Institute for Research in Economic and Fiscal issues

IREF Working Paper Series

An Impossibility Result on Nudging Grounded in the Theory of Intentional Action.

Sergio Beraldo

IREF Working Paper No. 201801

January 2018

Institute for Research in Economic and Fiscal issues

An Impossibility Result on Nudging Grounded in the Theory of Intentional Action. *

Sergio Beraldo
Department of Economics and Statistics
University of Napoli 'Federico II' & CSEF & IREF

January 8, 2018

Abstract

The critical literature on nudging lacks a formal, smoking-gun argument against the use of nudges that trigger automating behaviour without involving reflective thinking, such as setting up a default option or arranging the alternatives available to consumers in such a way as to make a given choice more likely. I offer an impossibility result, grounded in the theory of intentional action, that provides such an argument. I prove that if individuals are not open to money-pump manipulation and nudges are motivationally irrelevant, any induced choice is unintentional and just reflects the preferences of the choice architect. Autonomy is therefore violated, and nudging proves to be inconsistent with liberal principles at a fundamental level.

Keywords: *Nudging*, *Manipulation*, *Autonomy*.
JEL codes: D03, D6

*I thank Luigino Bruni, Enrico Colombatto, Till Grüne-Yanoff, Valentino Larcinese, Mozaffar Qizilbash, Annalisa Scognamiglio, Robert Sugden and all the seminar participants at the Network for Integrated Behavioural Science Conference (UEA, April 2016), the IREF workshop (St. Mary University, London, June 2016), The Italian Society of Law and Economics (University of Turin, December 2016), the Center for Studies in Economics and Finance (CSEF, Napoli, May 2017), for discussion. The usual disclaimers apply.

> In a sense, behavioral economics extends the paternalistically protected category of idiots to include most people, at predictable times (Camerer et al., 2003).
>
> Enlightement is man's emergence from his self-incurred immaturity. Immaturity is the inability to use one's own understanding without the guidance of another (Kant, 1784).

1 Introduction

Do flaws in human decision-making legitimate governments and agencies to nudge people towards better choices? A tendency to give an affirmative answer to this question is nowadays very popular both in US and in Europe, as witnessed by the establishment of some influential agencies such as the *Social and Behavioral Sciences Team* (US) and the *Behavioural Insight Team* (UK), as well as by the European Commission-funded project *Nudge-It*.

There are some reasons behind this popularity. First, nudging people has a very high benefit to cost ratio. In times of austerity this is certainly attractive. Second, nudging does not seem to violate any way individual freedom, making regulation appealing even to conservatives (Camerer et al., 2003). Third, nudging people is cool; social scientists are not the only ones who seem to experience an extraordinary sense of surprise when behavioural findings are discovered to work at the end.

Unfortunately, not all that glitters is gold.

Scholars have emphasized that nudging might be not compatible with liberal positions (Grüne-Yanoff, 2012); threatens people's control over their own evaluations and deliberation (Hausman and Welch, 2010); may produce infantilisation as a long-term effect, i.e. decreased responsibility in matters

regarding one's own welfare (Bovens, 2009). It has also been argued that the libertarian paternalistic approach of nudging uses a model of the human being that is psychologically and philosophically problematic (Infante et al., 2016).

The critical literature on nudging lacks, however, a formal, smoking-gun argument against its use. I provide such an argument.

I take for granted that nudges are called for because of flaws in individual decision-making, and *work by making use of those flaws* (Hausman and Welch, 2010). This helps distinguishing nudges from many other things, notably from the provision of information. As it helps in distinguishing between two types of nudges: one directed at stimulating reflective thinking; the other at triggering automating behavior without involving reflective thinking, what Hansen and Jespersen (2013) call type 1 nudges in homage to Dual Process Theory (e.g. Kahneman, 2011). My argument does not apply to nudges directed at stimulating reflective thinking (e.g. displaying a reminder, setting up a mandatory cooling-off period); these nudges might in fact be devised to make deliberation more conscious. My argument is instead directed towards those nudges that trigger automating behaviour without involving reflective thinking. Setting up default options in such a way as to increase organ donations or contributions to a pension scheme are pertinent examples, as well as arranging the options available to consumers to increase the probability that a given choice will take place.

With regard to nudges of this kind, I prove that if individuals are not open to money-pump manipulation and nudges are motivationally irrelevant, any induced choice is unintentional and just reflects the preferences of the choice

architect; autonomy is therefore violated.

The assumption that denies money-pump manipulation precludes that individuals are open to exploitation and saves their opportunity for intentional action (if money-pump is possible, individuals can be induced to choose whatever option). For what concerns the inability of nudges to motivate behaviour, this assumption naturally follows from what is commonly understood by nudge (Thaler and Sunstein, 2008; Hausman and Welch, 2010).

As far as autonomy is concerned, it can be interpreted in different ways. In a beautiful and for some aspects controversial contribution, Isaiah Berlin explains the closeness between autonomy and positive freedom (Berlin, 1969). He also clarifies why autonomy, meant as positive freedom, can be compatible with a situation in which there is no autonomy in a fundamental sense, i.e. independence of motivations that induce a person to act (Christman, 2015).

Hausman and Welch (2010) refer to the control an individual has over her own evaluation, deliberation and choice. According to their perspective, autonomy is not granted by the absence of coercion, and it is certainly possible that the lack of autonomy goes hand in hand with formal libertarianism.

I approach autonomy by negation. I say that an individual does not act autonomously or that her autonomy is violated whenever her choices just reflect the preferences of another subject who deliberately acts to shape them.

The condition requiring the choice architect to deliberately act to shape people's choices is crucial, for, as Thaler and Sunstein (2008) reminds us, there is no such thing as a *neutral* design. Therefore, even if it is unavoidable that individuals' choices are conditioned by some features of the choice

environment, what in my view is key for autonomy is that these features are not intentionally modelled.

The paper is organized as follows. In Section 2 I introduce the setting and justify my approach, grounded in the use of extended psychological preferences. In Section 3, grounding in the modern theory of action, I set a necessary and sufficient condition for intentional action. Section 4 states the main results. Section 5 concludes.

2 The setting

Let $\hat{A} = \{x, y, z, \dots\}$ be the universal set of alternatives. A choice problem is a nonempty subset of $\hat{A}$. Following Bernheim and Rangel (2009) and Salant and Rubinstein (2008), I enrich the standard model by a set $\Sigma \equiv \{\sigma\}$ of ancillary conditions (or frames). Examples of ancillary conditions are the manner in which options are presented or the labeling of a particular option as the *status quo*. A generalized choice situation $G = (A, \sigma)$ is a choice set $A \subseteq \hat{A}$ coupled with an ancillary condition, σ.

I consider a binary relation $\succeq_i^{\Omega}$ representing i's extended preferences over $\Omega = \hat{A} \times \Sigma$. I write $(x, \sigma) \succeq_i^{\Omega} (y, \sigma')$ to mean that alternative x coupled with ancillary condition σ is at least as good as alternative y coupled with σ'. The binary relation $\succeq_i^{\Omega}$ is not assumed to be complete.

Once the ancillary condition is set, only i's contingent preferences, i.e. i's preferences contingent to the choice of the ancillary condition, are relevant. Hence, for any $x, y \in A$, $x \succeq_i^{\sigma} y$ means that, given ancillary condition σ, x is at least as good as y for individual i ($\succ_i^{\sigma}$ and $\sim_i^{\sigma}$ indicate the asymmetric

and symmetric component of $\succeq_i^\sigma$ respectively).

Clearly, $x \succeq_i^\sigma y \leftrightarrow (x, \sigma) \succeq_i^\Omega (y, \sigma)$.

2.1 A motivational argument to justify the use of extended preferences

In what is generally considered a paradigmatic case of Libertarian Paternalism (Bovens, 2009), Sunstein and Thaler consider a cafeteria manager, Carolyn, who is aware that the way in which the food is displayed and arranged affects the choices that consumers make (Thaler and Sunstein, 2008). In particular, the manager knows that consumers have a tendency to choose food items which come first in the line.

As the manager has the power to change the order in which items are displayed, she is then able to increase the consumption of some foods at the expense of others. How should she decide which item to present first?

As there is no such thing as a neutral design, according to Sunstein and Thaler it is legitimate for the manager to try to influence people's behavior in order to make their lives better, *as judged by themselves*. So they argue in favour of practices, such as nudging, that steer people's choices in appropriate directions, where appropriateness is to be evaluated from the privileged standpoint of the individuals' superhuman ego.

If we suppose that there are two food items (Fruit, Dessert), it is possible to read the cafeteria example as a generalized choice situation in which the ancillary condition specifies which item is displayed first, i.e. $G = (A, \sigma)$ where $A = \{Fruit, Dessert\}$ and $\sigma \in \{\sigma_{Dessert}, \sigma_{Fruit}\}$.

Sunstein and Thaler, as well as any other supporter of libertarian paternalism, take for granted that ancillary conditions are irrelevant to judgment, hence they assume that ancillary conditions do not motivate choice.

There are valuable reasons to argue, however, that the value consumers place on a given option might change along with the ancillary condition.

As an example, consider that preferences might be such that an individual prefers choosing *Fruit* when *Fruit* comes next, rather than choosing *Fruit* when it comes first, i.e. $(Fruit, \sigma_{Dessert}) \succ_i^{\Omega} (Fruit, \sigma_{Fruit})$. An individual might place value on choosing Fruit when it comes next, because this is seen as a way to temper her character. In a similar fashion, and more relevantly, although the evidence suggests that an opt-out system of organ donations makes many more persons *willing* to be donors than an opt-in system (Johnson and Goldstein, 2003), the value of being an organ donor might differ both at the individual and the societal level when such choice is made in the context of an opt-in system. Thus an individual could well prefer being an organ donor in an opt-in system rather than being a donor when the system allows to opt-out, $(donor, \sigma_{Opt\text{-}in}) \succ_i^{\Omega} (donor, \sigma_{Opt\text{-}out})$.

My proposal of considering extended preferences defined both on options and ancillary conditions is similar to Kőszegi and Rabin's proposal of assuming behaviour as choice set dependent, for this would allow to reconcile choices with standard axioms (Kőszegi and Rabin, 2008, 2007). As they argue, somebody could choose *Dessert* from the choice set $\{Dessert, Fruit\}$ but be better off with the choice set $\{Fruit\}$ if she has choice-set-dependent preferences, where

$$(Fruit|\{Fruit\}) \succ (Dessert|\{Dessert, Fruit\})$$

and

$$(Dessert|\{Dessert, Fruit\}) \succ (Fruit|\{Dessert, Fruit\}),$$

for the option of eating dessert creates an unpleasant sensation of temptation. Similarly, a consumer could choose *Fruit* from the choice set $\{Fruit, Dessert\}$ when the ancillary condition is σ_{Fruit} but be better off when the ancillary condition is $\sigma_{Dessert}$. This calls into question whether the hypothesis that ancillary conditions are motivationally irrelevant, i.e. that ancillary conditions are of no value whatsoever as long as the motivations driving behaviour are concerned, is an innocuous hypothesis as it is pretended to be.

Given the discussion above, a suitable definition of motivationally-irrelevant ancillary condition is provided by the following:

Definition 1 (*Motivationally-irrelevant ancillary condition*)**.** A pair of ancillary conditions $(\sigma, \sigma') \in \Sigma$ is said to be motivationally-irrelevant to i's choice whenever for any alternative x, neither $(x, \sigma) \succ_i^{\Omega} (x, \sigma')$ nor $(x, \sigma') \succ_i^{\Omega} (x, \sigma)$.

In other words, ancillary conditions are motivationally-irrelevant when they are not able to make a given option more desirable. Going back to a previously discussed example, the kind of system in force, whether opt-in or opt-out, should not affect the decision of being an organ donor, for neither $(donor,\sigma_{Opt\text{-}in}) \succ_i^{\Omega} (donor,\sigma_{Opt\text{-}out})$ nor $(donor,\sigma_{Opt\text{-}out}) \succ_i^{\Omega} (donor,\sigma_{Opt\text{-}in})$.

2.2 The publicity principle

In an attempt to reply to their critics, Thaler and Sunstein endorse a publicity principle. They attribute the definition of this principle to John Rawls (Rawls, 1971), but in fact it was stated by Immanuel Kant (1795) as a *trascendental formula of public law*, according to which

> All actions relating to the rights of men are wrong, if the maxims from which they follow are inconsistent with publicity [p. 185]

The publicity principle bans governments from selecting a policy that it would not be able or willing to defend publicly to its own citizens. As emphasized by Thaler and Sunstein (2008):

> 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...the publicity principle is a good guideline for constraining and implementing nudges...the same conclusion holds for legal default rules. If government alters such rules to encourage organ donation, it should not be secretive about what it is doing ... government officials...[encouraging] people to register as organ donors... should be happy to reveal both their methods and their motives [p.15]

Methods and motives must indeed be agreed upon by citizens. Governments should not be secretive about what they are doing.

This disposition to be open about the techniques that are being employed to influence people's choices, is what, according to Sunstein and Thaler, ultimately allows to distinguish nudges from other, more intrusive forms of persuasion.

Perhaps unintendedly, this disposition to be open implies that governments *de facto* trigger people's reflection about the features of the choice environment. Some of these features, although unable to provide relevant economic incentives, can in principle characterize the options in such a way as to motivate choice.

To be more explicit about this, let us consider a legal default rule set up to encourage organ donation. One of such rules has been implemented in France in January 2017. For the publicity principle to be respected, government officials should not be secretive about the reasons why such a legal default rule was issued; so they should be *happy* to reveal that such rule produces a greater number of individuals who are *willing* to be organ donors (Johnson and Goldstein, 2003).

Revealing the reasons why an opt-out system for organ donation is preferred over an opt-in system, naturally induces people to meditate on what being an organ donor means in either of the two systems. This triggers people's reflection about the features of the choice environment that are able to characterize options from a motivational perspective, even though such features do not provide any economic incentive.

Thus one may be perfectly coherent in preferring to be a donor only when the system is opt-in and not being a donor when the system is opt-out.

Formally, let $G = (A, \sigma)$ be the generalized choice situation, where $A =$

$\{donor, non\ donor\}$ and $\sigma \in \{\sigma_{Opt-in}, \sigma_{Opt-out}\}$. It is plausible to have the following preferences

$$(donor,\ \sigma_{Opt\text{-}in}) \succ_i^{\Omega} (non\ donor, \sigma_{Opt\text{-}in})$$
$$(non\ donor,\ \sigma_{Opt\text{-}out}) \succ_i^{\Omega} (donor,\ \sigma_{Opt\text{-}out}).$$

Moreover

$$(donor, \sigma_{Opt\text{-}in}) \succ_i^{\Omega} (donor, \sigma_{Opt\text{-}out})$$
$$(donor, \sigma_{Opt\text{-}in}) \succ_i^{\Omega} (non\ donor, \sigma_{Opt\text{-}out}).$$

In other words, an individual might prefer to be a donor in an opt in system. However, her willingness to be a donor is not unlimited. If the government frustrates her motivations by setting up an opt-out system, then the individual might prefer not to be a donor.

Clearly, libertarian paternalists might argue that the government - once its plans become public - will be able to implement a legal default rule for organ donation only if it is supported by the majority of the people. Indeed, as the government is constrained not to be secretive about what is doing, a public debate would presumably follow. This is democracy. My answer is pragmatic. As in many other cases - and libertarian paternalists should agree with this - minorities would be obliged to accept an inferior (from their perspective) state of affairs. This partially contradicts the fascinating promise of Libertarian Paternalism, what makes it really appealing, i.e. that only Pareto improvements will be delivered: no one looses, some gain. Put in this way, there are reason to believe that even the least ardent libertarian would reject the libertarian paternalistic perspective.

3 Intentional action

In contemporary analytical philosophy, it is due to Margareth Anscombe (1957) and Donald Davidson (1963), the merit of having revitalized the debate on the nature of agency (the exercise or manifestation of an agent's capacity to act). As emphasized by Schlosser (2015), Anscombe and Davidson have contributed to define a standard conception of action, grounded in the following two claims: the notion of intentional action is a primitive of the notion of action, it is more fundamental in that any action has to be an intentional action under some description of the action itself; there is a strict connection between an intentional action and acting for a reason, i.e. intentional actions are performed for reasons (Mele and Moser, 1994).

As clarified by Margareth Ascombe

> What distinguishes actions which are intentional from those which are not? The answer that I shall suggest is that they are the actions to which a certain sense of the question 'Why?' is given application; the sense is of course that in which the answer, if positive, gives a reason for acting [p. 9]

Intentional actions are actions done for a reason (Davidson, 2003).

In a similar vein, Mele and Sverdlik (1996) write:

> An agent, S, A-ed intentionally only if S A-ed for a reason.

This is probably the most widespread way of understanding an intentional action.

Davidson (2003) quotes Stuart Hampshire (1965) who writes:

> A wants to do X is equivalent to *other things being equal*, he would do X, if he could.

The *other things being equal* clause is meant by Davidson as signifying *provided there is not something he wants more.*

With this interpretation in mind, Davidson puts Hampshire's principle as follows:

> If an agent wants to do x more than he wants to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally (p.27).

An agent who believes herself free to choose between x and y and intentionally chooses either x or y, will intentionally choose x if she wants x more than she wants y, i.e. if she prefers x to y. Notice *en passant* that from these premises Davidson draws the conclusion that it is false that there are incontinent actions, for, by definition, incontinent actions are those actions that are taken although they are not preferred in a relevant sense.

The previous passage establishes a necessary condition for intentional action. Suppose the agent believes himself free to choose among different alternatives. Also suppose that the agent wants alternative x more than alternative y. If the agent chooses y, then such choice does not count as intentional. Clearly, this conclusion is strictly related to Davidson's premise that intentional actions are actions done for a reason and is rooted in a very long philosophical tradition that can be traced back to Aristotle and David Hume.

By saying that if an agent intentionally picks up an alternative in the choice set, then such an alternative must be the one she wants most, it is said that a necessary condition for intentional action is that the agent chooses what she prefers.

Definition 2 (Intentional action). A necessary condition for alternative x to be intentionally chosen by i at t, is that, at t, i prefers x over other feasible alternatives, including not acting, by virtue of her desires and beliefs.

Given one's desires and beliefs about the foreseeable consequences of any of the feasible choices, if one prefers x over other feasible alternatives and chooses intentionally, she will intentionally choose x. This is a necessary, although not sufficient, condition for intentional action [1]: i might unintentionally choose x even if x is the alternative i wants most (e.g. Mele and Moser, 1994).

Clearly, intentionality requires that the agent chooses what she prefers in the relevant sense, not in the most fundamental sense. According to Plato, Socrates died drinking a cup of poison hemlock; he intentionally drank the poison, so - in a relevant sense - he preferred drinking the poison; however, this was not what Socrates probably wanted in a most fundamental sense. Given the set of circumstances, Socrates drank the poison because he preferred to do so, and his action counts as intentional.

[1] See Bratman (2009) for an original discussion of this point.

4 An impossibility result

In this Section I state my main results. Remember that I consider a binary relation $\succeq_i^{\Omega}$ representing i's extended preferences over $\Omega = \hat{A} \times \Sigma$, where $\hat{A}$ is the universal set of alternatives and Σ is the set of ancillary conditions or frames. I write $(x, \sigma) \succeq_i^{\Omega} (y, \sigma^{'})$ to mean that alternative x coupled with ancillary condition σ is at least as good as alternative y coupled with $\sigma^{'}$.

As the binary relation $\succeq_i^{\Omega}$ is not assumed to be complete, it might be the case that neither $(x, \sigma) \succeq_i^{\Omega} (y, \sigma^{'})$ nor $(y, \sigma^{'}) \succeq_i^{\Omega} (x, \sigma)$ holds, what I indicate by $(x, \sigma) \perp (y, \sigma^{'})$.

As it is not assumed completeness of $\succeq_i^{\Omega}$, it is consequently not required that $\succeq_i^{\Omega}$ respects transitivity, for when preferences are not complete, a violation of transitivity can occur. This may indeed happen whenever the individual weakly prefers (x, σ) to $(y, \sigma^{'})$ and $(y, \sigma^{'})$ to $(z, \sigma^{''})$, but is unable to compare (x, σ) with $(z, \sigma^{''})$, i.e. $(x, \sigma) \perp (z, \sigma^{''})$. A violation of transitivity opens the way to the possibility that the individual is money pumped. To avoid manipulation and to allow for the possibility of intentional action, I assume *behavioural transitivity* (Mandler, 2005):

Assumption 1 (Behavioural Transitivity, **BT**). *$\succeq_i^{\Omega}$ is behaviourally transitive, i.e. there do not exist three alternatives x, y, z and three ancillary conditions $\sigma, \sigma^{'}, \sigma^{''}$, such that $(x, \sigma) \succeq_i^{\Omega} (y, \sigma^{'})$, $(y, \sigma^{'}) \succeq_i^{\Omega} (z, \sigma^{''})$ and $(z, \sigma^{''}) \succ_i^{\Omega} (x, \sigma)$.*

As clarified above, the following assumption naturally follows from what is commonly understood by nudge.

Assumption 2 (Motivational irrelevance, **MI**). *A nudge is a motivationally irrelevant ancillary condition.*

According to Thaler and Sunstein (2008), the libertarian aspect of their strategies lies in the straightforward insistence that people should be «free to do what they like and to opt out of undesirable arrangements if they want to do so». In stating that people are free to choose what they prefer, their strategy apparently ensures that the necessary condition for intentional action is respected.

Assumption 3 (Intentionality, **I**). *Choice is intentional. If x is systematically chosen over y when the ancillary condition is σ, i.e. $c_{i\sigma} = x$, then $(x, \sigma) \succ_i^{\Omega} (y, \sigma)$.*

Now I prove the main result of the paper, i.e. that the previous three assumptions are incompatible.

Proposition. *For any given pair of alternatives (x, y) and any given pair of ancillary conditions $(\sigma, \sigma') \in \Sigma$ such that $c_{i\sigma} = x$ and $c_{i\sigma'} = y$, if $\succeq_i^{\Omega}$ is behaviourally transitive and ancillary conditions are motivationally irrelevant, then i is acting unintentionally.*

Proof. If $c_{i\sigma} = x$ and $c_{i\sigma'} = y$ but $(x, \sigma) \perp (y, \sigma)$ and $(x, \sigma') \perp (y, \sigma')$ by (**MI**) and (**I**) the individual is acting unintentionally and the proof is complete. Suppose therefore that $(x, \sigma) \succ (y, \sigma)$ and $(y, \sigma') \succ (x, \sigma')$. As ancillary conditions $(\sigma, \sigma') \in \Sigma$ are motivationally-irrelevant to i's action, for any pair of alternatives (x, y): $(x, \sigma) \sim_i^{\Omega} (\text{or} \perp) (x, \sigma')$ and $(y, \sigma) \sim_i^{\Omega} (\text{or} \perp) (y, \sigma')$.

Suppose without loss of generality that $(x,\sigma) \sim_i^\Omega (x,\sigma')$ and $(y,\sigma) \sim_i^\Omega (y,\sigma')$ (the proof is the same whenever either $(x,\sigma)\perp_i^\Omega(x,\sigma')$ or $(y,\sigma)\perp_i^\Omega(y,\sigma')$ or both).

By (**I**) and (**MI**)

$(y,\sigma') \succ_i^\Omega (x,\sigma') \sim_i^\Omega (x,\sigma)$, $(x,\sigma) \succ_i^\Omega (y,\sigma) \sim_i^\Omega (y,\sigma')$,

and, by (**BT**),

$(y,\sigma') \succ_i^\Omega (x,\sigma') \sim_i^\Omega (x,\sigma) \rightarrow (x,\sigma) \not\succ_i^\Omega (y,\sigma')$

$(x,\sigma) \succ_i^\Omega (y,\sigma) \sim_i^\Omega (y,\sigma') \rightarrow (y,\sigma) \not\succ_i^\Omega (x,\sigma)$

or $(y,\sigma) \not\succ_i^\Omega (x,\sigma) \not\succ_i^\Omega (y,\sigma') \rightarrow (y,\sigma) \preceq_i^\Omega (x,\sigma) \preceq_i^\Omega (y,\sigma')$

implying $(x,\sigma) \preceq_i^\Omega (y,\sigma') \sim_i^\Omega (y,\sigma)$, and, by (**BT**), $(x,\sigma) \not\succ_i^\Omega (y,\sigma)$

which violates (**I**), the assumption that choice is intentional.

□

One's autonomy is necessarily violated whenever her choice reflects someone else's preferences. This is the case when the ancillary condition is set by the choice architect. Let $\succeq^P$ be its preference relation, defined over ancillary conditions.

Corollary. *For any pair of alternatives (x,y) and any pair of motivationally - irrelevant ancillary conditions $(\sigma,\sigma') \in \Sigma$, such that $c_{i\sigma} = x$ and $c_{i\sigma'} = y$, if the ancillary condition is intentionally chosen, i's autonomy is violated.*

Proof. By the previous proposition, for any pair of alternatives (x,y) and any pair of motivationally-irrelevant ancillary conditions $(\sigma,\sigma') \in \Sigma$ such that $c_{i\sigma} = x$ and $c_{i\sigma'} = y$, i is acting unintentionally. Suppose σ is intentionally chosen. This implies that $c_{i\sigma} = x \succ^P c_{i\sigma'} = y$. As i's choice reflects the preference of the choice architect, i's autonomy is violated. □

Notice that it is not from the fact that an ancillary condition is selected that violation of individual autonomy is brought about, but by the fact that the ancillary condition is intentionally chosen. This implies a superimposition of the choice architect's preference over the individual' preferences.

5 Conclusion

In this paper I offer a formal proof that nudged people act unintentionally and that their autonomy is violated. More precisely, I prove that if individuals are not open to money-pump manipulation and nudges are unable to motivate behaviour (as it is required by the very definition of nudging), then they act unintentionally. If this is the case, their choices just reflect the preferences of the choice architect; individuals are, in other words, manipulated.

My argument does not apply to nudges directed at stimulating reflective thinking (e.g. displaying a reminder, setting up a mandatory cooling-off period); these nudges are devised to make deliberation even more conscious. My argument is directed towards those nudges that trigger automating behaviour without involving reflective thinking. Setting up default options, in such a way as to increase organ donations or contributions to a pension scheme is an example of these kind of nudges, as well as arranging the options available to consumers in such a way as to increase the probability that a given alternative will be chosen.

My analysis shows that the use of these nudges is incompatible with liberal principles at a very fundamental level.

References

Berlin, I. (1969). *Four essays on Liberty*. Oxford University Press.

Bernheim, B. D. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics*, 124(1):51–104.

Bovens, L. (2009). *The Ethics of Nudge*, pages 207–219. Springer Netherlands, Dordrecht.

Bratman, M. E. (2009). Intention, practical rationality, and selfgovernance. *Ethics*, 119(3):411–443.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for 'asymmetric paternalism'. *University of Pennsylvania Law Review*, 151:1211–1254.

Christman, J. (2015). Autonomy in moral and political philosophy. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2015 edition.

Davidson, D. (2003). *Essays on Actions and Events*. Oxford University Press.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 38(4):635–645.

Hampshire, S. N. (1965). *Freedom Of The Individual*. Harper Row.

Hansen, P. G. and Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1):3âĂŞ28.

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

Infante, G., Lecouteux, G., and Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, 23(1):1–25.

Johnson, E. J. and Goldstein, D. (2003). Do defaults save lives? *Science*, 302(5649):1338–1339.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Kant, I. (1917/1795). *Perpetual Peace*. GEORGE ALLEN & UNWIN LTD, London.

Kant, I. (1991/1784). An answer to the question: "what is enlightenment? In Reiss, H., editor, *Kant: Political Writings*. Cambrige: Cambridge University Press.

Kőszegi, B. and Rabin, M. (2007). Mistakes in choice-based welfare analysis. *American Economic Review*, 97(2):477–481.

Kőszegi, B. and Rabin, M. (2008). Choices, situations, and happiness. *Journal of Public Economics*, 92(8-9):1821–1832.

Mandler, M. (2005). Incomplete preferences and rational intransitivity of choice. *Games and Economic Behavior*, 50(2):255–277.

Mele, A. and Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, 82(3):265–287.

Mele, A. R. and Moser, P. K. (1994). Intentional action. *Noûs*, 28(1):39–68.

Rawls, J. (1971). *A theory of justice*. Oxford University Press.

Salant, Y. and Rubinstein, A. (2008). (a, f): Choice with frames -super-1. *Review of Economic Studies*, 75(4):1287–1296.

Schlosser, M. (2015). Agency. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2015 edition.

Thaler, R. and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.