 voltage electroshock to count as a 'nudge.'

Since this would be a rather uncharitable interpretation of Thaler and Sunstein, it seems most sensible to follow Hausman and Welch's suggestion to broaden the definition so that it encompass all other types of incentives as well. Thus, they define a nudge as follows:

> "Nudges are ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth".[39]

To this definition, Hausman and Welch add the qualification that:

> "They [nudges] are called for because of flaws in individual decision-making, and work by making use of those flaws".[40]

Finally, Thaler and Sunstein describe a "good nudge" as one in which intervention is carried out by one agent to influence the choice and behaviour of another, in accordance with the interests of the latter, as judged by this person. In fact, every time Thaler signs a copy of *Nudge*, he signs with "nudge for good." When applied in public policy-making, Thaler and Sunstein label "nudging for good" as Libertarian Paternalism.[41]

Since behavioural economics offer a wide range of insights and tools to influence behaviour and decision-making of people without limiting their existing choices, or controlling them by incentives, it seemingly provides public policy-makers with the ultimate tool: an ethical, politically noncontroversial approach to influencing the choices and behaviour of citizens in accordance with their own interests.

2. The responsibilities of choice architects

From the original definition it is obvious that the notion of "choice architecture" is a central one to the nudge approach as well as Libertarian Paternalism. According to Thaler and Sunstein, "A choice architect has the responsibility for organizing the context in which people make decisions".[42] Thaler and Sunstein's suggestion is that public policy-makers should perceive themselves as "choice architects."

Given this definition, many people are choice architects, whether realizing it or not.[43] Cafeteria managers arranging food, doctors presenting alternative treatments for their patients, and people who design election ballots, are just a few of the many choice architects mentioned by Thaler and Sunstein.

36 *Ibid*, at. p. 7.

37 *Ibid*, at p. 6.

38 Daniel Hausmann and BrynnWelch, "Debate: To Nudge or Not to Nudge", 18 *Journal of Political Philosophy* (2010), pp. 123–136.

39 *Ibid*, at p. 126.

40 *Ibid*, at p. 126.

41 Thaler and Sunstein, *Nudge, supra* note 1, at p. 2.

42 *Ibid*, at p. 3.

43 *Ibid*, at p. 3.

However, in their introduction of the notion of a "choice architect," Thaler and Sunstein only mention examples where choice architects design, construct, or organize context without changing the original choice sets or fiddling with incentives – examples of choice architects who work according to the nudge approach. Yet, as Thaler and Sunstein's definition of a nudge indicates, there seems to be no reason to reserve the concept to interventions limited in this respect (Thaler explicitly endorsed this view in a tweet made on June 1, 2012). Thus, the policy-maker who adjusts the range of choices available to a citizen, or makes adjustments to the incentive structure in a decision-making context, seems to be just as much a choice architect. Hence, it is important to emphasize that nudging is a particular approach in the design and re-design of choice architecture. This raises the question whether there are any special obligations associated with this approach to Behaviour change – that is, are there special obligations to consider for choice architects working with the nudge approach to behaviour change?

A point emphasized repeatedly by Thaler and Sunstein in this respect is that there "are many parallels between choice architecture and more traditional forms of architecture".[44] One such parallel is that just as a traditional architect must eventually build a particular building, a choice architect must likewise choose some way of organizing the context that she is responsible for, and in which people make decisions.[45] Another, and perhaps even more "crucial parallel is that there is no such thing as a 'neutral' design".[46] Thus, in all those situations where an agent must make some choice in how to organize a given decision-making or behavioural context, it becomes impossible to avoid influencing people's choices and behaviour.[47] Therefore, they argue, the idea that it is impossible for a choice architect not to influence people's choices in one way or another is a misconception,[48] and thus an "anti-nudge position" is a "literal non-starter".[49] Hence, a choice architect, whether she likes it or not, simply can't avoid influencing the decisions and behaviour in the context she's responsible for organizing. The only responsible thing to do, it seems, is to recognize this and actively incorporate such knowledge when designing the choice architecture she's responsible for.[50] Thaler and Sunstein's suggestion in how to go about this is by using the nudge approach in the service of Libertarian Paternalism.

Granted that the "anti-nudge position" is a literal non-starter, the corollary of Thaler and Sunstein's argument seems to be that the only viable option is to embrace a "pro-nudge position". But if one can only be pro-nudge, it may also seem that one can have no serious complaint if nudged for one's own benefit; thus, it seems always permissible to nudge people's choices and behaviour, as long as one is certain that it will be to their benefit as judged by themselves. This seems to be the sole special responsibility of the choice architect working with the nudge-approach. After all, and as is repeatedly emphasized within the pro-nudge literature, one is neither forcing citizens nor limiting their freedom to choose otherwise.[51]

3. Criticism

Yet, as mentioned in the introduction, criticism of the nudge approach to behaviour change has been widespread in both the public and academic spheres. A persistent and central tenet in the political and ethical criticism has been the claim that nudging works by "manipulating people's choices".[52] Also, as we saw even Thaler and Sunstein seem to subscribe to this view.[53] In fact, this claim is a fundamental one from which several other criticisms of the nudge approach may be seen as fully or partially derived. For instance, claims such as: that Libertarian Paternalism is an oxymoron[54]; that the nudge doctrine is ultimately just paternalism in disguise;[55, 56, 57, 58] that

44 *Ibid*, at p. 3.

45 *Ibid*, at p. 4.

46 *Ibid*, at p. 3.

47 *Ibid*, at p. 10–11.

48 *Ibid*, at p. 10.

49 *Ibid*, at p. 11.

50 *Ibid*, at p. 10.

51 Dolan, Hallsworth,Halpern et al., "MINDSPACE", *supra* note 3, at p. 3.

52 Luc Bovens, "The Ethics of Nudge", *supra* note 14, at p. 4

53 Thaler and Sunstein, *Nudge, supra* note 1, at p. 2 *et sqq*., p. 82 and p. 239.

54 Mitchell, "Libertarian Paternalism is an oxymoron", *supra* note 16, at p. 4.

55 Vallgårda, "Nudge a new and better way to improve health?", *supra* note 15, at p. 4

56 Burgess, "'Nudging' Healthy Lifestyles", *supra* note 18

57 Furedi, "Defending moral autonomy against an army of nudgers", *supra* note 23, at p. 4

58 Rebonato, *Taking Liberties*, *supra* note 17, at p. 4.

nudging is a return to present day behaviourism;[59] that the techniques used in nudge "works best in the dark"[60,61] and that the effect of a nudge disappears if it is recognized by people;[62] that nudging encourages the abuse of power by technocrats;[63,64] and that nudging impairs our autonomy and our ability to make moral choices for ourselves.[65,66] In the end, many critics seem to have missed the purported fact that an anti-nudge position is a "literal non-starter."

As this paper aims to show, the characterization of nudging as the manipulation of choice, and the policy recommendations that result from this characterization, depend on the theory of agency that one subscribes to, and how this attaches to ethical considerations and normative responsibilities. In particular, it is a problem that much of the criticisms are rooted in theories of agency and ethical considerations quite different from the theory and concepts underpinning the nudge approach to behavioural change. Thus, to properly engage with the criticism based on the claim of nudging as the manipulation of choice, it is necessary to examine the scientific foundations of the nudge approach to see if this criticism actually sticks to it on its own premises. That is, whether nudging is best characterized as the manipulation of choice as viewed from the premises of behavioural economics and modern cognitive psychology. To this end, we intend to take a closer look at the foundations with the aim of setting up a viable framework for determining if and when nudging is about the manipulation of choice, and what follows from our conclusions.

Still, it only makes sense to embark on such a journey insofar that the anti-nudge position is a viable position in the first place, and not, as Thaler and Sunstein claim, "a literal non-starter."

59 Burgess, "'Nudging' Healthy Lifestyles", *supra* note 18.

60 Bovens, "The Ethics of Nudge", *supra* note 14, at p. 4.

61 Burgess, "'Nudging' Healthy Lifestyles", *supra* note 18.

62 Evan Selinger and Kyle P. Whyte, "Competence and trust in choice architecture", 23(3–4) *Knowledge, Technology & policy (2010),* pp. 461–482.

63 Farrell and Shalizi, "'Nudge policies' are another name for coercion", *supra* note 21, at p. 4.

64 Rebonato, *Taking Liberties*, *supra* note 17 at p. 4.

65 Furedi, "Defending moral autonomy against an army of nudgers", *supra* note 23, at p. 4.

66 Bovens, "The Ethics of Nudge", *supra* note 14, at p. 4.

III. The anti-nudge position, a literal non-starter?

Above we saw how Thaler and Sunstein define a nudge as "any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives." Given Hausman and Welch's definition in the previous section, it should be clear by now that these incentives need to be understood more broadly.

Returning to Thaler and Sunstein's definition relative to the viability of the anti-nudge position, this further indicates that choice architecture may exist without an architect ("any aspect of") and by implication that nudges may exist without a "nudger." It appears then that to the extent that the choice architecture of a decision-making context influences our behaviour beyond what the available options relative to incentives may account for, we are being nudged by the architecture, albeit not necessarily towards any particular ends.

In turn, this is what allows Thaler and Sunstein to argue by analogy to architecture that there is no such thing as a "neutral design" even when the architecture and its effects are accidental. They seem to suggest that though no one may have intended for the choice architecture in question to nudge us toward particular ends, we are nevertheless always being nudged toward some behaviour in a predictable and consequential way. This is true even if a choice architect has not interfered; but of course, in such cases, the effects will not be intended nor directed towards any well-defined or consistent end.

1. An attractive defence

The argument that neutral designs do not really exist provides an attractive line of defence for the nudge approach, especially relative to the claim (and others like it) that nudge theory works by manipulating choices. If this fundamental premise is accepted, it seemingly follows that (1) we are always being nudged, whether we like it or not and regardless of anyone intended it so. Nudges are an inescapable feature of any decision-making context. It seems unreasonable then to argue that we should take measures to avoid the unavoidable.

Of course, this prompts the question: How can someone legitimately influence the choices and be-

haviours of others? To this end, the line of defence emphasizes that (2) the nudge approach, when based on principles of libertarian paternalism – as suggested by Thaler and Sunstein – seeks to improve citizen behaviour and decision-making, as judged by the citizens. It seems reasonable to assume that citizens would prefer to be nudged according to these principles rather than the alternatives, in which our decision-making contexts are often determined by profit or pure chance. In fact, an ethical argument may be constructed that the recognition of her capacity to nudge people in this direction imposes a moral obligation on the choice architect to do so.

Finally, this line of defence is supported by the observation that the nudge approach to behavioural change is liberty preserving. It does not promote behaviour by regulating the existing freedom of choice or by manipulating incentives. In fact, (3) one can always reject the behaviour that a given nudge is devised to promote.[67] One can always choose to do otherwise. Therefore, it seems that one cannot reasonably be concerned about the nudge approach to behavioural change. That is, so long as you accept this three-lined defence as valid, which the next three subsections call into doubt.

2. Why we are not always being nudged

As we just saw, the first premise is a conditional one, indicating that:

(1) If there is no such thing as a neutral design, then we are always being nudged.

However, something important seems to be lost when accepting this premise at face value.

Specifically, given that the criticism of the nudge-approach as working by the manipulation of choice questions the normative justification for this approach to behavioural change, there seems to be a clear and important distinction to be made between a given context that *accidentally* influences behaviour in a predictable way, and someone – a choice architect – *intentionally* trying to alter behaviour by fiddling with such contexts.

In matters of justification one simply cannot dispense with the issue of intentionality, and by extension agency, in the way premise (1) does. Intentionality is a conceptual precondition of normative evaluation. Ignoring it would render the notion of responsibility superfluous. At its extreme, dispensing with the issue of intentionality as related to responsibility would permit one to make such arguments as, "Because everyone must eventually die someday, one is justified in taking another's life." Unfortunately, this is exactly what is indicated with the assumption that: if contexts always influence your behaviour in predictable ways, this implies that you are always being nudged. Said differently, such a conceptual move blurs a crucial distinction at the heart of normative justification as to the notion of responsibility.

This point is somewhat obscured even in the definition of a nudge, as formulated by Thaler and Sunstein. Remember, according to their definition, a nudge is "any aspect of the choice-architecture that alters people's behaviour" in a predictable way. Hausman and Welch's definition, on the other hand, takes intentionality into account by defining nudges as "ways of influencing choice." Whereas Thaler and Sunstein's definition characterizes a nudge as an objective relation between the settings of a given context and human behaviour as this unfolds, the latter definition characterizes a nudge as an intentional intervention enacted by one part to influence the choice of another part (where it is possible that these are the same, but the case that this is usually not so).

Obviously, this slight difference of words makes a crucial distinction because it introduces the normative dimension of responsibility for one's actions. Rather than discussing here which sense to reserve for the notion of nudge, we simply opt for Hausman and Welch's definition to ensure this important distinction is not lost. Thus, we suggest that a nudge henceforth is best understood as the intentional attempt at influencing choice, while it is accepted that the settings of any given decision-making context may accidentally influence choice and behaviour in predictable ways as well. This also implies that the conditional premise (1) may be evaluated as misleading, in the sense that it ignores a crucial distinction in the discussion to which it is applied. The notion of "nudge" then, should only apply when someone intentionally tries to influence our behaviour without the use of regulation or fiddling around with incentives.

67 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 2 *et sqq*., p. 8.

3. The ends and means of libertarian paternalism

The intentional character of nudging seems to be at the heart of concern for critics: if nudging is the result of deliberative attempts to influence on choices, it must be occurring to promote certain ends and values. But what guarantees that such ends and values are consistent with one's own?[68, 69, 70, 71]

If combined with the apparent acceptance, even by advocates of the nudge approach, that nudge works by manipulation, this leads to an even more serious concern: What distinguishes nudging from similar methods that influence citizens to act – if not against their will, then in absence of consent – as mere tools for those in power? It is this combination of concerns that leads blogs like Spiked to accuse the Behavioural Insights Team of propagating an "elitist politics of the brain",[72] and Rebonato as well as Farell and Zalizi to ask what keeps nudgers in check from their own biases, such as over-optimism and power-blindness.[73, 74]

However, before considering what these concerns amount to in their combination, it is wise to address the standard defence of nudging when faced with the isolated issue of the ends promoted. To this end, Thaler and Sunstein have repeatedly emphasized that policy-makers should apply nudging in the service of Libertarian Paternalism.[75] What this means is that policy-makers should nudge to promote ends that are in the interest of citizens, as judged by themselves.[76] Thus, Libertarian Paternalism presents a policy-paradigm falling within the category of "soft paternalism," as opposed to "hard paternalism".[77] Granted, for now, that nudging by definition allows citizens to choose as they please (see next section), this in turn means that the nudge approach to behavioural change and Libertarian Paternalism fits like hand in glove besides avoiding the above concerns.

In fact, as Thaler and Sunstein take pains to point out from the outset of *Nudge*, nudging guided by Libertarian Paternalism seems preferable to any existing alternative. It appears to be in the interest of citizens that relevant choice architecture, e.g. cafeterias, are designed to favour their reflected preferences, for instance, with regard to health and nutrition, rather than the interest of mere profit or at random. In fact, an argument could be made that given that choice architects know how decision-making contexts influence choice in subtle ways, they are obligated by the mandates of their work not to ignore such knowledge. Given only the standard toolbox of public policy, Libertarian Paternalists would usually want to limit their efforts to the provision of information. The addition to this toolbox of nudging and the insights from modern behavioural sciences that it builds upon provides them with the most non-intrusive means to behavioural change that simultaneously takes seriously that information provision is not always as effective as previously believed. Working within the nudge approach to behavioural change, public policy-makers and other choice architects may seek to alter behaviour in the interests of the citizens they serve, without introducing further regulation or preventing citizens from choosing as they like.

Of course, even if public policy-makers remain true to this ideal, problems remain. As noted by Rebonato[78] as well as Vallgårda[79] there may be severe obstacles to determining what people judge to be in their own interests, especially in cases where they have not expressed or thought much about their preferences. Even if one asks for their opinion, there may be problems caused by the 'plasticity' of preferences, which in itself results from how choices are presented.[80] For instance, when it comes to asking citizens about when and for what purposes they deem state intervention acceptable to influence their behaviour, answers may depend on the level of abstractness of questions to such an extent that even reflected answers may become inconsistent.[81] Yet it is essential

68 Mitchell, "Libertarian Paternalism is an oxymoron", *supra* note 16, at p. 4.

69 Farrell and Shalizi, "'Nudge' policies are another name for coercion", *supra* note 21, at p. 4.

70 Vallgårda, "Nudge a new and better way to improve health?", *supra* note 15, at p. 4,

71 Rebonato, *Taking Liberties*, *supra* note 17, at p. 4.

72 O´Neill, "A message to the illiberal Nudge Industry: Push off", *supra* note 13, at p. 3.

73 Rebonato, *Taking Liberties*, *supra* note 17, at p. 4.

74 Farrell and Shalizi, "'Nudge' policies are another name for coercion", *supra* note 21, at p. 4.

75 Thaler and Sunstein, *Nudge*, *supra* note 1, at pp. 4–5.

76 *Ibid*, at p. 5.

77 *Ibid*, at p. 5.

78 Rebonato, *Taking Liberties*, *supra* note 17, at p. 4,

79 Vallgårda, "Nudge a new and better way to improve health?", *supra* note 15, at p. 4.

80 Scott Plous, *The Psychology of Judgement and Decision Making (New York:* McGraw-Hill, 1993).

81 Chris Branson, Bobby Duffy, Chris Perry et. al., "Acceptable Behaviour: Public Opinion on Behaviour Change Policy", Ipsos MORI. 2012, available on the Internet at: <http://www.ipsos-mori.com/researchpublications/publications/1454/Acceptable-Behaviour.aspx> (last accessed on 09 January 2013).

to notice that these problems do not pertain to nudging, but to Libertarian Paternalism; just as important to notice, they are problems that any public policymaker faces in one form or another, whether a libertarian paternalist or not.

This leads us to believe that the justifications of the nudge approach, and the justifications of Libertarian Paternalism, are two distinct issues. One may apply nudging without being a Libertarian Paternalist, and one may be a Libertarian Paternalist without endorsing the nudge approach to behavioural change. While nudging is a means to promote behavioural change, Libertarian Paternalism is a guide, or a series of constraints on what ends may be promoted. If this were not the case, there would be no need for Thaler and Sunstein to work with a notion of "evil nudges." Nor would it be necessary for Thaler to sign copies of Nudge with the caveat, "nudge for good." Thus, the concern of citizens being nudged towards certain ends, which are not universally embraced, splits into two distinct problems. For the Libertarian Paternalist, this amounts to how she can know what ends citizens prefer as judged by themselves, and be motivated to respect these in public-policy making. For the nudge approach to behavioural change, the problem amounts to whether this approach, considered as a means, is compatible with democratic public policy-making, and in particular, with its cornerstone of democratic consent.

Still, the latter question may be considered to be the primary one. Whether nudging works by manipulation determines if there is a natural fit between the nudge approach and Libertarian Paternalism. After all, the notion of manipulation seems incompatible with the freedom of choice. The issue also raises the question of whether or not the nudge approach is compatible with public policy making within democracy in general. If nudging really works by manipulation, thereby undermining the freedom of choice, any errors or harmful intent by policy makers – libertarian paternalists or not – brings the approach of nudging dangerously close to resembling an "elitist politics of the brain."

4. The principled freedom to choose differently

So far we have seen that the *raison d'être* of the nudge approach is that of turning to our own advantage the heuristics and biases that otherwise often make humans fall short of acting according to their reflected preferences. In addition, we have also seen that nudging is intentional by definition, promoting certain ends. Together this makes for the intentional intervention by means to an end, implying that the anti-nudge position is not a literal non-starter, but imposes special obligations on choice architects. Yet, as the last section demonstrated, this does not mean that the justification of ends nudged toward rest on the nudge approach as such. In particular, the defence of the nudge approach to behavioural change adds that this is so because, by definition, the nudge approach considered as a means allows the possibility for citizens to choose differently if they wish.

However, there seems to be something inconsistent in this line of reasoning. Appealing to a principled freedom of choice sits uncomfortably with the insights underlying the nudge approach to behavioural change.[82]

On the one hand, these insights into our human fallibility are used to justify public intervention by the libertarian paternalist. The baseline is no longer that citizens are always capable of acting according to what is in their own interests as judged by themselves. Hence the relevant baseline in evaluating a policy measure is no longer an idealized world inhabited by perfectly rational citizens. Rather, a given measure should be evaluated relative to existing alternatives in the toolbox of public policymaking and the reflected preferences of the citizens that policymaking is devised to serve.

On the other hand, this commitment to a realistic theory of human agency implies that public policymakers can no longer appeal to the fictive capacity of humans that always act on their reflected preferences. This is especially true against the background of state intervention, when seeking to justify the means chosen. It seems inconsistent to claim that while nudging allows citizens, in principle, to choose differently, they will also be capable of doing so in practice. Insofar as nudging turns out to work by manipulating people's choices, it seems that citizens are not really free to choose differently, since behavioural change that comes about by nudging will occur, if not necessarily against the will of citizens, then at least without their active consent and knowledge.

82 This issue is also discussed by Riccardo Rebonato in *Taking Liberties*, supra note 17, at p. 4, under the heading of the 'reversability' of libertarian paternalism.

Ultimately, then, this is the problem we face. *Because we have just discarded the world where citizens act as hyper-rational beings as a relevant baseline in real world public policy-making, we can no longer appeal to what hyper-rational agents would be capable of (for instance, easily rejecting a given nudge) as part of a defence for the non-problematic character of the nudge-approach*. Thus, though nudging by definition does not promote ends by constraining the existing freedom of choice (regulation), or by controlling incentives (incentives control), the fact that one is always free as a matter of *principle* to refuse to conform with the behaviour or decision that the nudge is devised to promote does, contrary to (3), not make this true in *practice*. Hence, this abstract notion of freedom does not exempt public policy-makers from taking responsibility for the ends promoted, nor criticism of the means chosen, which in this case is the nudge-approach to behavioural change.

In conclusion, the anti-nudge position is not a literal non-starter. According to the *raison d'être* of the nudge approach, and to the extent that nudging works by manipulation, this does not necessarily leave citizens free to choose differently from the ends nudged towards. Rather, in such cases, if they exist, the nudge approach needs to be justified relative to the democratic processes from which the mandates of public policy-makers flow. Once again, this means that the question of whether nudging works by manipulating choice is crucial in determining the responsibilities incurred by public policy-makers when adopting the nudge approach to behavioural change.

IV. Does nudging "manipulate choice"?

To answer this question, we begin by focusing on the part of this claim that nudging works by manipulating "choice". As a starting point, we return to the difference in the available definitions. Remember, while Hausman and Welch define nudges as "ways of influencing choice," Thaler and Sunstein's original definition defines nudges as aspects of choice architecture altering "people's behaviour." However, since the notions of "choice" and "behaviour" are not interchangeable, this makes a further clarification of the definition, as well as the theories and insights behind it, necessary to determine whether nudging is a actually a means to manipulate "choice".

1. Dual Process Theory

One of the first things introduced by Thaler and Sunstein in Nudge is the dual process theory, which underpins much of modern psychology and neuroscience.[83] Recently, psychologist Daniel Kahneman – the Nobel laureate, good friend, and long-time collaborator with Thaler – made this theory a cornerstone of his celebrated book, *Thinking, Fast and Slow*.[84]

Dual process theory asserts that the human brain functions in ways that invites for a distinction between two kinds of thinking: one, which is intuitive and automatic, and another, which is reflective and rational. Kahneman dubs these ways of thinking *System 1* and *System 2*, respectively; we choose, however, to follow the lead of Thaler and Sunstein[85] when referring to these modes of thinking as automatic thinking and reflective thinking.

Automatic thinking is characterized by being fast, instinctive, and usually not associated with experiences that one would describe as thinking. Reflective thinking is associated with the deliberate and conscious processing of information. It is slow, effortful and needs concentration. It is associated with self-awareness, the experience of agency, autonomy, and volition.[86, 87, 88] The key features of each system are shown in Table 4.1.

Table 4.1 Two cognitive modes of thinking

Automatic thinking	Reflective thinking
Uncontrolled	Controlled
Effortless	Effortful
Associative	Deductive
Fast	Slow
Unconscious	Self-aware
Skilled	Rule following

The point of dual process theory is that a given behaviour can result from either mode of thinking. Breathing is a good example. On the one hand,

83 Thaler and Sunstein, *Nudge*, *supra* note 1, pp. 19–22.

84 Daniel Kahnemann, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

85 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 19.

86 *Ibid*

87 Stanovich, *Rationality and the Reflective Mind,* supra note 29, at p. 6,

88 Kahnemann, *Thinking Fast and Slow,* supra note 84, at p. 13.

breathing is usually maintained automatically, or by our 'automatic pilot', if you wish. On the other hand, reflective thinking may engage and control this bodily activity; a good example is the decision to hold your breath when passing a bad smell on the street.

However, it is crucial to notice that dual process theory does not imply that a given behaviour is maintained or results exclusively by one or the other mode of thinking. While automatic thinking operates on its own, reflective thinking operates on premises and in a context provided by automatic processes.[89] Somewhat simplified, it is automatic modes of thinking that prompts reflective thinking concerning the presence of a bad smell; it is also automatic thinking that informs its reflective counterpart that its decision to hold your breath is working out fine. It is also automatic thinking that tells the reflective one when it is safe to breathe again. In other words, automatic and reflective thinking may interact, and the latter always seems to depend in one way or another on the former, while the opposite is not true.

2. Actions and Causes

In order to solve the conceptual question caused by the difference between the two definitions of nudge mentioned above with regard to the use of the notions of "choice" and "behaviour," an additional, but parallel distinction needs to be observed. This is the fundamental distinction made in the theory of action between *actions* and *causes*.

In the theory of action, actions are often defined in terms of the "states-of-the-world" that an agent intentionally seeks to bring about; an intention resulting from the active deliberation about what is believed by the agent to be available courses of actions in the situation, and determined by the preferences over expected consequences associated with these. Given such a process, a resulting bodily movement, inference or judgment may be referred to as an action and described as a consequence of a process referred to as *choice*. A straightforward consequence to draw from this conceptual distinction seems to be that a choice may reasonably be interpreted as an end-result of the intervention of reflective thinking.

At the other side of this conceptual distinction we find "non-voluntary actions." These are in fact not to be considered as real actions, but only so in a derivative sense as being behaviour under the potential control of a given agent. A "non-voluntary action" is thus better referred to as an event that happens to you, but which could have been controlled. Examples are the blinking of your eyes when a ball is thrown at you (reflexes), covering your mouth when you cough (habit), or spotting just the right move in a chess game (expertise). Because such non-voluntary "actions" are unintentional and do not involve active deliberation, they are usually not conceived as resulting from choice as defined above. Instead they are said to happen and be "caused" by other events. As a consequence, the involved agent is usually not asked to assume responsibility for the "action" in question. Again, the straightforward conceptual consequence to draw from this, and complementary to the one above, is that, what is referred to in the theory of action as "non-voluntary actions" are strikingly similar to behaviours that do not result from deliberation. That is, they are strikingly similar to behaviours resulting solely from automatic thinking.

For this reason, we adopt, for the remainder of the paper, a definition of behaviour to encompass any bodily movements and cognitive processes, but reserve the concepts of 'choice' and 'action' to those movements or processes which results from reflective thinking.

3. Two types of nudges – only one aimed at choice

So is nudging a way of influencing "choice" as Hausman and Welch's definition states? Not always, we claim. The reason is that while nudging always affects automatic modes of thinking, it does not necessarily involve reflective thinking.

More specifically, we suggest that one may reasonably draw a distinction between two types of nudges: *type 1* nudges and *type 2* nudges. Both types of nudges aim at influencing automatic modes of thinking. But while type 2 nudges are aimed at influencing the attention and premises of – and hence the behaviour anchored in – reflective thinking (i.e. choices), via influencing the automatic system, type 1 nudges are aimed at influencing the behaviour maintained by automatic thinking, or consequences thereof without involving reflective thinking.

89 See e.g. Edward Cartwright, *Behavioural Economics*, (London: Routledge, 2011).

As an example of a type 2 nudge, one may think of the "fly-in-the-urinal" nudge featuring prominently as an example in *Nudge*.[90] This nudge aims at capturing the visual search processes continuously performed by automatic thinking. When this happens, the nudge works by attracting reflective attention. A direct consequence of reflective attention is that this either result in a decision to aim for the fly or not, but either way increases the likelihood of the agent to focus on the current act of urinating. Another prominent example of a type 2 nudge is framing, such as in the "Asian Disease problem" studied by Tversky and Kahneman.[91] This experiment shows how the frame in which a decision problem is formulated can affect our reflective choices due to the emotional response by automatic associations made in relation to this frame.[92]

Type 1 nudges are equally familiar. Think e.g. of Brian Wansink's manipulation of the default plate-size to influence the calorie intake of cafeteria guests.[93] In this case, decreasing the size of plates in a cafeteria in order to reduce costumers' calorie intake works without engaging with reflective thinking. Customers are nudged by "change of default" to put less food on their plates. As a consequence, they consume fewer calories. This happens, because they engage in the mindless eating habits of first filling their plate and then finishing it without thinking about it. There is usually no conscious decision or choice made in this sequence of behaviour in regard to how much to eat.

Another example of a type 1 nudge is narrowing the side-lines on a road in order to get drivers to slow down. In this case, speed is decreased as an automatic response to what amounts to a warning from automatic responses. The road is seemingly narrowing and as a result, an experienced driver immediately begins to break. Only later does reflective thought processes find time to notice the fact that this response was caused by a visual illusion (more examples of type 1 and type 2 nudges are given below).

Given the distinction between actions and causes, and the distinction between type 1 and type 2 nudges, it is now evident that type 1 nudges are those influencing behaviours that do not involve deliberation, judgment, and choice. Type 2 nudges, on the other hand, are those influencing behaviours best characterized as actions, the results of deliberation, judgment, and choice. Against this background, it becomes imprecise to stipulate, as done by Hausman and Welch, that "nudges are ways of influencing choice without limiting the choice set ..."

Of course, this conclusion is due to our adoption of a broader conception of "choice," than that usually found in microeconomics and seemingly adopted by Hausman and Welch. Within such a framework, any behaviour is described as the result of choice. But as it is beginning to appear, adding relevant complexity becomes important when addressing whether or not nudging is about manipulating choice in the framework of real world applications and everyday moral language.

This point may be incorporated into Hausman and Welch's definition by a minor revision:

> A nudge is any attempt at influencing **behaviour** in a predictable way without forbidding any previously available courses of actions or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth.

In other words, nudging is not necessarily about the manipulation of "choice."

4. Nudging as the "manipulation" of choice

So far we have argued that nudging is not necessarily about the manipulation of "choice". However, the transparency of policy measures is by itself an important issue at the heart of democratic policy-making, upon which concepts such as "accountability," "respect," "deliberation," "consent," and "acceptance" depend. Hence, it remains crucial to pin down to what extent nudging works by "manipulation," whether this amounts to the manipulation of choice or the manipulation of behaviour.

a. Thaler and Sunstein on transparency

The problem is addressed by Thaler and Sunstein themselves in their discussion about transparency.[94]

90 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 4.

91 Daniel Kahneman and Amos Tversky, "Choices, values, and frames", 39(4) *American Psychologist* (1984), pp. 341–350.

92 Kahnemann, *Thinking Fast and Slow,* supra note 84, at p. 13 *et sqq*., pp. 363–376.

93 Brian Wansink, "Environmental factors that increase the food intake and consumption volume of unknowing consumers", 24 *Annual Review of Nutrition* (2004), pp. 455–479.

94 Thaler and Sunstein, *Nudge*, *supra* note 1, pp. 239–244.

Indeed, transparency is ultimately established as a "guiding principle" for handling the various normative objections to the nudge approach. In particular, they derive their notion of transparency from what philosopher John Rawls refers to as *the publicity principle*.[95] In its simplest form, this principle bans government from selecting a policy that it would not be able or willing to defend publicly to its own citizens.[96]

Thaler and Sunstein cite two separate reasons why they endorse this principle. One is pragmatic:

> "If a government adopts a policy that it could not defend publicly, it stands to face considerable embarrassment, and perhaps much worse, if the policy and its grounds are disclosed."[97]

The other, normative reason, involves the idea of respect:

> "The government should respect the people whom it governs, and if it adopts policies that it could not defend in public, it fails to manifest that respect. Instead, it treats its citizens as tools for its own manipulation. In this sense, the publicity principle is connected with the prohibition on lying. Someone who lies treats people as means, not as ends."[98]

Ultimately this leads Thaler and Sunstein to conclude that the "publicity principle is a good guideline for constraining and implementing nudges, in both the public and private sectors".[99] However, being able and willing to publicly defend a given policy, if it should become necessary, does not only seem awfully compatible with straightforward paternalism. The notion of transparency based upon this principle also appears to be insufficient to guarantee acceptable public policy, even in the eyes of Thaler and Sunstein themselves. This is revealed when taking a closer look at Thaler and Sunstein's own discussion of transparency in practice.

Thus, for some cases, Thaler and Sunstein seem to suggest a more pro-active approach than required by the publicity principle. In relation to changing defaults for registering as an organ donor, Thaler and Sunstein state that the government "should not be secretive about what it is doing".[100] In regards to government use of framing, e.g. the use of cleverly worded signs and choice-descriptions, Thaler and Sunstein say, "they should be happy to reveal both their methods and their motives".[101]

However, even if adopting a more pro-active attitude than required by the publicity principle, Thaler and Sunstein admit, hard cases will still be imaginable, such as in the case of subliminal advertising.

> "In the abstract, subliminal advertising does seem to run afoul of the publicity principle. People are outraged by such advertising because they are being influenced without being informed of that fact. But what if the use of subliminal advertising were disclosed in advance?"[102]

At his point, Thaler and Sunstein admit that even disclosure in such cases is not sufficient to ensure ethical legitimacy.

> "We tend to think that it is not – that manipulation of this kind is objectionable precisely because it is invisible and thus impossible to monitor."[103]

b. Strong transparency is too restrictive and may lead to an ethical paradox

By their rejection of subliminal advertisement Thaler and Sunstein implicitly introduce an alternative and stronger principle of transparency in terms of 'visibility' and 'the possibility of monitoring' for the acceptability of using the nudge approach to behaviour change in public policy as well as in the private sector. While Thaler and Sunstein's ruling out of subliminal advertisement, even if disclosed, is in in accord with most people's moral intuitions, this stronger principle of transparency becomes too restrictive and may even lead to paradox if applied to the nudge approach in general.

Think, for instance, of Brian Wansink's experiments on varying the size of plates in cafeterias or putting lean, tall glasses on the table instead of small, wide glasses.[104] Both interventions qualify as nudg-

95 John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971).

96 *Ibid*, at p. 49.

97 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 245.

98 *Ibid*, at p. 245.

99 *Ibid*.

100 *Ibid*.

101 *Ibid*.

102 *Ibid*.

103 *Ibid*, at p. 246.

104 Koert van Ittersum and Brian Wansink, "Shape of Glass and Amount of Alcohol Poured: Comparative Study of Effect of Practice and Concentration", *British Medical Journal* (2005), at p. 331.

es and may be used to reduce calorie and alcohol intake. Yet, like subliminal advertising, neither of these nudges are visible or easy to monitor; as such, they should be unacceptable according to the strong principle of transparency. The same goes for Thaler and Sunstein's own example of using stripes at the beginning of a dangerous curve on the Lake Shore Drive in Chicago to induce "a sensation that driving speed is increasing";[105] or the clever framing of the risks associated with different choices, such as when doctors present medical treatments. In other words a public agency would not be recommended according to the strong principle of transparency to choose such road stripes or a frame encouraging a particular treatment, even if it believes that it accords best with the interests observed among citizens. Ultimately, the same point may apply to the change of defaults for organ donors. Only people working one way or the other with issues of organ donation have given much thought about the role of the default in registration rates for the organ donor registry. Thus, one problem of Thaler and Sunstein's notion of transparency as based on the principle of visibility, is that it seems to become more restrictive than intended by these authors.

Turning to inconsistency, Thaler and Sunstein's recommended guidelines based on transparency and visibility seem unrealistic to apply and live up to given what they themselves have said about choice architecture. Did Thaler and Sunstein not just argue that the anti-nudge position was a "literal non-starter"?

As they argued themselves: if a choice architect, like a traditional architect, must choose some way of organizing the context that she is responsible for and in which people make decisions[106] – and if "there is no such thing as a 'neutral' design"[107] – then it is impossible to avoid influencing people's choices and behaviour.[108] But if a choice architect knows that even subtle and invisible features of the context matters – features which are difficult if not impossible to monitor – then it seems that she must either refrain from doing anything about this or take these into account. However, if she chooses the former she neglects the health, wealth and happiness of those she influences, which seems unethical given her responsibility. If she chooses the latter, she breaks the normative constraints that Thaler and Sunstein suggest for the acceptable use of nudging. Hence, adopting the normative guidelines suggested by Thaler and Sunstein seems bound to end one up in an ethical paradox for some instances.

In conclusion, the notion of transparency, when based on the publicity principle as well as Thaler and Sunstein's stronger notion of transparency, seems insufficient as guidelines for the responsible and acceptable use of the nudge approach to behaviour change by public policy makers.

c. An epistemic dimension of transparency

Still, as we shall see, Thaler and Sunstein's stronger principle may be given an important role in the framework for evaluating the acceptability of particular nudge-interventions to be suggested here. The first step in this direction is made by the suggestion that, in addition to type 1 and type 2 nudges, an epistemic dimension for evaluating transparency based on the stronger principle of transparency is adopted. In turn, this distinction also reveals why nudging is not necessarily about "manipulating" choices and behaviours.

The distinction we offer is one based on Thaler and Sunstein's allusion to the idea that an attempt at influencing other people's behaviour, including choices, may be objectionable, "because it is invisible and thus impossible to monitor".[109] This creates a distinction between transparent and non-transparent nudges that is not based on anything like Rawls' publicity principle, but on epistemic grounds instead.

With this view, a *transparent nudge* is defined as a nudge provided in such a way that the intention behind it, as well as the means by which behavioural change is pursued, could reasonably be expected to be transparent to the agent being nudged as a result of the intervention. This notion of transparency is very close to what Bovens refers to as "token interference transparency",[110] although the present notion is more specific.

105 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 41–42.

106 *Ibid*, at p. 4.

107 *Ibid*, at p. 3.

108 *Ibid*, at p. 10–11.

109 *Ibid*, at p. 246.

110 Bovens, "The Ethics of nudge", *supra* note 14, at p. 4, p. 13.

figure 1

As examples of transparent nudges we offer the fly-in-the-urinal;[111] stickers such as that provided in Figure 1, when placed next to a light-switch; footprints painted on the floor or street, leading to the stairs or a garbage bin;[112] "look right" painted on the streets of London; the change of printer defaults;[113] and the use of visual illusions in traffic, but which contrary to the stripes on the Lake Shore Drive just mentioned, are sufficiently obvious to get noticed as intended to create a visual illusion after the effect has taken place.

In all of these cases, the citizen nudged can reasonably be expected to be able to easily reconstruct the intention behind the nudge, and the means by which behaviour change is pursued.

A *non-transparent nudge*, on the other hand, will be defined as a nudge working in a way that the citizen in the situation cannot reconstruct either the intention or the means by which behavioural change is pursued.

As examples of non-transparent nudges we offer the use of stripes at the Lake Shore Drive in Chicago;[114] the shrinking of plate sizes aimed e.g. at reducing calorie intake;[115] the removal of trays in cafeterias, aimed to reduce food waste;[116] the clever use of words to frame decision-making on which medical treatment may be chosen;[117] and the change of defaults from opt-in to opt-out, for registering for organ donation.[118]

d. Transparency and manipulation

Our claim is that the notion of epistemic transparency may be used as a criterion for evaluating whether a nudge is a case of manipulation, especially in the sense relevant to critics.

The sense of "manipulation" that critics have been concerned with clearly seems to be a *psychological sense of manipulation*. That is, manipulation in the sense of intending to change the perception, choices or behaviour of others through underhanded deceptive, or even abusive tactics.[119] Thus, for instance, Bovens pitches nudging as working by the manipulation of choice.[120]

> "What these examples [of save more tomorrow and cafeteria re-arrangement] have in common is a manipulation of people's choices via the choice architecture, i.e. the way in which the choices are presented to them ... In all cases of Nudge, if the choice situation had not been so structured, then people would be less prone to make the choice that is either in their own or in society's interest.[121]

He then later explicitly asserts nudging as being in a latent conflict with epistemic transparency referred to by him as "token interference transparency":

> "The problem is that these techniques [nudges] do work best in the dark. So the more actual token interference transparency we demand, the less effective these techniques are."[122]

111 Thaler and Sunstein, *Nudge*, *supra* note 1, at p 2 *et sqq*.

112 The Economist, "Nudge nudge, think think", (2012), available on the Internet at: http://www.economist.com/node/21551032 (last accessed on 09 January 2013).

113 Rutgers, "the Print Green Program", available on the Internet at <http://www.nbcs.rutgers.edu/ccf/main/print/> (last accessed on 09 January 2013).

114 Thaler and Sunstein, *Nudge*, *supra* note 1, at p. 37.

115 Wansink, "Environmental factors that increase the food intake and consumption volume of unknowing consumers", *supra* note 93, at p. 22

116 Lisa W. Foderaro, Without Cafeteria Trays, Colleges Finds Savings, 2009, available on the Internet at <http://www.nytimes.com/2009/04/29/nyregion/29tray.html?_r=0> (last accessed on 09 January 2013).

117 Kahneman and Tversky, "Choices, values, and frames", *supra* note 91 at p. 22

118 Thaler and Sunstein, *Nudge*, *supra* note 1, pp. 175–182.

119 Harriet B. Braiker, *Who's Pulling Your Strings? How To Break The Cycle of Manipulation And Regain Control Of Your Life* (New York: McGraw-Hill, 2004).

120 Bovens, "The Ethics of nudge", *supra* note 14 at p. 4

121 *Ibid*, at p. 2

122 *Ibid*, at p. 13

It should be noticed right away that Bovens' claim that epistemic transparency is in conflict with the efficacy of nudging is not a direct claim as to nudging being psychological manipulation. Yet, it does lend credibility to this latter claim. If epistemic transparency undermines the approach, it must be because people would otherwise be acting against their will – they would have been "manipulated" in the sense that their perception, choices, or behaviour had been affected through underhanded deception, or abusive tactics.

However, if this is true it seems that a) using strong transparency, as a guideline would rule out all nudges, as well as b) render the possibility of any acceptable use of the nudge approach to behaviour change non-existent. Yet, in drawing the distinction between transparent and non-transparent nudges we just saw that there are several examples of transparent nudges that work undisturbed by such transparency. In fact, some of these – e.g. the-fly-in-the-urinal and stickers next to light switch – actually seem to work because of this. The save-more-tomorrow-program, prompted choice for organ donation, calorie-boards and energy bills allowing for social comparison of electricity consumption, may also be mentioned as counterexamples. What these interventions seem to share, rather than the "manipulation of choice" is a transparent reliance on consistency or 'ego', where an agent's broader or long term preferences – call them reflected preferences – are nudged into activity with reflective thinking and in turn behavioural change as a result.[123] Thus, Bovens clearly seems to overstate the case of a latent conflict between transparency and nudging.[124]

In conclusion, then, it seems that, contrary to a), there is not necessarily a latent conflict between nudging and transparency. Hence, a request on normative grounds for the epistemic transparency of nudges does not seem incompatible with the efficacy of nudging, nor, contrary to b) the possible acceptability of nudge-based policy-making. In fact, the exact opposite conclusion seems to hold, since it seems plausible that such transparency offers an immediate filter on behavioural changes not supported by the citizens nudged.

Further, if manipulation is taken, as the critics, in the psychological sense as that of intending to change the perception, choices or behaviour of others through underhanded deceptive, or even abusive tactics, it is clear from the above examples that nudging does not necessarily work by "manipulation", whether of choice or behaviour. In fact, as we shall see the distinction between epistemic transparent and non-transparent nudges may serve as a basis for evaluating nudges as working by manipulation.

Of course this prompts the question of how high-skilled critics have been able of not noticing this distinction. The answer might be found in the overt triviality of transparent nudges that are readily accessible to intuition and in accordance with preferences. For instance, for transparent nudges such as the fly-in-the-urinal, the save-more-tomorrow program of Thaler and Bernatzi and discussed by Bovens as well as prompted choice for registering as an organ donor it is not so much the outcome of decision-making, which is nudged as the event of decision-making in consistency with people's self-images. That is, these nudges consists of a decision being prompted, which leaves all the original opportunities open as well as the original overall structure of incentives intact, but which nudge the agent for a decision in consistency with her reflected (long-term or broad-perspective) preferences.

123 Bovens does in fact briefly comment on this type of nudges, where the preference for consistency between actions and reflected preferences lead to behavioural change. However, since Bovens overemphasizes the cases where one is nudged toward some end that one does not agree to, his point becomes that when behavioural change occur in these instances due to consistency, this may lead to a fragmented self.

124 The cause of his mistake seems partially to be found in a conflation between the psychological sense of manipulation and the more comprehensive, neutral and *technical* sense, i.e. the intentional manipulation of a straightforward cause-and-effect relationship. It should be noted that nudging usually only changes frequencies and thus the effect is probabilistic rather than deterministic. In regard to manipulation, this is both good and bad news. The good news is that a deterministic change would render nudging more intrusive/manipulative, since it would indicate that we have no way to avoid its influence. However, this is not the case. Looking at the above typology, the closest one comes to such a deterministic relationship seems to be type 1 nudges, where the cause-and-effect relationship may be conjectured to be more deterministic than for type 2 nudges, since there is no active decision-making that could interfere with the Behaviour change pursued. Especially, when a type 2 nudge is epistemic transparent does this possibility seem to arise. Hence, in this "technical" sense of manipulation, type 1 nudges in general seem more robust and thus manipulative than type 2 nudges in general. The case for nudging as manipulation in the "technical" sense seems more probable when applied to automatic behaviour than to choice.
Yet, the reason why transparency may undermine the efficacy or robustness of a nudge does not seem to hang solely on the distinction between type 1 and type 2 nudges. In the cases for which Bovens claims that transparency undermines effect, it rather seems to be the combination of the transparency of a type 2 nudge with the fact that the aim nudged towards do not square with the reflected preferences of the citizen, that is at fault. For instance, a reader of Spiked may recognize the fly-in-the-urinal and decide to pee on the wall as a response to the intervention. Bovens claim thus seems to result from an over-emphasis on transparent type 2 nudges that seek to promote behavioural changes the end or means of which citizens do not agree with, rather than from nudging as such.

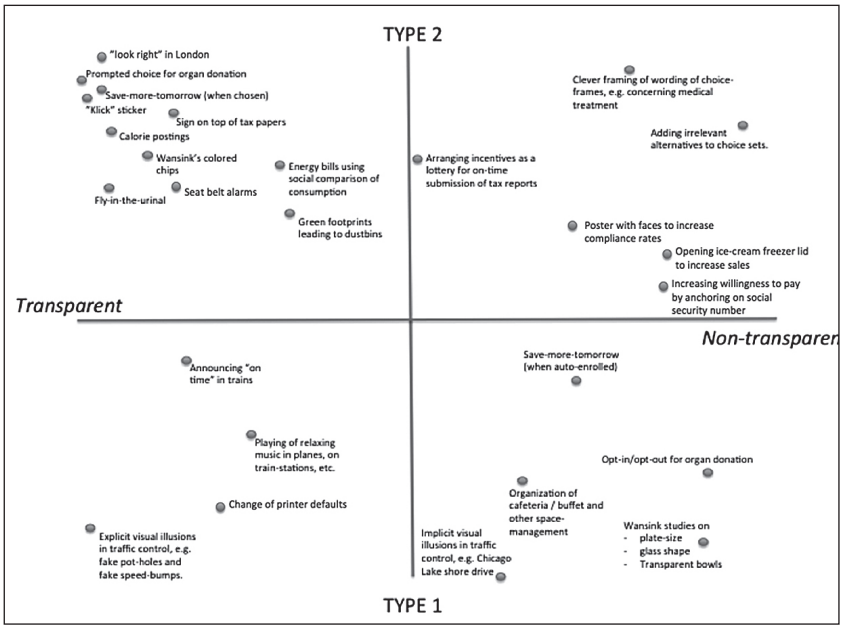
This use of consistency to explicitly nudge decision-making sometimes confuses scholars to the point where they fail to see the nudge in the intervention. However, this is not due to its subtlety, but rather to its overt triviality. It is ironic that prompted choice for organ donation and the Save More Tomorrow program are so epistemic transparent that high-skilled critics, who expect nudges to "work in the dark," fail to see nudges as nudging. Due to the transparent nature of these nudges, it seems difficult to maintain claims such as that alluded to by Bovens, that what nudges have in common is "manipulation" in the psychological sense of influencing the perception, choices, or behaviour of others through underhanded deception, or even abusive tactics.

Thus in so far as we take the relevant notion of manipulation referred to by critics of the nudge approach to behaviour change to be the psychological one of intending to change the perception, choices or behaviour of others through underhanded deceptive, or even abusive tactics, it seems obvious that the epistemic notion of transparency suggested here has an important role to play. In particular, if a nudge may be categorized, either pre- or post-intervention, as transparent, it seems that the notion of manipulation does not apply in this sense. It may thus be concluded that it would be misleading to describe the nudge approach to behaviour change as necessarily working through "manipulation", whether of choice or behaviour. However, to see the full implications of the notion of epistemic transparency on the issue of responsible and acceptable use of the nudge approach, we need to combine this with the distinction between type 1 and type 2 nudges.

V. Nudging and the manipulation of choice

Given the two distinctions developed (type 1 and type 2 nudges) and epistemic transparency and non-transparency, respectively, it is now possible to produce a matrix delineating four different types of nudges.

We recognize that there will be some nudge interventions that will be difficult to place. Some may fall into grey zones or seem to qualify for being several types due to their multi-layered structure of mechanisms and long-term dynamics that allow them to wander between categories. Amongst other things, this means that the issue of "Fuzzy Nudges" raised by Selinger and Whyte is not dealt with by the typology.[125] Yet the matrix form is valuable in guiding a responsible use of the nudge approach to behavioural change. It provides a basis for a typology of four types of nudges, each with their own characterization, evaluation, and policy recommendation.



1. Transparent type 2 nudges

In the top-left corner of the matrix we have epistemic transparent type 2 nudges. This type of nudge intervention engages the reflective system in a way that makes it easy for the citizen to reconstruct the intentions and means by which behaviour change is pursued.

125 Evan Selinger and Kyle Whyte, "Is there a right way to nudge? The Practice and Ethics of Choice Architecture", 5(10) *Sociology Compass* (2011), pp. 923–935.

A prominent example is the fly-in-the-urinal intervention. Most likely this influences behaviour by first engaging the citizen's automatic perceptual search processes; attention is drawn to the fly by the means of a contrast shadow resembling that of a living insect. The citizen's cognitive capacities for interpretation, which – when it's first there – then focuses on the current action (urination) in a way that prompts the conscious, albeit low-level, decision to make "aiming" adjustments. (Of course, we have yet to see someone actually investigating this in a brain scanner).

In the context and nature of the intervention – when not misinterpreting the fly as a poor company brand – the citizen recognizes the-fly-in-the-urinal as a deliberate attempt to influence decision-making. The ends and means are transparently epistemic prior to the decision. This also means that if the citizen does not agree with the ends or means, he may actively resist behaviour change.

Other examples of type 2 transparent nudges are interventions that like the-fly-in-the-urinal prompt decision-making by making aspects patently clear (seat belt alarms); making particular actions salient ("look right" painted on the streets of London, or the provision of nutritional advise by showing how to combine food on a plate, as done by ChoseMyPlate.gov, in lieu of the traditional food-pyramid); making preferences salient (e.g. by the use of green arrows or footprints to nudge people to take the stairs or throw litter in dustbins); making consequences salient (e.g., displaying disturbing pictures on cigarette boxes, putting calorie postings on menus, or providing real-time feedback on energy-use); or using social salience (e.g., electronic boards that depict one's real-time speed in a way that makes this speed public knowledge). Prompted choice for organ-donation and the Save More Tomorrow program also qualifies as transparent type 2 nudges. These two examples both work by prompting decisions in consistency with self-image, ego, or reflected long-term preferences. Finally, there are also transparent nudges that work by commitment (e.g., getting people to verbally repeat a scheduled appointment with their doctor, having gardeners hand out garbage bags to park visitors while having them verbally commit to leaving no trash behind, or putting signatures up front to increase honesty in tax reporting), as well as elicitation of descriptive norms with a clear messenger (e.g. the explicit private provision by a public agency of information about how your energy consumption compares with other people's consumption) counts as type 2 transparent nudges.

2. Transparent type 1 nudges

In the bottom-left corner we have epistemic transparent type 1 nudges. For this type of nudges reflective thinking is not engaged in what causes the behaviour change in question. Rather, reflective thinking occurs as a by-product, but in a way that easily allows for the reconstruction of ends and means.

A paradigm case of this type of nudges is the playing of relaxing music while passengers board a plane in order to calm them. By playing such music most passengers automatically begin to relax without thinking about it. However, the intention behind and use of playing relaxing music when boarding a plane, is easily recognized by passengers, but without this is necessary or prevents the behaviour change.

Another interesting example of a transparent type 1 nudge is one used by the Danish National Railway agency. Speakers in city trains are used to announce "on time" when trains arrive on time. This nudge has been devised in order to get people to easily remember not just the negative, for example, when a train is delayed, but also the positive, when trains are on time. Again, passengers easily recognize the intention behind and means by which such behaviour change – more precisely, attitude change – is pursued.

Other examples of transparent type 1 nudges are nudges that work by activating instinctive automatic responses (e.g., the use of the colour red, or flashing lights to draw attention to a sign, and the use of a car horn); nudges that work by activating learned responses (e.g., the fictive and somewhat dangerous use of fake potholes painted on the road to slow driver speed); nudges that work by changing the consequences of defaults in ways you are bound to notice (changing printer defaults from one-side to double-sided printing); or the more curious nudges, such as writing "you are now breathing manually" in a text, like this, which automatically causes you to breathe manually. For all of these nudges the behaviour change is more or less unavoidable to begin with, but transparent in a way that allows the influenced person to recognize the intention and means by which this is achieved as a direct consequence of the intervention.

3. Non-transparent type 2 nudges

In the top-right corner we have the non-transparent type 2 nudges. For this type of nudges to be suc-

cessful, the reflective system has to be engaged, but it doesn't happen in a way that by itself gives people epistemic access to the intentions and means by which influence is pursued.

A paradigm example is the clever framing of risks aimed to influence one's decision-making, e.g. when choosing between medical treatments. Most likely, such framing works by providing automatic processes with emotional associations, categories, and relations that in turn are handed over as the relevant premises for reflective decision-making, which is then called upon to make a decision on the basis of these premises. It is obvious that only very cautious or suspicious people will ever perceive such influences, if pursued in a subtle way. Different from transparent type 2 nudges, the recognition of such interventions is thus not a necessary (nor necessarily a detrimental) condition for this kind of nudge to succeed. Hence, the nudge works without epistemic transparency, while still engaging the reflective system.

Other examples of non-transparent type 2 nudges are nudges in general aimed at affecting decision-making by the clever framing of risks (e.g., when choosing between two medical treatments)[126]; nudges aimed at improving compliance rates in subtle ways (e.g., by posting posters with human faces to increase compliance rates with norms, such as cleaning up after oneself or paying for coffee);[127] using subtle cues to activate preferences for making particular choices (e.g., taking the lid off the ice-cream freezer, leading more costumers to crave and ultimately buy ice-cream);[128] using lotteries to get people to overestimate the chance of obtaining a rare effect (e.g., lotteries to encourage tax reporting); anchoring people's willingness for what price to pay for chocolate on their social security number;[129] and the subtle hint of scarcity or behavioural norms (e.g., having people queuing in front of a shop to lead others to believe that whatever is being sold must be good).[130,131]

4. Non-transparent type 1 nudges

Finally, in the bottom-right corner of the matrix we have non-transparent type 1 nudges. This type of nudges cause behaviour change without engaging the reflective system and in a way that does not make it likely to be recognized and transparent.

A long series of examples of non-transparent type 1 nudges is found in Brian Wansink's work, much of which is summarized in the bestseller *Mindless Eating*.[132] For instance, Wansink has found that by reducing the size of plates in a cafeteria from a 12-inch dinner plate to a 10-inch dinner plate leads people to serve and eat 22 % less calories.[133] The mechanism behind this change is the automated habit of first filling a plate and then finishing it. The habit of eating up was shown by Wansink in another study, where soup bowls would fill up without the subjects noticing it, ultimately leading subjects to eat 73 % more than subjects eating from normal soup bowls.[134] However, nudging people to lower (or increase) calorie intake by such measures work without engaging reflective thinking. It seems unreasonable to say that we have made a conscious decision to fill up the plate in the first place, as well as to say that we have decided to finish up. Further, a point repeatedly emphasized by Wansink is that we usually never notice influences like these and often find them unlikely when told about them. Even for cases where such nudges are explicitly pointed out to us, we have a hard time finding the resulting effects credible. It is safe to say, then, that nudges like these influence behaviour in a non-transparent way.

Other examples of non-transparent type 1 nudges are: changing of background defaults (e.g., changing from an opt-in to an opt-out procedure for registering as an organ donor); subtle and seemingly irrelevant changes to objects or arrangements in the behavioural context (e.g., changing the shape of glasses to reduce calorie intake, the removal of trays in cafeterias to reduce food waste, and rearranging cafeterias to get people to head for the salad buffet first rather

126 Kahneman and Tversky, "Choices, values, and frames", *supra* note 91, at p. 22.

127 Chris Branson; Bobby Duffy; Chris Perry et al., "Acceptable Behaviour: Public Opinion on Behaviour Change Policy", *supra* note 81, at p. 15.

128 A.W Meyers, A.J Stunard, M. Coll, "Food accessibility and food choice. A test of Schachter´s externality hypothesis", 37(10) *Archives of General Psychology* (1980), pp. 1133–1135.

129 Dan Ariely, George Loewenstein and Drazen Prelec, "'Coherent Arbitrariness': Stable demand curves without stable preferences", 118(1) *Quarterly Journal of Economics* (2003), pp. 73–105.

130 Cass Sunstein, *Infotopia: How Many Minds Produce Knowledge*, (Oxford: Oxford University Press, 2006).

131 Pelle Guldborg Hansen and Vincent Fella Hendricks, *Oplysningens Blinde Vinkler*, (Frederiksberg: Samfundslitteratur, 2011).

132 Brian Wansink, *Mindless Eating. Why we eat more than we think* (New York: Bantam, 2010).

133 Brian Wansink, "Environmental factors that increase the food intake and consumption volume of unknowing consumers", *supra* note 93, at p. 22.

134 *Ibid.*

than the meat); and the use of anchoring expectations (e.g., announcing a longer waiting time than actually expected, so people become pleasantly surprised).

Thus, this categorization lends itself to summing up the four different types of nudges relating to how they affect us and how acceptable their use is:

	Transparent	*Non-transparent*
System 2 thinking	Transparent facilitation of consistent choice	Manipulation of choice
System 1 thinking	Transparent influence (technical manipulation) of behavior	Non-transparent manipulation of behavior

Table 1: Suitable labels of intervention types

VI. A framework for the responsible use of the nudge approach to behaviour change

The paper began with Thaler and Sunstein's characterization of the anti-nudge position as a literal non-starter. However, it was argued that the intentional intervention by policy-makers and other choice architects in the lives of citizens confers special responsibilities; responsibilities that cannot be ducked simply because options and incentives in principle are left untouched by this approach. In addition, because nudging has been widely described as the "manipulation of choice," this seems to make the approach particularly difficult to reconcile with fundamental democratic values such as the respect for citizens, as well as their active participation and consent. Yet we have also argued that nudging is much more nuanced than this characterization. This seems to call for more nuanced considerations for the responsible use of the nudge approach in public policy-making.

Section 4.4 showed that Thaler and Sunstein seem to concede this. Thus, as we have seen in their discussion on the ethics of nudge,[135] they argue for the adoption of Rawls' publicity principle as a guideline for the responsible use of the nudge approach to behaviour change.[136] Their reason is partly pragmatic and partly normative. Specifically, they claim that public policy-makers should not be allowed to use citizens as mere tools.[137] However, if nudging is a question of manipulating citizens' choices, even the kind of passive disclosure implied by Rawls' publicity principle seems to be insufficient for certain cases. Seemingly recognizing this, Thaler and Sunstein then considered the adoption of more active forms of disclosure as a possible guideline.[138] But they also recognize that even active disclosure seems insufficient for responsible policymaking given that some cases of nudging, such as subliminal advertisement, are invisible to citizens and difficult to monitor.[139] In the end, public policy-makers recognizing themselves as choice architects could easily end in an ethical paradox: responsible for certain decision-making contexts that they know will influence citizens' choices and behaviours, but uncertain about the acceptability of applying the nudge approach for behavioural change.

However, given the typology developed in the previous section, we believe that important policy implications for the responsible use of nudging emerge. The primary one is that the typology provides a framework for evaluating whether a nudge is manipulative (non-transparent) or not, as well as whether manipulation pertains to choice (reflective thinking) or not. This makes the framework an important one for generating guidelines for the responsible use of the nudge approach by providing a typology of four types of nudges, each with its own characterization and policy recommendations. Of course, the credibility of these recommendations depend on how they square with robust normative intuitions, as well as whether they succeed in clarifying the considerations put forward by Thaler and Sunstein. We now turn to this issue.

1. Transparent Type 2 nudges: Prompting of reflected choice

Looking at transparent type 2 nudges, it is obvious that although this type of nudges tries to influence behaviour anchored in reflective thinking, such nudges are not aimed at doing so by means of psychological manipulation. Rather, nudges of this type aim at promoting decision-making in ways that are

135 Thaler and Sunstein, *Nudge*, *supra* note 1, pp. 236–251.

136 *Ibid*, at p. 244.

137 *Ibid*, at p. 245.

138 *Ibid*, at p. 246.

139 *Ibid*, at p. 246.

transparent to the agents influenced. These nudges work by prompting choices consistent with the reflected preferences of citizens by making features, actions, preferences, and/or consequences salient, or by providing feedback and decision- or commitment-mechanisms. Thus, to the extent that we take such preferences to constitute the core of autonomous decision-making, transparent type 2 nudges may be said to facilitate the "freedom of choice" and the empowerment of citizens in complex environments.

Viewed this way, transparent type 2 nudges may ultimately be characterized as a certain type of 'libertarian' nudge. They actually allow citizens to be nudged, to change their actions and behaviour in a predictable way, *while simultaneously leaving them free to choose otherwise – not just as a matter of principle, but also in practice*. Hence, nudges of this type may be characterized as 'empowerment' nudges, which promote decision-making in the interests of citizens, as judged by themselves, without introducing further regulation or incentives.

Due to these features we believe that when it comes to transparent type 2 nudges the anti-nudge position is truly, as originally claimed by Thaler and Sunstein, a "literal non-starter." While it is true that policy-makers and other choice architects who use this type of nudges intentionally intervene in the lives of citizens, nudges of this type are the least invasive. People are nudged towards reflective decision-making in a certain context, but in ways that allow for the full freedom of choice consistent with their reflected preferences. Also, since nudges of this type are transparent, the personal responsibility, which usually accrues to citizens when their behaviour is rooted in their reflective decision-making, is not misplaced. Citizens are nudged to consider choices without the use of manipulative measures usually because decision-making in itself is regarded as valuable. Of course, policy-makers and choice architects should carry the responsibility for the mild disturbances that this type of interventions cause in the lives of citizens, but this responsibility will be clear due to the transparency of the nudge, and usually acceptable due to the non-intrusiveness compared to the traditional policy measures.[140]

A real world exemplification comes by means of the suggestion of prompted choice for registering as an organ donor. Different from the traditional procedures of opt-in and opt-out, prompted choice for registering as an organ donor – for example, when requiring citizens to answer either "yes", "no" or "undecided" in order to obtain a driver's license – calls for active decision-making, aimed only at reducing the gap between reported attitudes and actual participation in the register. Of course, this requires that policy-makers take responsibility for the mild disturbance that this prompt causes in the lives of citizens, with the argument that such decision-making is regarded as sufficiently important to justify it. Given the transparency of the intention behind, and means by which this disturbance is performed, citizens are thus free to dispute the decision of policy-makers as well as exercise choice. As a matter of fact, the public discussions about prompted choice for registering as an organ donor in the US, UK and Denmark seem to reflect that these issues are actually those that arise.

2. Transparent type 1 nudges: Influencing behaviour

Turning to transparent type 1 nudges, the framework allows us to conclude that nudges of this type do not try to influence citizens' "choices". Rather, they are about influencing automatic behaviours and the consequences thereof in a transparent way. This type of influence is difficult, if not impossible to avoid, because it activates instinctive or learned responses. Nevertheless, we choose to characterize such interventions as the "influencing", rather than the "manipulation" of behaviour. Admittedly, these nudges do in a sense work by manipulation. But it is important to notice that this is in the sense of "technical" manipulation, not "psychological" manipulation, cf. (note 124).

Still, it should be emphasized that citizens are not trivially free to ignore the ends nudged toward, and choose otherwise if they prefer to do so. While this freedom remains in principle, the effect is usually unavoidable in practice – at least to begin with. Being exposed to a transparent type 1 nudge does not allow the citizen to avoid the effect because it works through automatic behaviour. As a consequence, nudges of this type are not truly libertarian.

140 In fact this road to behavioural change may be evaluated as even less invasive and less manipulative than the provision of information. Information is hard to provide in an objective way, and for instance in the case of prompted choice for organ donation, it is only the act of taking a stand on the issue which is highlighted as important, rather than what public policy-makers deem as the right information about this.

This characterization of transparent type 1 nudges points to a distinct set of responsibilities for the policy-maker. Since the influence pertains to automatic behaviour and not reflected choice, the policy-maker acquires full responsibility for the effect of the nudge. Of course, citizens may over time learn to recognize such interventions, allowing citizens to avoid them. However, we believe that this does not exempt policy-makers from their responsibility for the nudges' effects, especially because citizens are not usually required to learn avoidance techniques against nudges. Also, it is important to emphasize that learned avoidance of nudges might have the unintended effect of undermining otherwise appropriate behavioural responses. Finally, it should be noted that such considerations, plus transparency by itself, are not sufficient to ensure the responsible use of transparent type 1 nudges. While transparency makes it possible for citizens to recognize the intention behind and means by which their behaviour is influenced, it does not easily allow them to avoid this. Thus, in order to fulfil her responsibilities, the policy maker should be required to provide passive disclosure and transparent paths to filing grievances. In other words, it is to this type of nudges that Thaler and Sunstein's suggestion of Rawls' Public Publicity Principle as a guideline is appropriate. So long as these recommendations are adhered to, the use of transparent type 1 nudges should be regarded as generally acceptable; they allow citizens to easily dispute their influence within the democratic process, and assign proper responsibilities to public policy-makers.

A real world exemplification comes by means of a tentative attempt by road planners in Philadelphia to encourage careful driving by painting illusions of speed bumps on streets. In an experiment, 10 sets of illusions were burned into a half-mile stretch of road to test their effects on drivers. Before the intervention drivers were clocked averaging 38 miles per hour, 13 mph above the posted speed limit. A month later, that figure had dropped to 23 mph.[141] Obviously, the use of such illusions becomes transparent to drivers as soon as they pass the illusion. Yet, according to our view, the road planners will have to take responsibility, not just for the immediate effects, but also for possible side effects. As noted by Tom Vanderbilt, author of *Traffic* "One of the main drawbacks [of this experiment] is that people who live in the neighbourhood or use the road regularly ... will become familiar with the visually confusing speed bumps."[142] If this happens, it follows that road planners will be responsible for subsequent ignorance by drivers of actual speed bumps caused by the intervention – at least as long as these effects occur within the posted speed limits. In addition, the recommendations state, relative to this particular nudge, that road planners should provide passive disclosure consistent with Rawls' Publicity principle, as well as easy paths to filing complaints. Consulting the general discussions of the experiment in the media confirms that this is actually what has occurred. In general, such issues of unintended consequences may be noticed to overlap with the issue referred to by Selinger and Whyte as "semantic variance".[143]

3. Non-transparent type 1 nudges: Manipulating behaviour

However, passive disclosure and making complaints easy to file should not be regarded as a sufficient precaution when it comes to non-transparent type 1 nudges. Since these are non-transparent to citizens, their application constitutes the use of both technical and psychological manipulation. Intending to change people's behaviour or the consequences thereof by e.g., decreasing the size of plates, substituting small and fat glasses with high and lean ones, and rearranging the cafeteria, may not be regarded as transparent to those whose behaviour one is trying to influence. Yet, it is just as important to notice that this type of manipulation is not one of "choice", but of automated behaviours and their consequences. The automatic, instinctive and learned behaviours that non-transparent type 1 nudges operate upon are generally not the result of any conscious and reflective decision-making. If one neglects this difference a conflation of issues pertaining to the manipulation of active and reflected choice, or autonomous decision-making, with issues pertaining to the manipulation of behaviour readily occur.

In conclusion, non-transparent type 1 nudges are nudges that qualify for the popular characterization

141 Sean D. Hamill, "To Slow Speeders, Philadelphia Tries Make-Believe, 2008, available on the Internet at: <http://www.nytimes.com/2008/07/12/us/12bump.html> (last accessed on 10 January 2013).

142 Tom Vanderbilt, *Traffic: Why we drive the way we do (and what it says about us)* (New York: Vintage Books, 2009).

143 Evan Selinger and Kyle Whyte, "Is there a right way to nudge?", *supra* note 125, at p. 29.

of 'operating under the radar' of citizens.[144] Their application results in behaviour change by means of the technical, as well as the psychological manipulation of citizens. As a result, citizens are in general only capable of avoiding their effects as a matter of principle. Avoiding it within a complex everyday setting seems much more difficult, if not impossible. Thus, non-transparent type 1 nudges may be evaluated as truly paternalistic interventions. Yet, they do not intervene with the reflected thinking or conscious choices of citizens, but rather operates in the background of their private and public lives.

We believe that this evaluation of non-transparent type 1 nudges widens the responsibilities for policy-makers and other choice architects even more than transparent type 1 nudges. While the use of nudges of this type cause no direct disturbances to the lives of citizens, they nevertheless leave citizens unaware of their influence on behaviour and the consequences thereof. This means that, besides being responsible for the effect of these interventions and their possible side effects, choice architects are also responsible for ensuring that interventions are rooted in democratic procedures. Hence, they should not only make sure that the ends pursued are carefully calibrated with what citizens judge to be in their interests. They should also make sure that the intentions behind and the means by which behavioural change is pursued are actively disclosed and possibly consented to at least in general (i.e. type-consent). This will clarify that choice architects are responsible for the nudges' effects as well as possible side effects, not the individual citizen, who cannot be expected to manoeuvre around such nudges in complex everyday behavioural contexts.

One may take as an example an employer deciding to decrease plate-sizes, substitute glasses and re-design the cafeteria in order to reduce calorie intake among employees. While it may seem a bit of overkill, the right precaution in this case seems to us to be that the choice architect, i.e. the employer, actively provides information to employees stating the reasons for and measures by which such steps are taken. Only in this way may the employer avoid the accusation of manipulating with the behaviour and consequences thereof since she has provided the possibility for open debate and the awareness amongst employees of possible effects and subsequent side effects. If, for instance, employees should notice that they are beginning to snack unhealthy foods in between meals, such active disclosure will allow them to make the connection between intervention and this side effect, and in turn to engage with the employer in order to actively play a part in shaping the environment of their own lives. Still, one should not describe the use of this type of nudges as the 'manipulation of choice' as done e.g. by the blog Spiked or Burgess[145] since what is actually manipulated is automated behaviours and non-reflective habits.

4. Non-transparent type 2 nudges: Manipulating choice

With non-transparent type 2 nudges, the framework makes it clear that nudges of this type may be rightfully characterized as straightforward "manipulation of choice". This type of nudges works by psychologically manipulating citizens through underhanded, deceptive, and possibly even abusive tactics. Affecting behaviour change by the clever framing of risks, subtle goal-substitution, the use of subliminal cues, anchoring and priming, or using lotteries to get people to overestimate the chance of obtaining a rare effect are all measures that fall into this category. It is with regard to this type of nudges that token-transparency, as noted by Bovens,[146] may be expected to undermine effect or cause controversy in so far the intention behind and the means by which behaviour change is pursued are in conflict with the interests of citizens as judged by themselves. Assuming that individual reflected preferences made in accordance with the self-images of citizens constitute the core of autonomous decision-making, as judged by the norms of democracy, as well as public opinion, non-transparent type 2 nudges constitute the most controversial type of nudges, because citizens are used as mere tools rather than treated as ends.

In this way non-transparent type 2 nudges may be characterized as cases of straightforward paternalism. While citizens in principle are free to choose otherwise, the lack of transparency makes this unlikely in practice. Furthermore, because the choices of citizens are results of reflective decision-making, basic norms for assigning responsibility to decision-makers are ascribed to citizens, whose actions are ac-

144 Luc Bovens, "The Ethics of Nudge", *supra* note 14, at p. 4.

145 O´Neill, "A message to the illiberal Nudge Industry: Push off", *supra* note 13, at p. 4

146 Bovens, "The Ethics of Nudge", *supra* note 14, at p. 4.

tually results of manipulation by the choice architect. In this sense non-transparent type 2 nudges may be perceived as even more invasive than traditional policy measures, such as explicit regulation and the control of incentives and sanctions. The citizen is manipulated into compliance with the ends of the choice architect in a way that does not by itself promote debate or consent, while at the same time ascribing full responsibility of the action to the citizen.

Given these facts, it is difficult to find acceptable places for the responsible use of non-transparent type 2 nudges in democratic societies. Thus, we believe that policy-makers and other choice architects should generally abstain from the use of this type of nudges short of explicit individual token-consent. Insofar such nudges are used without obtaining consent; the policy-maker or choice architect takes full responsibility, not only for the effects and possible side effects of the intervention, but also for the use of citizens as mere tools. Even active disclosure and type-consent will not do. Individual responsibility of actions usually pertains to singular actions and not just types of actions. One of the only areas of policy-making where we find that the use of non-transparent type 2 nudges might be judged as acceptable is in ensuring compliance with certain important laws resulting from democratic public decision-making, where infringement may cause direct harm to other people. When citizens break the laws of society in ways that pose a threat to the safety and freedom of fellow citizens, there might be instances where the use of this type of nudges may be deemed responsible as a result of general consent, rather than individual token consent.

Besides the use of subliminal advertising discussed by Thaler and Sunstein, we offer attempts to influence a patient's medical decision-making concerning treatments by the subtle framing of risk and choice-options to influence choice in a non-transparent way. Such decision-making has usually been left for the patient because it has been deemed a relevant choice for the patient to make. Intentional framing of the choice is thus a way to ascribe responsibility to the patient for making a certain choice while simultaneously manipulating her to make that choice. Of course, as noted by Thaler and Sunstein, it is impossible to avoid framing decision-making one way or another. However, there is an important difference between trying to frame decision-making in a way that nudges people towards a particular decision, rendering autonomous choice a mere fiction, and framing decision-making in a way that respects the reasons for providing a person with the power and responsibility of making that choice. If one finds such 'neutral' framing difficult to imagine, one need only think of voting ballots (possibly using randomized order of candidates) and the relatively neutral framing of choice options on formulas for registering as an organ donor found in most countries as examples. Thus, again, it seems that as a matter of everyday practice, the responsibilities described with regard to non-transparent type 2 nudges are usually recognized and followed.

VII. Summary

In this paper we have argued that the anti-nudge position is not a literal non-starter. While it is true that our choices and, more generally, our behaviour are always being influenced by context, intentional intervention aimed at affecting behaviour change ascribes certain responsibilities to the public policy-maker or choice architect that are not addressed by Thaler and Sunstein. Further, we argued that these responsibilities cannot be waived by pointing out that nudges are liberty preserving and thus in principle leave citizens free to choose otherwise, since while this in principle is true, one can hardly appeal to it in a practical context where the nudge approach to behavioural change is applied exactly because we tend to fall short of such principles.

This ultimately seemed to leave us to the critics, with a public policy approach based on the manipulation of citizens' choices. However, against this we have argued that the characterization of nudging as the manipulation of choice is too simplistic. Both classical economic theory and behavioural economics describe behaviour as always resulting from choices, but the psychological dual process theory that underpins behavioural economics, used by Thaler and Sunstein, distinguishes between automatic behaviours, and reflective choices. Nudging always influences the former, but it only sometimes affects the latter. The conceptual implication of this is that nudging only sometimes targets choices.

However, this still left the accusation that nudging is about "manipulation." Against this claim we argued that Thaler and Sunstein's appeal to Rawls' Publicity Principle is insufficient as a safeguard against non-legit state manipulation of people's choices. Instead, we introduced an epistemic distinction between

transparent and non-transparent nudges, which serves as a basis for distinguishing the manipulative use of nudges from other kinds of uses. In the end the result is a conceptual framework for describing the character of four broad types of nudges that may provide a central component for more nuanced ethical considerations and a basis for various policy recommendations. It is our hope that this framework may clear up some of the confusion that surrounds the normative discussion of the nudge approach to behavioural change and better inform its adoption in public policy-making.