Nudging and Phishing:

A Theory of Behavioral Welfare Economics

(Job market paper)

David Jimenez-Gomez*

January 14, 2017

Click here for the most recent version

Abstract

Nudges, which are interventions that do not restrict choice, have become widespread in policy applications. I develop a general and tractable framework to analyze the welfare implications of nudges. In this framework, individuals suffer from internalities (their utility when choosing is different from their welfare-determining utility) and choice and welfare depend on the environment, which can be altered by the nudge. I show that, in order to design the optimal nudge, no knowledge of environment-independent preferences is required. This means that the social planner does not need to fully recover individual preferences, something which is especially difficult in the presence of internalities. In heterogeneous populations, the optimal nudge trades off correcting the internalities of biased individuals with psychological costs imposed by the nudge on all individuals. When taxes are also available, nudging is generally optimal as long as the government is not fully efficient in collecting revenue from taxation. I also analyze phishing, when firms change the environment to take advantage of consumers’ internalities. Competition does not necessarily reduce phishing and, when firms have incentives to phish, competition can be welfare-decreasing. I analyze nudging and phishing in general equilibrium, and characterize the optimal nudge. In contrast to recent empirical work, which finds that nudging can backfire in general equilibrium because firms raise prices in response to a nudge, I show that under perfect competition nudging is generally welfare-enhancing.

JEL: D03, D11, D60, H21.

Keywords: behavioral economics, welfare economics, nudge, phishing, general equilibrium.

*I thank my postdoctoral supervisor John List for his support in pursuing this research. My thanks for useful comments to Seth Blumberg, Eszter Czibor, Cristian Dagnino, Daniel Tannenbaum, and Citlalli Vivar; as well as the members of John List’s research group and the attendants of the Micro lunch seminar at the University of Chicago and the Theory lunch at MIT. Email: jimenezd@mit.edu
1 Introduction

Nudges, which are interventions that do not restrict choice, have grown exponentially in policy in the last decade: they are now officially implemented in 51 countries,\textsuperscript{1} in a wide range of domains, from health to personal finance (Thaler and Sunstein, 2008). Proponents of nudges argue that they can improve people’s decisions without being excessively paternalistic, while a prominent argument against nudges is that they have psychological costs that should be taken into account. Unfortunately, the lack of a general framework has prevented a rigorous economic analysis of those arguments.\textsuperscript{2} In this paper I propose a general and tractable model for analyzing nudges and their welfare consequences, therefore providing a framework in which to evaluate the arguments that have been made about nudges.

Why nudge? Besides the traditional reasons in Economics (such as externalities), governments who nudge also attempt to correct people’s internalities, which happen when individuals underweigh or ignore consequences of their own behavior for themselves (Herrnstein et al., 1993).\textsuperscript{3} This paper joins a recent literature on Behavioral Welfare Economics, which tackles the problem of designing policies that maximize social welfare in the presence of internalities.\textsuperscript{4} In this literature, individuals use a utility function to make a choice $x \in X$ (their decision utility), that differs from the actual hedonic value or well-being they experience when they consume $x$ (their experienced utility). Because of this discrepancy between decision utility at the time of making the choice, and experienced utility at the time of consumption, individuals are subject to internalities.\textsuperscript{5}

I start with a framework in which the individual uses decision utility to make choices in

\textsuperscript{1}Whitehead et al. (2014).
\textsuperscript{2}There is a recent and quickly growing literature on Behavioral Welfare Economics, which analyzes the welfare effect of nudges in particular contexts (Baicker et al., 2015; Allcott and Taubinsky, 2015). In contrast, my aim is to develop a more general model that can be applied to a variety of nudges.
\textsuperscript{3}The Behavioral Insights Team, which is the United Kingdom’s “Nudge Unit”, has as one of its main three objectives to act “wherever possible, enabling people to make better choices for themselves”, http://www.behaviouralinsights.co.uk/about-us.
\textsuperscript{4}Recent papers include Mullainathan et al. (2012b), Allcott and Taubinsky (2015), Allcott et al. (2014), Alcott and Kessler (2015), Baicker et al. (2015).
\textsuperscript{5}This framework can also be interpreted in a broader context, where the social planner wants to design policies using a different utility from that with which the individual chooses.
set $X$, but feels the consequences of the choice under experienced utility, and both depend on the environment $e$. In doing so, I follow an old tradition in Psychology (Tversky and Kahneman, 1981), that has been recently incorporated into Economics (Salant and Rubinstein, 2008; Bernheim and Rangel, 2009), which considers that the choices people make are influenced by the environment in which they are made – the environment $e$ is sometimes known as a “frame” in the psychology literature, and more recently as “choice architecture” (Thaler and Sunstein, 2008). People make different choices in different environments, because their decision utility $u^{\text{dec}}(x|e)$ depends on $e$. This framework allows me to study nudges, which Thaler and Sunstein (2008) define as interventions that “alter people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives”. In my framework, nudges change the environment $e$ but not the choice set $X$.

A widespread criticism of nudges argues that “the social planner needs to know people’s preferences better than they themselves do” – what Sunstein (2014), in his defense of nudges against such criticism, called the epistemic argument. In support of Sunstein’s position, I show in Proposition 1 that the optimal nudge can be designed without knowledge of the environment-independent preferences of the individual. Knowledge of the individual’s internality and the nudge’s psychological cost is still needed, but as I argue in Section 3, this can be done more easily than fully recovering people’s preferences. This is a remarkable result, because it shows that even if people know their environment-independent preferences better than the social planner, this has no influence on whether the social planner is able to design the optimal nudge.

In heterogeneous populations, the optimal design of nudges must take into account the

---

6The reader should note that, in order to analyze nudges, we need the two characteristics of the framework already described: a discrepancy between decision and experienced utility (or between choice and welfare, more generally), and a dependence of choice on the environment. This is because if people already maximized social welfare, there would be no reason to change their choices; and if the environment did not matter, there could be no intervention that changed individual behavior without providing incentives, and therefore changing the choice set.

7Inspired by John Stuart Mill’s claim that “with respect to his own feelings and circumstances, the ordinary man or woman has means of knowledge immeasurably surpassing those [...] by any one else”.

3
costs and benefits for the different types of individuals. The concept of asymmetric paternalism endorses a nudge “if it creates large benefits for those who make errors, while imposing little or no harm on those who are fully rational” (Camerer et al., 2003). I show that when the nudge imposes psychological costs, the relative size of the biased population and the magnitude of their internality must be traded off with respect to the psychological costs, which will be imposed on the entire population. Therefore, in the presence of psychological costs, some nudges may not satisfy the criteria for asymmetric paternalism.

As a corollary to the criticism of nudges based on psychological costs, Glaeser (2006) argues that taxation should be preferred because a nudge is “an emotional tax on behavior which yields no government revenues”.

I show that this is only correct when taxation is fully efficient, in the sense that for each dollar taxed, the government collects no less than one dollar. However when taxation is not fully efficient (so for each dollar taxed the government collects less than one dollar), it is generally optimal to use nudges and taxes jointly.

The framework I developed also allows me to analyze what happens when firms change the environment to exacerbate consumers’ internalities in order to increase their profits. I follow Akerlof and Shiller (2015), who call such behavior “phishing”. A phish is a nudge that increases profits for the firm while reducing consumer welfare. Economists traditionally argue that a combination of competition and arbitrage would prevent phishing from happening in equilibrium, an argument that was criticized by Mullainathan and Thaler (2000). I show that competition does not necessarily reduce phishing, and can even result in a reduction in welfare, because with lower prices more consumers end up making mistakes.

While most analyses of nudging have been performed in partial equilibrium, a number

---

8 The argument is that taxing the behavior that the social planner wants to change (i.e. a “sugar tax” to avoid excessive consumption of sugar) is superior because it achieves the desired change in behavior while raising revenue.

9 Gabaix and Farhi (2015) also study taxation and nudging. The main difference between my framework and theirs, is that they are mainly interested in taxation, and have a particular case of nudge, which can be captured (under the assumption of quasilinearity of utility) by my framework. My main focus, on the other hand, is on nudges, and I consider only the simple case of linear taxation. Therefore, the two papers offer complementary approaches to public policy in the presence of internalities.
of authors have recently stressed the importance of analyzing nudges in general equilibrium, because firms’ reactions to the nudge can partially undo its effect (or even have a negative effect altogether). For example, Handel (2013) analyzes a nudge that reduces inertia in health insurance choice, and finds that this nudge backfires because it exacerbates adverse selection, resulting in a welfare loss. I characterize the optimal nudge in general equilibrium and show that, under perfect competition and some mild conditions (that the nudge actually works and that psychological and implementation costs are well-behaved), the optimal nudge is positive. In other words, under those conditions, there exists a nudge that improves social welfare. This result is important for two reasons. First, it shows that the results that nudges backfire in general equilibrium (such as Handel, 2013) might not be just due to the particular nudges used, but also because of market failures. Second, it highlights why having a general model is important: this result could not have been obtained in a partial equilibrium model, such as the ones that have been traditionally used in behavioral welfare economics. It is because of the tractability and generality of my framework that I can derive such result.

This paper is organized as follows. Section 2 develops the framework with experienced and decision utility, the dependence of utility on the environment, and the internality. Section 3 introduces nudges into the framework and characterizes the optimal nudge, in increasingly richer contexts. Section 4 introduces phishing, when firms profit from consumers’ mistakes, and shows that competition does not necessarily reduce phishing. Section 5 analyzes nudging and phishing in general equilibrium, and characterizes the optimal nudge, which will be positive under some mild conditions. Section 6 concludes. Any proofs not included in the main text are found in the Appendix.

2 Behavioral welfare economics

In a seminal paper, Paul Samuelson (1938) advocated using choice data to uncover individual preferences, in what is known as the revealed preferences paradigm. This paradigm

---

10 Spiegler (2014) and Grubb and Osborne (2015) also analyze nudges in a general equilibrium framework but do not derive general results about nudging and phishing.
assumes that the individual chooses the option (from a choice set) which maximizes her utility; therefore it is possible to find her preferences by “inverting” the process: if she chose $x$ rather than $y$ when both were available, then she must prefer $x$ to $y$. Recently, however, several economists have called this paradigm into question. They argue that if people make mistakes when making choices, then the revealed preference approach could be flawed.\textsuperscript{11} Indeed, the larger the mistakes people make in their choices, the less their choices represent their preferences.

I follow a framework from behavioral economics and psychology (Kahneman and Thaler, 2006), which considers that people have a belief about how much they will enjoy a good or an activity – this is called a hedonic forecast:\textsuperscript{12} a necessary condition for the revealed preferences paradigm to work is that people’s hedonic forecasts be accurate. However, a large number of studies show that people exhibit systematic biases in their hedonic forecast. For example, Kahneman and Snell (1992) show that people’s forecasts of how much they would enjoy consumption of various goods (such as yogurt or music) over the course of a week were basically uncorrelated with their actual enjoyment.\textsuperscript{13} As Kahneman and Thaler (2006) point out, the claim is not that people do not know what they like, but that on occasions (often, important ones) hedonic forecasts are very inaccurate. Even more, it seems that different brain circuits are involved in “wanting” versus “liking” (Berridge and Robinson, 2003). This distinction is especially relevant, because wanting is connected to the act of choosing, whereas liking is connected to our preference for things. This suggests that what we choose, and how much we actually enjoy what we choose, might be processed by overlapping, but distinct, neural circuits.

\textsuperscript{11}For example, Spiegler (2008) writes “non-standard decision models are often inconsistent with a narrow version of the revealed preference principle, according to which utility maximization and observed choice are synonymous”. Camerer et al. (2005) write “if we cannot infer what people like from what they want and choose, then an alternative method for measuring liking is needed, while avoiding an oppressive paternalism”. See also Campbell (2006).

\textsuperscript{12}Also known as affective forecast in the literature (Kahneman et al., 1997).

\textsuperscript{13}Similarly, Simonson (1990) and Loewenstein and Adler (1995) showed that people are bad forecasters of their own future preferences. For more on the early psychological literature on people’s biases forecasting their future tastes see Kahneman (1994).
Given the evidence against the revealed preferences paradigm, what should economists do? Campbell (2006) suggests “to abandon the framework of revealed preference and to consider the possibility that households may not express their preferences optimally”. That is the approach in this paper: I maintain the assumption that people are utility maximizers, but I relax the requirement that their hedonic forecast must be accurate. Moreover, I allow both decision and experienced utility to depend on the environment (such as the individual’s state of hunger).  

### 2.1 Experienced vs. decision utility

Inspired by this wealth of evidence, I explicitly consider that the individual makes a choice \( x \in X = \mathbb{R} \) in an environment \( e \in E = \mathbb{R} \). The environment affects the choice (in a way that will be described below) but cannot be altered by the individual, who takes it as given. I follow Kahneman et al. (1997) and Kahneman and Sugden (2005) in making a distinction between what people want and what they like: I define **experienced utility** \( u^{exp}(x|e) \) as the well-being or hedonic experience from outcome \( x \); this is how much the individual likes \( x \). For example, if she decided to buy a hamburger, the experienced utility would be the hedonic experience and related well-being from actually consuming the hamburger.  

On the other hand, I define **decision utility** \( u^{dec}(x|e) \) as the utility that the individual maximizes when she makes her choice – this is what the individual wants at the moment of making the choice. An interpretation of decision utility is that people have beliefs \( \beta \) about the rewards from each outcome \( x \), and that decision utility is simply the expectation of experienced utility, \( u^{dec}(x|e) = \mathbb{E}_{\beta}[u^{exp}(x|e)] \). Crucially, beliefs do not need to be correct.

---

14 There is a trend in Psychology and Economics, since the seminal work by Tversky and Kahneman (1981), where it is acknowledged that the environment matters when agents take decisions. For example, Bertrand et al. (2009) documented that showing the photo of an attractive female in the brochure for a loan increases demand by as much as a 200-basis point reduction in the interest rate.

15 Environment \( e \) should be understood as all the circumstances that might affect the choice of the individual, as well as the utility of the individual from making the actual choice, but not the utility from the outcome. For example, calorie information on restaurant menus will affect the choice of which meal to order, as well as the utility from the act of eating itself (it is possible that the consumer sees her hedonic experience from the meal diminished by the calorie information). However, the nudge does not affect the long-run utility from eating; the health consequences of over-eating that one time will be the same, with or without the nudge.

16 Kahneman et al. (1997) track the intellectual origins of this idea to Bentham and Edgeworth.

17 Formally, beliefs are probability distributions over states of the world \( s \in S \), and \( u^{dec}(x|e) = \mathbb{E}_{\beta}[u^{exp}(x|e,s)] \). Under this interpretation, projection bias means that those beliefs are biased in the direction of the current state.
and, as I indicated above, it has been documented that experienced and decision utility 
can differ and are even processed by different areas in the brain (Berridge and Robinson,
2003; Berridge and O’Doherty, 2014). Following the previous literature (Herrnstein et al.,
1993; Allcott et al., 2014), I define the internality $\Lambda(x|e)$ as the difference between 
the individual’s decision and experienced utility:

$$\Lambda(x|e) = u^{\text{dec}}(x|e) - u^{\text{exp}}(x|e).^{18}$$

The internality is important when analyzing individual decision-making, because it 
changes the relative incentives for different choices. For example, if the internality is a con-
stant, then the individual makes the same choices under her decision and her experienced 
utility, in which case there are no distortions. To measure the relative mistakes individuals 
make, we need to look at the marginal internality. I make the following normalizing 
assumption.

Assumption 1. The marginal internality satisfies:

$$\frac{\partial \Lambda(x|e)}{\partial x} \geq 0, \quad \frac{\partial^2 \Lambda(x|e)}{\partial x \partial e} \geq 0.$$ (1)

The first inequality in Assumption 1 states that individuals are over-consuming, such as 
in the case of over-eating at restaurants. This is because the internality $\Lambda(x|e)$ is increasing 
in $x$, and therefore higher choices of $x$ entail larger mistakes. When the marginal internality 
is 0 there is no distortion in choices; when that is the case we say the agent is rational.
The second inequality means that as the environment increases, the bias increases weakly,
i.e. higher values of $e$ represent environments where agents suffer from a larger internality, 
which is just a normalization. The marginal internality $\frac{\partial \Lambda(x|e)}{\partial x}$ encapsulates all the behavioral phenomena that would prevent an individual from fully pursuing her self-interest, some 
of which can be found in the earlier literature, such as projection bias,$^{19}$ misoptimization by

---

18 The concept of internality was introduced by Herrnstein et al. (1993): “a within-person externality, or internality, [occurs] when a person underweighs or ignores a consequence of his or her own behavior for him- or herself”. The concept is often used in Behavioral Welfare Economics, see for example Allcott and Taubinsky (2015) and Chetty (2015).

19 Projection bias is the (mistaken) belief that one’s preferences in the future will be similar to one’s current
consumers, projection bias, hedonic adaptation (which is the phenomenon by which people adapt to circumstances, good or bad, more than they had anticipated), and self-control problems. Going back to the interpretation of decision utility as the expected experienced utility, $u^{dec}(x|e) = E_\beta[u^{exp}(x|e)]$, it is tempting to argue that most people would eventually learn the correct beliefs and eliminate their biases. For example, List (2003) showed that individuals who gain experience in the market are less subject to the cognitive bias known as the endowment effect. However, there is also evidence against the idea that biases can be easily “unlearned”. First, people process information in a biased, and self-serving way, and are therefore not Bayesian learners (Lord et al., 1979). Moreover, some of these biases are evolutionarily ancient, and therefore unlikely to be the result of learning. Ultimately, internalities will be more relevant in those domains where there are limited opportunities for learning, or in those domains where despite frequent opportunities for feedback (such as in the case of eating), individuals do not seem to learn from their experience.

preferences. For example, people who do their grocery shopping when hungry, tend to buy disproportionately more food than when they just ate, even though their current state has no implications for their food needs for the following week (Nisbett and Kanouse, 1969; Gilbert et al., 2002). This is true even when stakes are higher: Busse et al. (2012) showed that people buy more convertibles when the weather is warmer or the skies are clearer than average, and a similar influence on hot weather on the purchase of houses with swimming pool. Although people understand the direction of the bias associated with their hedonic forecasts, they systematically underestimate its magnitude (Conlin et al., 2007).

When they do not fully understand the benefits or costs associated with purchasing a certain product; this has been studied previously in the literature: in the context of health (Baicker et al., 2015), energy efficient technology (Allcott et al., 2014; Allcott and Taubinsky, 2015), add-ons (Gabaix and Laibson, 2006) and overconfidence (Grubb and Osborne, 2015).

Several papers cited above show evidence of people’s inability to correct for their current mood and circumstances when taking decisions for their future. Busse et al. (2012) have a model where they explicitly consider the difference between experienced and decision utility.

Brickman et al. (1978) found that both lottery winners and victims of car accidents, reverted to the same level of happiness they had prior to their respective events after a few years. Hedonic adaptation is probably hard-wired in humans (Rayo and Becker, 2007), contributes to mistakes in hedonic forecast, and is one of the causes behind the projection bias (Loewenstein et al., 2003).

In models of self-control, the decision taken at a point in time might not reflect the individual’s well-being. Lack of self-control has been linked to a brain process with competing modules (McClure et al., 2004) – although see Kable (2013) for a critical review. In Jimenez-Gomez (2015) I analyze how a multi-modular self-control could have evolved in the human brain. When choices are the result of multiple mind modules, experienced and decision utility do not need to coincide.

The endowment effect is a bias that makes individuals value more things that they already own.

See also Ortoleva (2012) for a review of recent evidence from financial markets.

Capuchin monkeys, when properly trained to operate in markets, display reference dependence and loss aversion (Chen et al., 2006) and the endowment effect (Lakshminarayan et al., 2008). Because capuchin monkeys are relatively distant from humans (our last common ancestor is estimated to have lived around 40 million years ago), these biases are not culturally dependent, but rather universal in humans (Santos and Platt, 2013).
At this level of generality it is difficult to analyze nudges. Because of that, I add structure to the framework. Let procedural utility $\psi(x|e)$ be such that

$$\frac{\partial \psi(x|e)}{\partial e} = \frac{\partial u^{exp}(x|e)}{\partial e}, \text{ for all } x \in X, e \in E. \quad (2)$$

Note that, in principle, there are several functions that fulfill Equation 2. However, by definition, it must be that

$$u^{exp}(x|e) = \omega(x) + \psi(x|e) + K, \quad (3)$$

where $\omega(x)$ is outcome utility, and we can make $K = 0$ without loss of generality.\(^{27}\)

Note that outcome utility, $\omega(x)$, is the standard utility function considered traditionally in Economics. On the other hand, procedural utility $\psi(x|e)$ represents a “behavioral component” of utility, which depends on the environment: for example on the state of hunger of the individual or whether there are alluring pictures of food in the restaurant.

**Assumption 2.** I assume that $\omega(x)$ is concave, and that $\psi(x|e)$ is concave in $x$ and such that:

$$\frac{\partial \psi(x|e)}{\partial x} \geq 0, \quad \frac{\partial \psi(x|e)}{\partial e} \geq 0, \quad \frac{\partial^2 \psi(x|e)}{\partial x \partial e} \geq 0. \quad (4)$$

The interpretation of Assumption 2 is that procedural utility is increasing in both $x$ and $e$, and there are complementarities between consumption and the environment. This is consistent with my assumption that higher $e$ means an environment with more temptation and more immediate rewards.

**Example 1 (Calorie labels in restaurant menus).** The consumer chooses “excess” calories (i.e. calories beyond what is needed for survival) $x \in [0, +\infty)$. The consumer derives hedonic pleasure from the process of eating those calories $\psi(x|e) = e \cdot h(x)$, where $h(x)$ is a hedonic function. However, she must pay a monetary as well as a future health cost, $\omega(x) - px = -\frac{x^2}{2} - px$. The individual suffers from present bias, and therefore she does not fully internalize the future health costs, $\Lambda(x|e) = e \cdot \frac{x^2}{2}$. That means that her experienced

\(^{27}\)The concepts of outcome utility and procedural utility were introduced by Benz (2005, 2007); Frey et al. (2004).
utility is \( u^{exp}(x|e) = -\frac{x^2}{2} - px + e \cdot h(x) \), and her decision utility is \( u^{dec}(x|e) = -(1 - e) \frac{x^2}{2} - px + e \cdot h(x) \). Maximizing the decision utility, we find that the individual chooses \( x \) such that

\[
e \cdot h'(x) = p + (1 - e) \cdot x
\]

Consider two environments. In the status quo \( e_0 \), there are no calorie labels on the menu. On the other hand, in environment \( e_{CL} \leq e_0 \) calorie labels are included on the menu. Note from Equation 5 that when \( e \) decreases the consumer chooses a lower \( x \), which is good for her because she was overconsuming due to the internality. On the other hand, a lower \( e \) also reduces procedural utility (knowing that a meal has a lot of calories reduces its hedonic value). Therefore, there is a tradeoff in introducing calorie labels: they correct the internality, but they have a cost in terms of procedural utility. As we will see in Section 3 below, this is a general intuition that will determine how to nudge optimally.

3 Nudging

In the book “Nudge”, Thaler and Sunstein (2008) defined a nudge as an intervention “that alters people’s behavior in a predictable way without forbidding any options or significantly changing their economic incentives.” Since they published this influential work, nudges are used in a wide range of areas: nutrition (List and Samek, 2015), retirement savings (Bernheim et al., 2011), and energy efficiency (Allcott and Taubinsky, 2015), among others. The study of nudges has been spearheaded by policymakers and practitioners, with formal economic analysis lagging behind, partly because a neoclassical setup cannot be used to analyze nudges, because in such setup agents typically make no mistakes and, absent externalities, nudges are unnecessary. However, when people make mistakes (and as I have documented above in Section 2, this happens in several important domains), nudging can increase welfare. Equipped with the concepts of experienced and decision utility, we can make progress in analyzing when nudges will increase welfare, and when they will not.
3.1 A Theory of Nudges

I assume that the individual makes a single choice in a static context. The universe of potential choices is given by $X$, and the universe of environments by $E$. The individual cannot affect the environment, and therefore she takes it as given. There is a status quo environment $e_0$: this is the environment before the introduction of any nudge, for example menus with no calorie information.

**Definition 1.** A nudge $\nu \in \mathbb{R}_+$ is a variable that does not affect the choice set $X$, but changes the environment $e = e_0 - \nu$.

Note that the marginal internality is decreasing in $\nu$. This is arguably the case with overeating, where individuals overestimate the degree of pleasure they will derive from certain types of food consumption (for example due to projection bias, when individuals mistakenly purchase too much food due to their current state of hunger, Read and Van Leeuwen, 1998). In that context, $\nu$ can be understood as the intensity of anti-obesity campaigns, such as the ubiquity and salience of calorie and nutritional information. Moreover, because of Assumption 2 procedural utility is also decreasing in $\nu$, $\frac{\partial \psi(x|e)}{\partial e} \frac{\partial e}{\partial \nu} \leq 0$, capturing the psychological costs from the nudge.

Let $-\eta(x)$ be a negative externality generated by the individual, for example costs associated with diabetes or obesity which are borne by the social security system. I assume that, when making welfare analysis, the social planner maximizes experienced utility net of externalities. I consider that there is a representative agent (I drop this assumption in Section 3.3), and define welfare as

$$W = \omega(x) + \psi(x|e) - \eta(x).$$  \hspace{1cm} (6)
In principle, the fact that the social planner and the individual maximize different functions is without loss of generality, although Camerer et al. (2005) and Kahneman and Sugden (2005) make the stronger claim that welfare should be based on liking (which they identify with \( u^{exp} \), not on wanting, identified with \( u^{dec} \)). While I agree with this point of view, the skeptical reader can simply consider \( W = u^{exp} - \eta \) to represent any objective function for the social planner.\(^{33}\) Moreover, I am not including explicitly producer surplus in Equation 6, because for now I am implicitly assuming that producer surplus is zero (for example, because there is perfect competition), and therefore that welfare is simply the sum of experienced utility and externalities. In Section 5 I will consider producer surplus explicitly.

3.2 Optimal nudge: no knowledge of outcome preferences needed

We say that nudge \( \nu \) is optimal if it maximizes welfare. The main observation in order to design the optimal nudge, is that we can obtain information from the individual’s choice, as long as we know the internality and externality, because:

\[
\frac{\partial W}{\partial x} - \frac{\partial u^{dec}(x|e)}{\partial x} = - \frac{\partial}{\partial x} [\Lambda(x|e) + \eta(x)]
\] (7)

That is, the right-hand side of Equation 7 measures the gap between marginal welfare and the marginal incentives the individual faces. Proposition 1 below shows that the benefit from the nudge can be measured in terms of that gap.

**Proposition 1.** Nudge \( \nu \) is optimal if

\[
- \frac{\partial x}{\partial \nu} \cdot \frac{\partial [\Lambda(x|e) + \eta(x)]}{\partial x} = \frac{\partial \psi(x|e)}{\partial e},
\] (8)

where \( x = x(\nu, p) \) and \( e = e_0 - \nu \).

The left hand-side of Equation 8 expresses the benefit from the nudge, which is given in terms of the marginal decrease in the gap between welfare and decision utility which, as we

---

\(^{33}\)If the social planner maximizes one function (which is \( W \)), and the individual maximizes another function (which is \( u^{dec} \)), then this model is general enough to capture the standard neoclassical model (in which case \( u^{exp}(x|e) = u^{dec}(x|e) \)), as well as any other situation where social and individual welfare are not aligned. The mathematical results still hold, it is only their interpretation that changes.
saw in Equation 7, can be expressed in terms of the marginal internality and externality. The right-hand side of Equation 8 expresses the psychological cost of the nudge, which is a decrease in procedural utility (such as a decrease in the hedonic experience of eating a meal once we know its calorie content).\(^{34}\) The key intuition to understand Proposition 1 comes from the fact that given \(\nu\), at the optimal \(x\) the marginal increase in experienced utility must be equal to the marginal decrease in the internality:

\[
\frac{\partial u^{exp}(x|e)}{\partial x} = -\frac{\partial \Lambda(x|e)}{\partial x},
\]

(9) because at the optimal \(x\) we have that \(\frac{\partial u^{dec}(x|e)}{\partial x} = 0\). This has no parallel in standard Economics, and is therefore one of the innovations of Behavioral Welfare Economics.\(^{35}\) Note also that internalities and externalities are given the same weight in Proposition 1, the main difference being that internalities are relevant for choice but not for welfare, whereas externalities have the reverse property of being relevant for welfare but not for choice. Even absent internalities, nudges are an important tool to correct externalities, in addition to taxation (which I analyze in Section 3.4).

Proposition 1 endorses Sunstein’s view that the epistemic argument (which posits that the individual knows more about her own preferences than the social planner does) is not a valid criticism against nudges: the social planner does not need to know outcome preferences in order to design the optimal nudge. This has great importance in the design of nudges, because recovering preferences from choice data when agents are subject to internalities is notoriously difficult (Bernheim and Rangel, 2009).\(^{36}\) Indeed, I argue that the social planner can identify the optimal nudge without recovering people’s full preference

---

\(^{34}\)The proof in the Appendix is written for general experienced and decision utility functions (not necessarily additive), showing that this result does not depend on any particular functional form: in the general case, the social planner needs to know the form of \(u^{exp}\) and \(u^{dec}\), but does not need to know \(\omega(x)\).

\(^{35}\)In standard Economics experienced and utility decision utility coincide, and therefore in most papers the marginal internality is zero by assumption, meaning that nudges play no role in such models: since the left hand side of Equation 8 is 0, the optimal solution is at the corner \(\nu = 0\) as long as there are psychological costs.

\(^{36}\)See also Benkert and Netzer (2015) and Goldin and Reck (2015). Of course, Proposition 1 can also be interpreted in light of this literature: for example, Benkert and Netzer (2015) show that a nudge is optimal if it induces less mistakes than any other environment, which is a property of the optimal nudge \(\nu^*\) in the context of this paper too.
profile (a point made by Chetty, 2015). First, note that $\frac{\partial x}{\partial \nu}$ can be inferred from observing reactions of demand to similar nudges from historical data.\(^{37}\) Moreover, the externality $\eta(x)$ can be obtained by traditional methods such as contingent valuation (Cherry et al., 2001) or some more novel methods (Chetty, 2015). The main difficulty in finding the optimal nudge lies in identifying the internality $\Lambda(x|e)$ and procedural utility $\psi(x|e)$. One of the main ways in which economists are recovering such non-standard functions is through structural estimation coupled with field experiments (Allcott and Taubinsky, 2015). Moreover, some economists (Bernheim and Rangel, 2009) argue that neuroscience can be helpful in uncovering people’s preferences beyond choice data, because most biases are consistent and predictable (Kahneman, 2011).

### 3.2.1 Identifying $\frac{\partial x}{\partial \nu}$ for new nudges through price elasticity

Equation 8 poses an identification problem for new nudges, for which the marginal demand with respect to the nudge $\frac{\partial x}{\partial \nu}$ is unknown because of lack of historical data. However, we can get around this challenge by using the change in demand with prices. Let’s assume that there is a price $p$ per consumption unit, so that an individual that consumes $x$ must pay $px$, so that outcome utility is $\omega(x) - px$.

**Corollary 2.** The optimal nudge is given by

$$-\frac{\partial^2[\Lambda(x|e) + \psi(x|e)]}{\partial x \partial e} \cdot \frac{\partial x}{\partial p} \cdot \frac{\partial[\Lambda(x|e) + \eta(x)]}{\partial x} = \frac{\partial \psi(x|e)}{\partial e},$$

where $x = x(\nu, p)$ and $e = e_0 - \nu$.

Equation 10 continues to reflect the cost-benefit of the nudge, but can potentially be estimated even for a nudge that has never been implemented; as long as the social planner knows $\Lambda$ and $\psi$ (something which I have discussed above in Section 3.2), and has historical data about the response of demand to price.

\(^{37}\)See Corollary 2 below for an alternative method to compute $\frac{\partial x}{\partial \nu}$ using price elasticity.
3.3 Asymmetric paternalism: nudging a heterogenous society

So far we have considered nudging a representative individual, however the target population of most interventions is heterogeneous, and this heterogeneity needs to be taken into account when designing an optimal nudge (Conlin et al., 2007; Agarwal and Ambrose, 2008). I consider a population of agents who differ in their preferences, as summarized by parameter \( \theta \in [0, 1] \), so that decision utility is given by

\[
\begin{align*}
    u^{\text{dec}}(x|e, \theta) &= \omega(x|\theta) - px + \Lambda(x|e, \theta) + \psi(x|e, \theta). \quad (11)
\end{align*}
\]

Let \( \theta \) be distributed in the population according to distribution \( F \). I define a social welfare function, as a function which aggregates the utilities from society into a single social welfare measurement (Mas-Colell et al., 1995). We say that nudge \( \nu \) is socially optimal if it maximizes social welfare. Following the standard tradition in Economics, I consider the “utilitarian” social welfare function \( W(\nu) \) that results from the (potentially weighted) sum of individual utility:

\[
W(\nu) = \mathbb{E}_\theta [\omega(x|\theta) - px + \psi(x|e, \theta) - \eta(x)]
\]

The social welfare \( W(\nu) \) is a function of the nudge \( \nu \), and the socially optimal nudge is the one that maximizes it.

**Proposition 3.** The socially optimal nudge \( \nu \) is given by

\[
\mathbb{E}_\theta \left[ -\frac{\partial x}{\partial \nu} \left( \frac{\partial (\Lambda(x|e, \theta) + \eta(x))}{\partial x} \right) \right] = \mathbb{E}_\theta \left[ \frac{\partial \psi(x|e, \theta)}{\partial e} \right],
\]

where \( x = x(\nu, p, \theta) \), and \( e = e_0 - \nu \).

Proposition 3 expands Proposition 1 to a heterogeneous population, and preserves the result that the social planner does not need to have access to outcome preferences, only

---

38Although I assume quasilinearity of outcome utility on wealth for ease of exposition, the results in this paper (with the notable exception of Corollary 2) do not depend on this assumption.

39Formally, a social welfare function \( S \) is given by \( S : \mathbb{R}^{[0,1]} \rightarrow \mathbb{R} \).

40Even general social welfare functions (as in Saez and Stantcheva, 2016) are utilitarian although with heterogeneous weights. However the results of this paper can be equally applied to any other social welfare function, like maximin (Rawls, 1974), by changing the relevant calculations.
the internality and psychological cost. However, note that the social planner needs to have access to individual demand, and not just aggregate demand, in order to be able to compute the left-hand side of equation 11. The intuition we found in Section 3.2 translates to social welfare: the benefit from the nudge is measured by the decrease of the marginal internality and the per-unit decrease in nudging costs, and must equal the average marginal psychological cost of increasing the intensity of the nudge. Proposition 3 also hints to a new dimension that needs to be balanced in finding the optimal nudge, and that is taking into account the different types of individuals who benefit or are hurt by the nudge.

Camerer et al. (2003) advanced the concept of asymmetric paternalism, describing the case when “nudges are effective because they improve the decision of biased agents without hurting rational agents”. Here I will show that nudges are not asymmetrically paternalistic if they have large psychological costs that are borne by individuals who do not benefit from the nudge. In Section 2.1 I defined rational agents as those whose marginal internality is zero, i.e. those with \( \theta \) such that \( \frac{\partial \Lambda(x|e,\theta)}{\partial x} = 0 \) for all \( x, e \). Note that in the left-hand side of Equation 11 in Proposition 3, the change in demand is weighed by the marginal internality; therefore the weight given to rational agents in the left-hand side is zero because they have zero marginal internality. Suppose that there are two types of individuals, the rational \( \theta_R \) and the biased \( \theta_B \), where the latter have \( \frac{\partial \Lambda(x|e,\theta_B)}{\partial x} > 0 \). The respective weights of these individuals in society are \( \mu_R \) and \( \mu_B \). For simplicity assume that there are no externalities, so \( \eta(x) = 0 \) for all \( x \). Let \( x^i = x(\nu, p, \theta_i) \), Equation 11 becomes

\[
-\mu_B \cdot \frac{\partial x^B}{\partial \nu} \cdot \frac{\partial \Lambda(x^B|e,\theta_B)}{\partial x} = \mu_B \cdot \frac{\partial \psi(x^B|e,\theta_B)}{\partial e} + \mu_R \cdot \frac{\partial \psi(x^R|e,\theta_R)}{\partial e}.
\] (12)

Equation 12 indicates that the arguments of both proponents and detractors of nudges are correct about an aspect of the nudging tradeoff. Proponents of asymmetric paternalism are correct that the nudge should be designed as to influence biased agents, not rational ones, as showed in the left-hand side. Detractors of nudges, however, are right that all agents bear a similar point was made by Allcott and Taubinsky (2015) in the context of using sufficient statistics in the presence of behavioral biases.
the psychological cost of the nudge, that is in the right-hand side of the equation. Even though biased agents might bear more of this cost, rational agents should be taken into account, and this is especially important if the social planner believes that the fraction of biased agents is small. Indeed, as $\mu_B$ becomes small, the marginal internality has to be large to justify a nudge in the presence of psychological costs.\footnote{Note however that if rational agents do not have a psychological cost, $\frac{\partial \psi(x_R | e, \theta_R)}{\partial e} = 0$, as might be the case with certain nudges such as cigarette warning signs, then the nudge will be asymmetric paternalistic.}

**Example 2.** Individuals decide whether to get a dessert or not, $x \in \{0, 1\}$. There are two types of individuals: rational, who chose to consume if $\omega \geq p$; and biased, who consume if $\omega + \Lambda \geq p$ (where I have abused notation by writing $\omega$ and $\Lambda$ as constants). All individuals have experienced utility $x \cdot (\omega - p)$. There is fraction $\mu_R$ of rational, and fraction $\mu_B$ of biased individuals, with respective valuations $\omega_R$ and $\omega_B$, with $\omega_R > p$ and $\omega_B + \Lambda > p > \omega_B$. That is, both types of individuals consume the dessert, but it is only socially optimal for rational agents to do so. There is a nudge that does not affect the internality, but simply creates a psychological cost for all: $-c(1)$ is the psychological cost incurred by those who consume the dessert, and $-c(0)$ by those who don’t consume. Therefore, with the nudge rational agents consume if $\omega - c(1) > p - c(0)$, and biased individuals if $\omega + \Lambda - c(1) > p - c(0)$. Moreover, assume that the nudge is such that only rational individuals consume the dessert, i.e. $\omega_R - c(1) > p - c(0)$ but $\omega_B + \Lambda - c(1) < p - c(0)$. Assume that there is perfect competition, so all surplus is captured by consumers. The change in welfare from introducing the nudge is

$$\Delta W = \mu_R(\omega_R - p - c(1)) - \mu_B \cdot c(0) - [\mu_R(\omega_R - p) + \mu_B(\omega_B - p)] =$$

$$= \mu_B(p - \omega_B - c(0)) - \mu_R \cdot c(1).$$

Because $p > \omega_B$, the nudge increases the welfare of biased individuals by preventing them from consuming, whenever $c(0) \leq p - \omega_B$. On the other hand, the nudge imposes a cost $c(1)$ on rational individuals, who continue to consume dessert in any scenario. Therefore, the nudge will increase welfare as long as the social welfare gain from biased individuals, measured by $\mu_B(p - \omega_B)$ is larger than the social cost of the nudge, measured by $\mu_B \cdot c(0) +$
3.4 Nudging optimally when taxes are available

A recent literature analyzes optimal taxation when agents make mistakes, (Gabaix and Farhi, 2015; Lockwood and Taubinsky, 2015; Gerritsen, 2015). There is evidence that taxing cigarettes can increase the welfare of smokers (Gruber and Mullainathan, 2006), arguably by preventing them from making choices that do not maximize their experienced utility. In this section I analyze the optimal level of nudging and taxation when both policies are available, an important question that (with the exception of Gabaix and Farhi, 2015) has received little attention in the literature.  

Suppose that the government can set a linear tax $\tau$, in addition to the nudge $\nu$. However, I allow for the possibility that for each dollar taxed, the government collects less than one dollar. In particular, I assume that that revenue collected from taxes is measured by $\delta(\cdot)$, so that for each $y$ dollars that the individual pays in taxes, the government collects only $\delta(y)$, with $\delta(0) = 0$, $\delta(y) \leq y$, and $\delta(y)$ increasing and concave. Intuitively, $\delta'(y)$ measures the efficiency of the government in collecting revenue: when $\delta'(y) = 1$, the government is fully efficient (and when $\delta'(y) = 0$ the government is fully inefficient). Individual $\theta$ solves
\[
\max_x \omega(x|\theta) + \Lambda(x|e, \theta) + \psi(x|e, \theta) - (p + \tau)x.
\]

For simplicity I assume that the government redistributes its revenues from taxation, and that there are no implementation costs (I consider them in Section 5). The social planner maximizes social welfare $W(\nu, \tau)$:
\[
W(\nu, \tau) = \mathbb{E}_\theta [\omega(x) + \psi(x|e) - \eta(x) - (p + \tau)x + \delta(\tau \cdot x)].
\]

Gabaix and Farhi (2015) are interested in the question of optimal taxation in the presence of “behavioral agents” (and consider the limit case when taxes are close to zero). While their framework is more general than mine when it comes to taxation, their applications to nudges assume a particular form of internality (where the individual mistakenly believes the price of a good to be higher or lower than it really is, but the budget constraint must be satisfied). With quasilinearity (i.e. for purchases that are small, which are the ones I am considering in this paper), their assumption is analogous to having $\Lambda(x|e) = A_\Lambda \cdot e \cdot x$, for some constant $A_\Lambda$, making my framework more general when it comes to analyzing nudges. Therefore, both papers offer complementary approaches to analyzing taxation and nudges when individuals suffer from internalities.
Proposition 4. The optimal nudge and tax are given by

\[ E_\theta \left[ -\frac{\partial x}{\partial \nu} \cdot \frac{\partial \left[ \Lambda(x|e, \theta) + \eta(x) \right]}{\partial x} \right] = \partial E_\theta [\psi(x|\nu, \theta)] \cdot \tau \cdot E_\theta \left[ \frac{\partial x}{\partial \nu} \cdot \delta'(\tau \cdot x) \right], \]  
(13)

\[ E_\theta \left[ -\frac{\partial x}{\partial \tau} \cdot \frac{\partial \left[ \Lambda(x|e, \theta) + \eta(x) \right]}{\partial x} \right] = E_\theta \left[ x - \left( x + \frac{\partial x}{\partial \tau} \right) \cdot \delta'(\tau \cdot x) \right], \]  
(14)

where \( x = x(\nu, p, \theta, \tau) \).

On the left hand side of Equation 13, we have the usual benefit from the nudge. On the right-hand side, the first term is the psychological cost, and the second term is reduction in revenue from the decrease in demand induced by the nudge. With respect to Equation 14, the left hand side measures the benefit from taxation in terms of reducing the gap between welfare and decision utility. On the right hand side we see that taxation has two effects: one is to increase revenue directly, and the other is to reduce demand, and indirectly reduce revenue. Allcott and Taubinsky (2015) analyze a similar situation, in which taxation is fully efficient, \( \delta'(y) = 1 \) for all \( y \). Consider in addition that population is homogeneous, so there is a representative individual. In that case the optimal tax can be characterized as follows.

Corollary 5. When taxation is fully efficient and the population is homogeneous, the optimal tax \( \tau^* \) is given by:

\[ \tau^* = \frac{\partial \left[ \Lambda(x|e) + \eta(x) \right]}{\partial x}, \]  
(15)

where \( x = x(\nu, p, \tau) \).

When taxation is fully efficient and the population is homogeneous, the optimal policy is to set the tax equal to the sum of marginal internality plus externality. Note that this is a generalized Pigouvian tax, which includes the internality in addition to the externality. This solves the issue of the optimal tax, but what is the optimal nudge when taxation is available? Glaeser (2006) argued that “soft paternalism is an emotional tax on behavior which yields no government revenues”, hinting that old-fashioned taxation might be superior.

---

44 Allcott and Taubinsky (2015) analyze a similar setup, in which buyers have a bias that makes them purchase too few efficient lightbulbs, and they study the optimal subsidy. In their model each consumer purchases at most one lightbulb, and they obtain a formula equivalent to Equation 14, and reach the same conclusion about the optimal subsidy being equal to the marginal bias.

45 This observation has been anticipated in similar contexts by Allcott et al. (2014), Chetty (2015) and Baicker et al. (2015).
to nudging. It turns out that in this particular context, Glaeser’s argument is correct.

**Corollary 6.** In a homogenous population, if taxation is fully efficient, the optimal nudge is \( \nu = 0 \) (i.e. no nudge), as long as it has a psychological cost.

**Proof.** With a representative individual and when taxation is fully efficient, the calculation for the optimal nudge from Equation 13 simplifies to:

\[
\frac{\partial W}{\partial \nu} = - \frac{\partial x}{\partial \nu} \left( \frac{\partial (\Lambda + \eta)}{\partial x} - \tau \right) - \frac{\partial \psi}{\partial e} = - \frac{\partial \psi}{\partial e}.
\]

Therefore, if there is any psychological cost \(- \frac{\partial \psi}{\partial e} < 0\), the optimal nudge is \( \nu = 0 \).

In general, Glaeser’s argument is correct only in the special case when taxation is fully efficient and the population homogeneous: the optimal nudge is the one that maximizes procedural utility, and consequently if the nudge imposes psychological costs, the optimal amount of nudging is zero. However, taxation is usually not fully efficient; for example Fullerton and Metcalf (1997) estimated that for each dollar taxed, the government collects only 75 cents per dollar. Because of that, in general it is optimal to nudge, even when taxation is available, as long as the costs from nudging are not too high. To show this formally, I assume that \( \delta'(y) = \alpha \cdot y \), i.e. for each dollar the government taxes, \( \alpha \) dollars are collected.

**Proposition 7.** In a homogeneous population, if \( \alpha < 1 \) and the psychological costs are small for a small nudge, \( \frac{\partial \psi}{\partial e} \bigg|_{\nu=0} = 0 \), then the optimal nudge is positive, \( \nu > 0 \).

To summarize, when taxation is efficient, the social planner has fewer reasons to nudge; when taxation is not efficient, the social planner should use nudges which are not too costly.

In other words, whenever taxation is inefficient, then it is costly to use it as a policy instrument, and the optimal level of nudges and taxes will depend on their relative costs and benefits.

---

46 Note that \( \alpha \) corresponds to the marginal utility of public funds in Gabaix and Farhi (2015).

47 Note that in the limit when taxation is fully inefficient, i.e. \( \delta'(x) = 0 \) for all \( x \), we have that the optimal nudge is identical to the one we found in Proposition 3, which is always positive as long as \( \frac{\partial \psi}{\partial e} \) is small at \( \nu = 0 \), i.e. as long as the psychological costs do not increase dramatically when a small nudge is introduced.

48 This result is related to Proposition 3.3 in Gabaix and Farhi (2015). Footnote 43 summarizes the differences.
3.5 Choosing between nudges

So far I have assumed that there is only one type of nudge available. However, in some applications the social planner might be choosing between two distinct nudges: for example, nudge $\nu_1$ which includes calorie labels in restaurant menus versus nudge $\nu_2$ which increases the salience of healthy options. To analyze this problem, in this Section I assume that the environment $e$ is multidimensional, $e = (e^A, e^\psi)$, where one dimension of the environment affects internalities, and the other dimension affects procedural utility. However, the choice is still unidimensional, $x \in X = \mathbb{R}_+$. The individual maximizes her decision utility $u^{\text{dec}}(x|e) = \omega(x) - px + \psi(x|e^\psi) + \Lambda(x|e^\Lambda)$. Importantly, each nudge can affect the two different dimensions of the environment, so $e_i = (e^\Lambda(\nu_i), e^\psi(\nu_i))$. For simplicity, I assume that there is a representative individual (it is straightforward, but notationally intense, to extend the result to a heterogeneous population). Let $x^i$ be the choice of the individual under nudge $\nu_i$; the welfare function is defined only in terms of the dimension of the environment that affects procedural utility:

$$W(\nu_i) = \omega(x^i) - px^i + \psi(x^i|e^\psi_i) - \eta(x^i).$$

We say that nudge $\nu_1$ is better than nudge $\nu_2$ if $W(\nu_1) \geq W(\nu_2)$.

**Proposition 8.** A sufficient condition for nudge $\nu_1$ to be better than $\nu_2$ is

$$\psi(x^2|e^\psi_2) - \psi(x^2|e^\psi_1) \geq \Lambda(x^1|e^\Lambda_1) - \Lambda(x^2|e^\Lambda_2) + \eta(x^1) - \eta(x^2).$$

(16)

A necessary condition for nudge $\nu_1$ to be better than $\nu_2$ is

$$\psi(x^1|e^\psi_1) - \psi(x^1|e^\psi_2) \leq \Lambda(x^2|e^\Lambda_2) - \Lambda(x^1|e^\Lambda_1) + \eta(x^2) - \eta(x^1).$$

(17)

Proposition 8 provides sufficient and necessary conditions for nudge $\nu_1$ to be better than nudge $\nu_2$, which are very similar to the condition in Proposition 1: nudge $\nu_1$ will be better if the psychological costs (measured by changing $e^\psi$ while keeping the choice of the individual constant) are smaller than sum of the changes in internalities and externalities.
Note that nudge \( \nu_1 \) and \( \nu_2 \) could affect the environment very differently: for example, \( \nu_1 \) could affect \( e^\psi \) but not \( e^\Lambda \) if providing calorie information resulted in a psychological cost but not in a reduction of temptation, whereas nudge \( \nu_2 \) could affect \( e^\Lambda \) but not \( e^\psi \) if making healthy options salient reduced temptation without psychological costs. When we make the comparisons with respect to environment \( e_1 \) we obtain the sufficient conditions, and when those comparisons are made with respect to \( e_2 \), we obtain the necessary conditions. Proposition 8 is useful for precisely the same reasons I discussed with respect to Proposition 1: the social planner does not need to know outcome preferences \( \omega(x) \) in order to determine whether these conditions hold.

### 3.6 Limitations of the approach

The literature on “behavioral revealed preferences”, initiated independently by Bernheim and Rangel (2009) and Salant and Rubinstein (2008), analyzes what we can infer about people’s preferences from their choice when they make mistakes, which is the approach I have followed. There are two main threats to this approach, which I analyze next.

#### 3.6.1 When consumers do not optimize

A direct threat to the approach happens when people do not maximize their decision utility. This could happen when people are not aware of a subset of choices \( Y \subset X \): we, as modelers, believe that decision utility from \( y \in Y \) is \( u^{\text{dec}}(y|e) \), but truly it is \(-\infty\), since it is not in the effective choice set of the individual. This is especially relevant when designing nudges that might change the attention or awareness of the individual, such as defaults (in cases when they might be ignored) or other inconspicuous “choice architecture designs”. Consider the case of organ donation, where there is a default: if consumers are not paying attention to the option of opting in or out of being a donor, then they are not maximizing a utility function at all.\(^{49}\) Another potential reason why people might not optimize is a depletion of

\(^{49}\)Formally, suppose that individuals only pay attention to the choice with probability \( p \) (and when they pay attention they make the choice that they want), and with probability \( 1-p \) they do not pay attention to the choice, and they simply stick to the default. Note that this case cannot be modeled as an experienced vs. decision utility paradigm, because some people are not maximizing their utility at all (decision or otherwise), since they are not aware of having made a choice! Suppose that a fraction \( q \) of the population would like to be an organ
cognitive resources (like choice fatigue). If people are subject to too many choices, or are simply physically or mentally exhausted, they might use some “shortcut heuristic” in order to make a choice, that differs from what we believe is their decision utility.

### 3.6.2 When consumers’ preferences depend on more than \( x \) and \( e \)

The other main threat to the behavioral revealed preference approach arises when people value things that we have not included in the model. In the context of organ donation, if people want to conform to a social norm (which is not included in the model), it is possible to have an equilibrium where everybody becomes an organ donor to conform to the norm, even if everybody would prefer not to be a donor.\(^{50}\) When there is a default, while both the “all donor” and the “no donor” equilibria are possible, the default can serve as a focal point, where people form expectations from the default, therefore becoming a self-fulfilling prophecy. In this case, the social planner can generate the outcome that she wants simply by choosing the default, and *de facto* the social norm, but this is not informative about people’s preferences. In those cases, revealed preference cannot be used to justify the nudge.

### 4 Phishing

Akerlof and Shiller (2015) define *phishing* as “getting people to do things that are in the interest of the [person responsible for phishing], but not in the interest of the target” – in this spirit, phishing is very similar to conning. While the term “phishing” was coined only recently, the phenomenon has long been documented and studied by economists. Ausubel (1991) analyzed the credit card market, with thousands of firms and lacking regulatory donor. When being a donor is the default, then we will observe that a fraction of \( p \cdot q + (1 - p) \) are donors, and fraction \( p \cdot (1 - q) \) are not donors. If we knew \( p \) we could in theory find \( q \). However for most applications we do not know \( p \): we could try to find out \( p \) by asking people whether they were aware of the choice after the fact, although this method has its own problems.

\(^{50}\)Formally, and continuing the discussion from Footnote, 49 consider that people are consciously optimizing with probability \( p = 1 \); however now we are not going to assume that \( q \) is exogenous. People are known to conform to social norms, and therefore their choice should not be taken in isolation: consider the case where individuals have a preference \( \omega - \mu \cdot n \) to be an organ donor, where \( \omega > 0 \) is homogeneous in the population, and \( n \) is the fraction who choose not to become donors, and let’s normalize utility of not being a donor to 0. This utility function corresponds to the case when people have an altruistic desire to be organ donors, but they do not want to stand out from a norm of not donating. If \( \mu > \omega \) there are two pure Nash equilibria. In the “donor” equilibrium, everybody donates, because \( \omega > 0 \); in the “not donor” equilibrium nobody donates, because \( \omega < \mu \).
barriers (and so prima facie fulfilling the canonical conditions for perfect competition); he finds however that credit card interest rates are higher than what would be expected in a competitive market, and attributes this to consumers’ lack of sophistication.\footnote{In particular, to the failure to correctly predict the probability with which they would end up paying interest on their outstanding balances.} Financial markets provide several examples of phishing, supporting Mullainathan and Thaler (2000), who criticize the claim that irrationality could not exist in those markets, since arbitrageurs would profit from irrational agents, driving them out of the market. Carlin (2009) provides evidence that identical financial products can have significantly different prices, and this is partly driven by firms increasing the complexity of their products, making it more difficult for consumers to understand the actual price. Complexity increases with competition, because there is more profit to be made from phishing consumers than from debiasing them.\footnote{This observation was already made by Akerlof and Shiller (2015) and Thaler (2015).} Egan (2015) finds that in a market for bonds, consumers often purchase dominated bonds (i.e. with a lower yield, net of price, than another bond in any state of the world); moreover they purchase more of the more expensive bonds. He attributes this to brokers’ (misaligned) incentives, and consumer lack of sophistication. Mullainathan et al. (2012a) provide experimental evidence of broker phishing behavior: they randomly assigned auditors (with different portfolios) to brokers. They found that brokers fail to debias their clients and, on the contrary, reinforce biases that are in their own interests.\footnote{Biasing consumers towards strategies that satisfy them in the short run (in detriment of long-run returns) could be exacerbated if there is herding (Banerjee, 1992).}

4.1 A Theory of Phishing

Consider a firm that sells a good at price $p$, and has a technology with constant returns to scale for production of the consumption good with marginal cost of production equal to 0. Unlike in standard models in Economics, this firm can alter the environment to increase its profits.

**Definition 2.** A phishing is a variable chosen by the firm, which does not affect the choice set $X$, but alters the environment $e = e_0 + \phi$.

Note that $\phi$ is analogous to the nudge $\nu$, except that it changes the environment in the
opposite direction. The firm can choose $\phi$, at a per-unit cost $\kappa(\phi)$: a firm that produces $y$ units of the good, has cost $y \cdot \kappa(\phi)$, with $\kappa(0) = 0$ and $\kappa$ increasing and convex. There is a mass 1 of individuals, who purchase $x \in \{0, 1\}$ units. I assume for the moment that procedural utility is $\psi(x|e) = 0$ for all $x$; I relax this assumption in Section 5. Therefore, individuals have decision utility

$$u^{dec}(x|\phi) = x \cdot [\omega - p + \Lambda(e_0 + \phi)]$$

for $x \in \{0, 1\}$ units of consumption from a firm with price $p$ and environment $\phi$, where $\omega$ follows distribution $F$ in the population, and $\Lambda(e_0 + \phi)$ is a concave function. In other words, a consumer purchases one unit of the good if $\omega + \Lambda(e_0 + \phi) \geq p$. The firm maximizes profits:

$$\pi = (p - \kappa(\phi)) \cdot D(p,e)$$

where $D(p,e)$ is the demand for the good at price $p$ and environment $e$.

**Example 3** (Phishing consumers into overeating). Akerlof and Shiller (2015) give the example of how a restaurant can increase the smell of the food that it produces, so that it makes those who pass by feel hungry. Let $x \in \{0, 1\}$ be the choice of eating a tasty pastry. Suppose that $\phi$ represents the smell of the pastry, which exacerbates present bias $\Lambda(e_0 + \phi)$. Individuals’ outcome utility from the pastry is given by $\omega$.

Some economists have argued that competitive pressures in the market will reduce or eliminate behavioral biases (Laibson and Yariv, 2007; Becker, 2012; Blume and Easley, 2006; the last acknowledged that in incomplete markets this argument might not hold). Others (Mullainathan and Thaler, 2000; Issacharoff and Delaney, 2006) have been more skeptical about the ability of the market to correct behavioral biases. Loewenstein and Haisley (2008) point out that the sellers of commercial diets “make the most money by selling hope rather than actual results”, as in the in the market for quacks of Spiegler (2006). Gabaix and Laibson (2006) show that in a market with shrouded attributes and enough naive consumers, competition does not make firms disclose their shrouded attributes, but rather exploit naive consumers. There is evidence that market competition alone does not weed out products that are ineffective or harmful for consumers (Akerlof and Shiller, 2015). Ausubel (1991) argues that competition in the credit card market does not seem to reduce phishing; Carlin
(2009) shows that, in financial markets, competition might even increase phishing. In the section below I analyze this problem and show that competition indeed does not necessarily reduce phishing.

4.2 Competition does not reduce phishing

In this section I analyze a very simple and tractable setup, and prove that competition does not lead to less phishing in that setup. Suppose there is a monopolist that produces the good. Valuation $\omega$ for the good is distributed according to $F$ in the population. Because an individual only purchases if $\omega + \Lambda(e_0 + \phi) \geq p$, demand for the item at price $p$ in environment $\phi$ is given by $D(p, \phi) = 1 - F[p - \Lambda(e_0 + \phi)]$, and therefore profit maximization is given by:

$$
\max_{p, \phi} (p - \kappa(\phi)) \cdot (1 - F[p - \Lambda(e_0 + \phi)]),
$$

where the monopolist chooses both $p$ and $\phi$. Maximizing with respect to $p$ we find

$$
1 - F[p - \Lambda(e_0 + \phi)] = (p - \kappa(\phi)) \cdot f[p - \Lambda(e_0 + \phi)]. \tag{18}
$$

Maximizing with respect to $\phi$:

$$
\kappa'(\phi) = \Lambda'(e_0 + \phi) \cdot \frac{(p - \kappa(\phi))f[p - \Lambda(e_0 + \phi)]}{1 - F[p - \Lambda(e_0 + \phi)]} = \Lambda'(e_0 + \phi),
$$

where the second equality comes from Equation 18. Therefore, the monopolist invests in phishing up to the point where the marginal increase in per-unit cost equals the marginal internality.

Consider now the case with phishing competition, and let $J$ be the set of firms. There is free entry, meaning that profits must be zero: $p = \kappa(\phi_j)$ for any firm $j \in J$. I assume that there is an environment $e_j = e_0 + \phi_j$ that is particular to each firm $j \in J$, and only affects the products sold by that firm. This would be the case if the firms were restaurants, in which case the alluring pictures of food, or the smell of tasty food emanating from the kitchen, would be specific to that particular restaurant. Consumers only buy from the firm with the lowest
price, and therefore firm \( j \) must choose \( \phi_j \) such that \( \phi_j = \operatorname{arg\, max}_\phi \omega + \Lambda(e_0 + \phi) - \kappa(\phi) \), where I have used the fact that \( p = \kappa(\phi) \). Differentiating we find \( \kappa'(\phi_j) = \Lambda'(e_0 + \phi) \). From these calculations we have the following result.

**Proposition 9.** Phishing is identical with a monopoly and with phishing competition.

This conclusion was derived under some assumptions, the most crucial of which is that the phishing costs are per-unit costs, which is arguably a good approximation for the credit card or financial broker industry, where each client receives personalized treatment. If there are fixed costs in phishing, however, then we would expect a monopolist to phish more than a number of competitive firms, because the demand might not justify the investment in phishing in the case of many firms. However, there might be other reasons (which I have not included in the model) that make firms phish more when there is more competition, a phenomenon observed by Carlin (2009).

### 4.3 Competition can decrease welfare

Let \( \phi^* \) be the choice of \( \phi \) by firms, i.e. \( \kappa'(\phi^*) = \Lambda'(e_0 + \phi^*) \). While competition does not affect firms’ choices of \( \phi^* \), it still decreases the price of the good. To see this, note that the price in competition is \( p^C = \kappa(\phi^*) \); however, the price chosen by the monopolist must be higher, because profits are increasing in \( p \) at \( \phi^* \) and \( p = p^C \), and therefore the monopolist chooses a price \( p^M > p^C \). Whether competition increases consumer welfare is \textit{a priori} ambiguous: competition reduces prices, what increases the welfare of those who purchase; but it also induces some people to purchase who would be better off not doing so, hence reducing welfare. I assume for simplicity that there are no externalities, i.e. \( \eta(x) = 0 \).

In order to compute the change in welfare \( \Delta W \) from going from a monopolist to perfect competition, note that consumer welfare under a monopolist is expressed as \( \int_{\phi^M - \Lambda(\phi^*)}^{1}(\omega - p^M)dF(\omega) \), and under perfect competition as \( \int_{\phi^C - \Lambda(\phi^*)}^{1}(\omega - p^C)dF(\omega) \). Therefore, \( \Delta W \) is given by

\[
\Delta W = \int_{\phi^M - \Lambda(\phi^*)}^{1}(\omega - p^M)dF(\omega) - \int_{\phi^C - \Lambda(\phi^*)}^{1}(\omega - p^C)dF(\omega).
\]

\(^{54}\)Because the derivative of profits at \( \phi = \phi^* \), \( p = p^C = \kappa(\phi^*) \) is positive:

\[
\frac{\partial(p - \kappa(\phi))(1 - F(p - \Lambda(e_0 + \phi)))}{\partial p} \bigg|_{\phi = \phi^*, p = \kappa(\phi^*)} = 1 - F(0) > 0.
\]

28
\[
\Delta W = \int_{p^C - \Lambda(\phi^*)}^{p^M - \Lambda(\phi^*)} (\omega - p^C) dF(\omega) + \int_{p^M - \Lambda(\phi^*)}^{1} (p^M - p^C) dF(\omega). \tag{19}
\]

The second term is always positive because those consumers who would have purchased in either case prefer to do it at a lower price under competition. However, the first term can be negative for those consumers with \( \omega < p^C \) and if the internality \( \Lambda(e_0 + \phi^*) \) is relatively large. Suppose that \( \omega \) is distributed uniformly in the interval \([0, 1]\), so that \( F(\omega) = \omega \) if \( \omega \in [0, 1] \), that the internality is linear, so that \( \Lambda(e_0 + \phi) = A_\Lambda \cdot \phi \), and that the phishing cost is quadratic, \( \kappa(\phi) = A_\kappa \cdot \phi^2 \). Then, we have the following result.

**Proposition 10.** Under the assumptions above, a sufficient condition for \( \Delta W \) to be negative is that \( A_\Lambda^2 > A_\kappa \).

Note that in this case the internality is given by \( A_\Lambda \cdot \phi \), therefore \( A_\Lambda \) is the marginal change in the internality caused by phishing: when this change is large enough, compared to the cost of phishing, then there is a large amount of phishing in equilibrium, and a monopoly would increase consumer welfare (and hence total welfare). The reason is that higher prices have a duplicitous effect on consumer welfare: on the one hand, those who purchase are worse off, but on the other hand higher prices deter some consumers from making a purchase that would make them worse off (because their experienced utility would have been negative, had they purchased). In a sense, by restricting demand, the monopolist is helping precisely those who are hurt most by purchasing.\(^{55}\)

The fact that a monopolist can increase welfare in a situation where there are “frictions” is a well-known phenomenon (Lipsey and Lancaster, 1956). For example, if consumers generate large negative externalities, a monopoly can increase welfare by reducing consumption and hence the emission of externalities. Therefore, it should come as no surprise that, when consumers have internalities, a monopolist could raise welfare, applying the same reasoning as in the case of externalities and other frictions. However, this fact has not been sufficiently appreciated when it comes to internalities. This example suggests that, under some circum-

\(^{55}\)An interesting extension could assume that firms, instead of choosing \( \phi \) directly, can change \( \phi \) only through advertising. It is possible that perfect competition then reduces welfare if competition in advertising exacerbates people’s biases.
stances, deregulating markets and allowing for firms to compete in sectors where people’s behavior tends to be most biased (or firms can cheaply induce such biases) could actually reduce social welfare.

5 Nudging and Phishing in General Equilibrium

The welfare analysis of nudges has been performed mostly in a partial equilibrium context (if at all), where firms’ choices are taken as given. Recently, however, some economists (Spiegler, 2014; Handel, 2013; Grubb and Osborne, 2015) have argued for a deeper analysis taking general equilibrium effects into account, because firms’ reactions to a governmental nudge can potentially undo part (or all) of its intended effects. It is also possible, at least in principle, that a nudge could hurt social welfare partially by reducing demand for the firms, hence reducing producer surplus. Even if that is the case, some economists (Camerer et al., 2003) have argued that total surplus must increase with a well-designed designed nudge. In this section I will formalize that intuition, by analyzing what happens when the government has a nudge \( \nu \) at its disposal, but the firm can react to this nudge by changing its price \( p \) or phishing \( \phi \).

5.1 The theory in general equilibrium

As in Section 3.5, I assume that the environment \( e \) is multidimensional, \( e = (e^\Lambda, e^\psi) \), where one dimension of the environment affects internalities, and the other dimension affects procedural utility. The individual maximizes her decision utility \( u^{dec}(x|e) = \omega(x) - px + \psi(x|e^\psi) + \Lambda(x|e^\Lambda) \) and has experienced utility \( u^{exp}(x|e) = \omega(x) - px + \psi(x|e^\psi) \), where the environment \( e(\nu, \phi) \) depends on both the nudge and the phish, and I simply assume that \( \frac{\partial e^\psi}{\partial \nu} \leq 0 \), \( \frac{\partial e^\Lambda}{\partial \nu} \leq 0 \) and \( \frac{\partial e^\psi}{\partial \phi} \geq 0 \), \( \frac{\partial e^\Lambda}{\partial \phi} \geq 0 \). There is a representative firm, which has a per-unit cost \( \kappa_\phi(\phi) \), and solves:

\[
\max_{p, \phi} \quad (p - \kappa_\phi(\phi)) \cdot x(p, \phi, \nu),
\]

where demand \( x(p, \phi, \nu) \) is decreasing in price \( p \) and increasing in \( \phi \).
Definition 3. Given a nudge \( \nu \), a general equilibrium is a tuple \((x, p, \phi)\) of consumer choice \( x \), price \( p \) and phish \( \phi \), such that:

1. consumers maximize their decision utility given \( p \), \( \phi \) and the nudge \( \nu \), and
2. firms maximize their profits given consumer behavior and the nudge \( \nu \).

Until this point I have ignored the costs of nudges derived strictly from their implementation; however we need to include all costs into our analysis, if we want to obtain a true general equilibrium effect. Therefore, I assume that the social planner has a per-unit implementation cost for the nudge \( \kappa(\nu) \). Both \( \kappa(\nu) \) and \( \kappa(\phi) \) are increasing and convex.

The social planner maximizes social welfare \( W \), which is the sum of consumer and producer welfare (net of implementation costs). Moreover, I assume that there are no externalities:

\[
W = \mathbb{E}_\theta \left[ \omega(x|\theta) + \psi(x|e^\psi, \theta) - (p + \kappa(\nu)) \cdot x + (p - \kappa(\phi)) \cdot x \right],
\]

where \( x = x(p, \phi, \nu, \theta) \).

Proposition 11. If positive, the optimal nudge in general equilibrium is given by

\[
E_\theta \left[ \frac{\partial x}{\partial \nu} \left( -\frac{\partial \Lambda}{\partial x} + \kappa(\nu) \right) + \frac{\partial \phi}{\partial \nu} \frac{\partial e^\psi}{\partial \phi} \frac{\partial \psi}{\partial e^\psi} - \frac{\partial p}{\partial \nu} \cdot x \right] = \mathbb{E}_\theta \left[ - \frac{\partial x}{\partial \nu} (p - \kappa(\phi)) - \frac{\partial e^\psi}{\partial \nu} \frac{\partial \psi}{\partial e^\psi} + \kappa'(\nu) \cdot x \right].
\]

(20)

where \( \frac{\partial x}{\partial \nu} = \frac{\partial x}{\partial \nu} + \frac{\partial x}{\partial p} \frac{\partial p}{\partial \nu} + \frac{\partial x}{\partial \phi} \frac{\partial \phi}{\partial \nu} \), \( \Lambda = \Lambda(x|e^\Lambda, \theta) \) and \( \psi = \psi(x|e^\psi, \theta) \).

Proposition 11 is a generalization of Proposition 3, taking into account firms’ reactions in general equilibrium. In the left-hand side of Equation 20, the first term measures the reduction of the internality, net of the cost of nudging; the second term is the the impact on procedural utility from the nudge through the change in the firm phishing behavior; and the last term measures the gain in consumer surplus from the reduction in prices induced by the nudge. On the right-hand side, the first term measures the loss in producer surplus from the reduction of demand from the nudge, the second term is the usual impact of the nudge on procedural utility (a psychological cost) and the last term is the marginal increase in the

---

\(^{56}\)Including externalities in Proposition 11 is straightforward: the first term of the left-hand side of Equation 20 would be \( - \frac{\partial x}{\partial \nu} \left( \frac{\partial \Lambda + \eta}{\partial x} + \kappa(\nu) \right) \).
per-unit part of the cost of the nudge. Camerer et al. (2003) argued that implementation costs should be taken into account when discussing the benefits of nudges, and Proposition 11 captures the intuition that nudging more is more expensive (term $\kappa'(\nu) \cdot x$), but at the same time some of those costs are directly offset by the reduction in the demand for the nudged behavior, which appears in the term $-\frac{\partial x}{\partial \nu} \cdot \kappa_{\nu}(\nu)$.

5.2 Competition: nudging is optimal under mild conditions

Proposition 11 informs us about the characteristics of the optimal nudge, when this nudge is positive. However, a simpler and very relevant question is to know when is it that the optimal nudge is positive. Therefore I will look at sufficient conditions under which some amount of nudging would be optimal. I consider the limit of perfect competition, where the firm’s profits tend to zero, $(p - \kappa_{\phi}(\phi)) \rightarrow 0$.

Proposition 12. In general equilibrium in the limit of perfect competition, the firm

- reacts to the nudge in such a way that quantity demanded does not change: $\frac{\partial E_{\theta}[x]}{\partial \nu} \rightarrow 0$,
- keeps the amount of phishing constant, $\frac{\partial \phi}{\partial \nu} \rightarrow 0$.

Moreover, if the nudge is effective $\left(\frac{\partial E_{\theta}[x]}{\partial \nu} < 0\right)$, the firm lowers prices: $\frac{\partial p}{\partial \nu} \rightarrow -\frac{\partial E_{\theta}[x]}{\partial \phi} \frac{\partial \phi}{\partial p} < 0$.

In other words, when the nudge is effective in partial equilibrium $\left(\frac{\partial E_{\theta}[x]}{\partial \nu} < 0\right)$, then in the limit of perfect competition, the firm reacts in such a way that the quantity demanded and the amount of phishing both remain constant, and lowers the price as to compensate for the effect of the nudge. It turns that out that the limit of perfect competition, together with some mild assumptions on the nudge, are sufficient for some amount of nudging to be optimal.

Corollary 13. Under perfect competition, the following conditions are sufficient for the optimal nudge to be positive:

1. the nudge is effective: $\frac{\partial E_{\theta}[x]}{\partial \nu} < 0$,
2. for a small nudge, the implementation costs are small: $\kappa_{\nu}(0) = 0$, and $\kappa'_{\nu}(0) = 0$,
3. For a small nudge the psychological costs are small: 
\[ \frac{\partial e}{\partial \psi} \frac{\partial \psi}{\partial \nu} \bigg|_{\nu=0} = 0. \]

In the limit of perfect competition, price equals approximates per-unit cost of phishing and therefore social welfare approximates consumer welfare. In that case, nudging is desirable whenever it is effective and has a low cost (both in terms of implementation and psychological costs). Corollary 13 is remarkable, because it shows that under circumstances often assumed in economic applications (like perfect competition), when the nudge works and is not too costly, the optimal nudge intensity is positive. Revisiting the concept of asymmetric paternalism that we saw in section 3.3, the Corollary provides a formal justification for the argument of Camerer et al. (2003) that “any asymmetrically paternalistic policy that helps boundedly rational consumers make better choices must, on net, increase economic efficiency as measured by the sum of consumer and producer surplus”.

5.2.1 When nudging backfires in general equilibrium

Nudging can backfire if firms react to the nudge in a such a way that, in the new equilibrium with the nudge, individuals are worse off. Handel (2013) argues that consumers’ inertia (when choosing a health plan) helps alleviate problems that arise from asymmetric information; he finds that when a nudge is introduced that reduces inertia, firms react by rising prices in a way that ultimately hurts consumers. These arguments are not surprising and are not confined to behavioral welfare economics and nudging: since at least Lipsey and Lancaster (1956), it is known that in the presence of market failures (imperfect competition, externalities, etc.), solving one of them in isolation can result in a welfare loss. For example, if a there is a single firm that is polluting a river, introducing competition might exacerbate the contamination of the river, increasing the externality and ultimately reducing welfare.

Corollary 13 sheds light on these “negative” results (Handel, 2013; Grubb and Osborne, 2015) argue that bill shock alerts (that warn consumers when they are exceeding a certain threshold in their cellphone consumption), if introduced during the period 2002-2004, would have reduced annual consumer welfare by $33. Again, this is because firms would have raised prices, hurting consumers in the process.

57 That is, there is a nudge which increases welfare but, as we saw in Section 3.3, not any nudge will have that effect.

58 Grubb and Osborne (2015) argue that bill shock alerts (that warn consumers when they are exceeding a certain threshold in their cellphone consumption), if introduced during the period 2002-2004, would have reduced annual consumer welfare by $33. Again, this is because firms would have raised prices, hurting consumers in the process.
suggesting that they might be the result of market imperfections, and not necessarily the particular nudged employed. This result could not have been obtained in a model that was more tailored to a particular application, and this highlights the advantages of building a general and tractable framework to analyze nudging and phishing. The result in Corollary 13 is true in the absence of market failures (and some additional assumptions); not surprisingly if there are multiple market failures such as asymmetric information and internalities, fixing only one of them can result in a welfare loss. This means that, when using nudges as a policy tool, there is a need to seriously consider their general equilibrium effects.

6 Conclusion

The contribution of this paper is to provide a tractable and general framework to analyze the welfare consequences of nudges. My framework encompasses several of the previous models found in the literature, and provides new insights that can be translated into the policy applications of nudges. An important result for implementing nudges is Proposition 1, which shows that the social planner does not need to know outcome preferences (which do not depend on the environment, hence are presumably “deeper”) in order to design the optimal nudge. This finding is of great importance because it addresses one of the main arguments against nudges: that the social planner needs to know people’s preferences better than they do (which, in the presence of internalities, is particularly difficult to accomplish).

As I argued in Section 3.2, the social planner could in principle estimate the parameters that she needs to know either from historical data (change of demand with the nudge, externalities) or from structural estimation coupled with experiments (internalities, procedural utility), without ever needing to recover individuals’ full preferences.

Having a general framework for nudges is fundamental to obtain results that are valid across domains. As an example, Corollary 13 shows that under perfect competition and some mild conditions, nudging is optimal in general equilibrium. This result can only be
obtained in a framework which is general enough to incorporate nudges and firms’ reactions within equilibrium. The conditions of the Corollary can then be checked for any specific nudge that we want to implement. Indeed, one of the main problems in the application of nudges is that there is no “taxonomy” to classify nudges, and where to search for a nudge that is tailored to a particular situation. By providing a general framework, I aim to contribute to a systematic analysis of nudges, where they are classified according (among other characteristics) on their effects on demand, on internalities, and on psychological costs. Doing so would allow a more systematic analysis of the arguments about using a nudge in a specific policy application.

This paper opens new directions for further research. I have assumed that all the choices happen in a static context; however many decisions are dynamic, with consequences becoming apparent over time, and with possibilities of learning, as well as habit formation. Another interesting direction of research is connected to random choice: in the framework presented in this paper I have assumed that choice is deterministic, and therefore that differences in choice could only result from differences in the characteristics of the choice problem. However, actual choice has a degree of randomness (Webb, 2013), and a direction for research would be integrating random choice (Gul et al., 2014) into the framework. Moreover, in Proposition 1 there are two components that have not traditionally been measured by economists: the marginal internality and the psychological cost of the nudge. Recently, researchers have started combining structural models and field experiments to estimate internalities, and the same can be done to estimate psychological costs. A second direction of research stems from using non-choice data, in particular data from psychology and neuroscience, where advances have been made in underpinning the neural substrates of decision vs. experienced utility (Berridge and Robinson, 2003; Berridge and O’Doherty, 2014). This knowledge can guide economists in the design of structural models with which to estimate internalities and psychological costs, which are the necessary components to design optimal nudges for the most common interventions, in order to increase people’s well-being.
7 Appendix

GENERAL PROOF OF PROPOSITION 1 FOR ARBITRARY UTILITY FUNCTIONS.

Suppose that experienced utility is given by a general function \( u^{\text{exp}}(x|e) = \Xi^{\text{exp}}(\omega(x,p), \psi(x|e)) \), and decision utility is given by \( u^{\text{dec}}(x|e) = \Xi^{\text{dec}}(u^{\text{exp}}(x|e), \Lambda(x|e)) \). Moreover, suppose that \( u^{\text{exp}} \) and \( u^{\text{dec}} \) are concave in \( x \). The individual solves \( \max_x u^{\text{dec}}(x|e) \), and the (necessary and sufficient) first order conditions are

\[
[x] \quad \partial_x \Xi^{\text{dec}} \cdot \left[ \partial_1 \Xi^{\text{exp}} \cdot \frac{\partial \omega(x,p)}{\partial x} + \partial_2 \Xi^{\text{exp}} \cdot \frac{\partial \psi(x|e)}{\partial x} \right] + \partial_2 \Xi^{\text{dec}} \cdot \frac{\partial \Lambda(x|e)}{\partial x} = 0, \quad (21)
\]

where

\[
\partial_i \Xi^{\text{exp}} = \frac{\partial \Xi^{\text{exp}}(x_1, x_2)}{\partial x_i} \bigg|_{x_1 = \omega(x,p), x_2 = \psi(x|e)}
\]

and analogously for \( \partial_i \Xi^{\text{exp}} \). The social planner chooses \( \nu \) to maximize \( W \):

\[
[\nu] \quad \frac{\partial x}{\partial \nu} \cdot \left[ \partial_1 \Xi^{\text{exp}} \cdot \frac{\partial \omega(x,p)}{\partial x} + \partial_2 \Xi^{\text{exp}} \cdot \frac{\partial \psi(x|e)}{\partial x} - \eta'(x) \right] - \partial_2 \Xi^{\text{exp}} \cdot \frac{\partial \psi(x|e)}{\partial e} = 0. \quad (22)
\]

We can substitute for the terms between brackets in Equation 22 by using Equation 21:

\[
- \frac{\partial x}{\partial \nu} \left[ \partial_1 \Xi^{\text{dec}} \cdot \frac{\partial \Lambda(x|e)}{\partial x} + \eta'(x) \right] = \partial_2 \Xi^{\text{exp}} \cdot \frac{\partial \psi(x|e)}{\partial e}. \quad (23)
\]

When \( u^{\text{exp}}(x|e) = \omega(x) - px + \psi(x|e) \), and \( u^{\text{dec}}(x|e) = u^{\text{exp}}(x|e) + \Lambda(x|e) \), then \( \partial_i \Xi^{\text{exp}} = \partial_i \Xi^{\text{dec}} = 1 \), and therefore Equation 23 simplifies to the expression in Proposition 1.

PROOF OF COROLLARY 2. The first order conditions of the individual are given by

\[
\omega'(x) - p + \frac{\partial \Lambda(x|e)}{\partial x} + \frac{\partial \psi(x|e)}{\partial x} = 0, \quad (24)
\]

where \( e = e_0 - \nu \). Using the Implicit Function Theorem on Equation 24 with respect to \( \nu \) we find

\[36\]
\[ \frac{\partial x}{\partial \nu} = -\frac{\partial^2 x}{\partial \nu^2} \left[ \Lambda(x|e) + \psi(x|e) \right] \]  

Using the Implicit Function Theorem, this time with respect to \( p \), we have

\[ \frac{\partial x}{\partial p} = \frac{1}{\frac{\partial^2 x}{\partial \nu^2} \left[ \Lambda(x|e) + \psi(x|e) \right]} . \]  

Therefore, from Equations 25 and 26 we have that

\[ \frac{\partial x}{\partial \nu} = \frac{\partial x}{\partial p} \cdot \frac{\partial^2 \left[ \Lambda + \psi \right]}{\partial x \partial e} \]  

Note that the term \( \frac{\partial^2 \left[ \Lambda + \psi \right]}{\partial x \partial e} \) is always positive, because of Assumptions 1 and 2. Plugging the expression in Equation 27 into Proposition 1, we obtain the desired result.

**Proof of Proposition 3.** Proposition 3 is a particular case of Proposition 11.

**Proof of Proposition 4.** Optimizing welfare with respect to \( \nu \) and \( \tau \):

\[ \frac{\partial W}{\partial \nu} = \mathbb{E}_\theta \left[ \frac{\partial x}{\partial \nu} \cdot \left( \frac{\partial [\omega(x) + \psi(x|e) - \eta(x)]}{\partial x} - (p + \tau) \right) - \frac{\partial \psi(x|e)}{\partial e} + \tau \cdot \frac{\partial x}{\partial \nu} \cdot \delta'(\tau \cdot x) \right] , \]  

\[ \frac{\partial W}{\partial \tau} = \mathbb{E}_\theta \left[ \frac{\partial x}{\partial \tau} \cdot \left( \frac{\partial [\omega(x) + \psi(x|e) - \eta(x)]}{\partial x} - (p + \tau) \right) - x + \left( x + \tau \cdot \frac{\partial x}{\partial \tau} \right) \cdot \delta'(\tau \cdot x) \right] . \]  

The first order conditions for the individual are given by

\[ \frac{\partial \omega(x|\theta)}{\partial x} + \frac{\partial \Lambda(x|e, \theta)}{\partial x} + \frac{\psi(x|e, \theta)}{\partial x} - (p + \tau) = 0 . \]  

Using Equation 30, we can substitute the terms between the inner brackets in Equations 28 and 29, and obtain

\[ \frac{\partial W}{\partial \nu} = \mathbb{E}_\theta \left[ -\frac{\partial x}{\partial \nu} \cdot \frac{\partial [\Lambda(x|e, \theta) + \eta(x)]}{\partial x} - \frac{\partial \psi(x|e)}{\partial e} + \tau \cdot \frac{\partial x}{\partial \nu} \cdot \delta'(\tau \cdot x) \right] , \]  

\[ \frac{\partial W}{\partial \tau} = \mathbb{E}_\theta \left[ -\frac{\partial x}{\partial \tau} \cdot \frac{\partial [\Lambda(x|e, \theta) + \eta(x)]}{\partial x} - x + \left( x + \tau \cdot \frac{\partial x}{\partial \tau} \right) \cdot \delta'(\tau \cdot x) \right] . \]
**PROOF OF COROLLARY 5.** When taxation is fully efficient, we have that \( \delta(x) = x \), and therefore \( \delta'(x) = 1 \). From Equation 14, the optimal tax is given by

\[
\mathbb{E}_\theta \left[ -\frac{\partial x}{\partial \tau} \cdot \frac{\partial [\Lambda(x|e, \theta) + \eta(x)]}{\partial x} \right] = -\tau \cdot \mathbb{E}_\theta \left[ \frac{\partial x}{\partial \tau} \right].
\] (33)

When the population is homogeneous, Equation 33 becomes

\[
-\frac{\partial x}{\partial \tau} \cdot \frac{\partial [\Lambda(x|e, \theta) + \eta(x)]}{\partial x} = -\tau \cdot \frac{\partial x}{\partial \tau},
\]

and the terms \(-\frac{\partial x}{\partial \tau}\) cancel on both sides of the equality, obtaining the desired result. \(\square\)

**PROOF OF PROPOSITION 7.** We have two cases: that the optimal tax is \( \tau^* = 0 \), or \( \tau^* > 0 \). If \( \tau^* = 0 \), then

\[
W'(\nu) = -\frac{\partial x}{\partial \nu} \cdot \frac{\partial [\Lambda(x|e) + \eta(x)]}{\partial x} - \frac{\partial \psi(x|e)}{\partial e},
\]

and \( W'(0) > 0 \). If \( \tau^* > 0 \), then

\[
\tau^* = \frac{-\frac{\partial x}{\partial \tau} \cdot \frac{\partial [\Lambda + \eta]}{\partial x} - (1 - \alpha)x}{-\alpha \cdot \frac{\partial x}{\partial \tau}}
\] (34)

Note that \( \tau \) is increasing in \( \alpha \), because

\[
\text{sign}(\tau'(\alpha)) = \text{sign} \left( -\frac{\partial x}{\partial \tau} \cdot \left[ \alpha x + \left( \frac{\partial x}{\partial \tau} \cdot \frac{\partial [\Lambda + \eta]}{\partial x} - (1 - \alpha)x \right) \right] \right) = +.
\]

That means that for any \( \alpha < 1 \),

\[
\alpha \tau^*(\alpha) < \tau^*(1) = \frac{\partial [\Lambda + \eta]}{\partial x},
\]

and therefore
\[ W'(\nu) = \left[ -\frac{\partial}{\partial \nu} \left( \frac{\partial [\Lambda(x|e) + \eta(x)]}{\partial x} - \alpha \tau^*(\alpha) \right) \right] > 0, \]

so \( W'(0) > 0 \) because \( \frac{\partial \psi}{\partial e} \bigg|_{\nu=0} = 0. \)

PROOF OF PROPOSITION 8. Because the individual chooses \( x^1 \) over \( x^2 \) in environment \( e_1 = (e_1^\Lambda, e_1^\psi) \), it must be true that

\[
\omega(x^1) - px^1 + \Lambda(x^1|e_1^\Lambda) + \psi(x^1|e_1^\psi) \geq \omega(x^2) - px^2 + \Lambda(x^2|e_1^\Lambda) + \psi(x^2|e_1^\psi),
\]

or equivalently

\[
\omega(x^1) - px^1 + \psi(x^1|e_1^\psi) - \eta(x^1) \geq \omega(x^2) - px^2 + \psi(x^2|e_1^\psi) - \eta(x^2) \geq \omega(x^1) - px^1 + \Lambda(x^1|e_1^\Lambda) + \psi(x^1|e_1^\psi) - \eta(x^1).
\]

A sufficient condition for \( \nu_1 \) to be better than \( \nu_2 \) is that the expression marked with ♣ be positive, and that is exactly Equation 16 in Proposition 8. On the other hand, because the individual chose \( x^2 \) over \( x^1 \) in environment \( e_2 = (e_2^\Lambda, e_2^\psi) \), it holds that

\[
\omega(x^2) - px^2 + \Lambda(x^2|e_2^\Lambda) + \psi(x^2|e_2^\psi) \geq \omega(x^1) - px^1 + \Lambda(x^1|e_2^\Lambda) + \psi(x^1|e_2^\psi),
\]

or equivalently,

\[
\omega(x^2) - px^2 + \psi(x^2|e_2^\psi) - \eta(x^2) \geq \omega(x^1) - px^1 + \psi(x^1|e_2^\psi) - \eta(x^1) \geq \omega(x^2) - px^2 + \Lambda(x^2|e_2^\Lambda) + \psi(x^2|e_2^\psi) - \eta(x^2).
\]

A necessary condition for nudge \( \nu_1 \) to be better than \( \nu_2 \) is that the expression marked with ♠ be negative, and that is equivalent to Equation 17 in Proposition 8.

PROOF OF PROPOSITION 10. Solving the integrals in Equation 19 and simplifying, we obtain that the difference in welfare is
\[ \Delta W = \frac{(p^M - A\Lambda \phi^*)^2}{2} - \frac{(p^C - A\Lambda \phi^*)^2}{2} - \frac{(p^M - p^C)(p^C + p^M - 1 - A\Lambda \phi^*)}{2}. \tag{35} \]

Let \( p^M = p^C + \epsilon \). Simplifying in Equation \( 35 \), we have \( \Delta W = -\frac{\epsilon^2}{2} - \epsilon(p^C + 2A\Lambda \phi^* - 1) \).

Therefore, if \( p^C + 2A\Lambda \phi^* > 1 \), the change in consumer welfare would be negative. Note that \( \kappa'(\phi^*) = A\Lambda \phi^* = A\Lambda \), and hence \( \phi^* = \frac{A\Lambda}{\kappa} \). Moreover \( p^C = \kappa(\phi^*) = \frac{A\Lambda^2}{\kappa} \). But then we have \( p^C + 2A\Lambda \phi^* = \frac{A\Lambda^2}{\kappa} + 2\frac{A\Lambda^2}{\kappa} \), and a sufficient condition for this expression to be greater than 1 is \( A\Lambda^2 > \kappa \).

**PROOF OF PROPOSITION 11.** Social welfare is given by

\[ W = \int \omega(x) + \psi(x|e^\psi) - (p + \kappa_\nu(\nu)) \cdot x + (p - \kappa_\phi(\phi)) \cdot x \, dF(\theta). \]

Derivating with respect to \( \nu \), we obtain

\[ W'(\nu) = \int \frac{\partial x}{\partial \nu} \left( \frac{\partial \omega}{\partial x} + \frac{\partial \psi}{\partial x} - \kappa_\nu(\nu) - \kappa_\phi(\phi) \right) + \frac{\partial e^\psi}{\partial \nu} \frac{\partial \psi}{\partial e^\psi} - \kappa'_\nu(\nu) \cdot x + \]

\[ + \frac{\partial p}{\partial \nu} \left[ \frac{\partial x}{\partial p} \left( \frac{\partial \omega}{\partial x} + \frac{\partial \psi}{\partial x} - (p + \kappa_\nu(\nu)) \right) - x + (p - \kappa_\phi(\phi)) \frac{\partial x}{\partial p} + x \right] + \]

\[ + \frac{\partial \phi}{\partial \nu} \left[ \frac{\partial x}{\partial \phi} \left( \frac{\partial w}{\partial x} + \frac{\partial \psi}{\partial x} - (p + \kappa_\nu(\nu)) \right) + \frac{\partial x}{\partial \phi} (p - \kappa_\phi(\phi)) - \kappa'_\phi(\phi) \cdot x + \frac{\partial e^\psi}{\partial \phi} \frac{\partial \psi}{\partial e^\psi} \right] dF(\theta). \]

Note that, given a nudge \( \nu \), the firm maximizes

\[ (p - \kappa_\phi(\phi)) \cdot x(p, \phi) \]

where (abusing notation) I denote \( x(p, \phi) = E_\theta[x(\theta, p, \phi, \nu)] \) is the demand function.
Therefore, the first order conditions for the firm are

\[
\begin{align*}
[p] & \quad x + (p - \kappa_x(\phi)) \frac{\partial x}{\partial p} = 0, \\
[\phi] & \quad -\kappa'(\phi)x + (p - \kappa_x(\phi)) \frac{\partial x}{\partial \phi} = 0.
\end{align*}
\]  

(36) (37)

From the fact that firm maximizes, the expressions indicated with ▼ and ◊ cancel out due to Equations 36 and 37 respectively. Moreover, the individual solves the problem

\[
\max_x \omega(x|\theta) + \Lambda(x|\theta, e^x(\phi, \nu)) + \psi(x|\theta, e^y(\phi, \nu)) - p \cdot x,
\]

and hence

\[
\frac{\partial}{\partial x}(\omega + \Lambda + \psi) - p = 0.
\]  

(38)

Using Equation 38, we can substitute the term \(\frac{\partial \omega}{\partial x} + \frac{\partial \psi}{\partial x}\), and hence the expression for \(W'(\nu)\) becomes

\[
W'(\nu) = \mathbb{E}_\theta \left[ \mathbb{D}_\nu \left( -\frac{\partial \Lambda}{\partial x} - \kappa_\nu(\nu) \right) + \frac{\partial x}{\partial \nu} \left( p - \kappa_x(\phi) \right) - \kappa'_\nu(\nu) \cdot x + \frac{\partial e^x}{\partial \nu} \frac{\partial \psi}{\partial e} - \frac{\partial p}{\partial \nu} + \frac{\partial \phi}{\partial \nu} \frac{\partial e^\phi}{\partial \nu} \right],
\]

(39)

and hence the optimal nudge is given by Equation 20.

\[\square\]

**PROOF OF PROPOSITION 12.** Differentiating implicitly with respect to \(\nu\) in Equations 36 and 37, we obtain:

\[
\frac{\partial x}{\partial \nu} + \left( \frac{\partial p}{\partial \nu} - \kappa'_\phi(\phi) \cdot \frac{\partial \phi}{\partial \nu} \right) \frac{\partial x}{\partial p} + (p - \kappa_x(\phi)) \frac{\partial^2 x}{\partial p \partial \nu} = 0. 
\]  

(40)

\[
-\kappa''(\phi) \frac{\partial \phi}{\partial \nu} \cdot x - \kappa'_\phi(\phi) \cdot \frac{\partial x}{\partial \nu} + \left( \frac{\partial p}{\partial \nu} - \kappa'_\phi(\phi) \cdot \frac{\partial \phi}{\partial \nu} \right) \cdot \frac{\partial x}{\partial \phi} + (p - \kappa_x(\phi)) \frac{\partial^2 x}{\partial \phi \partial \nu} = 0.
\]  

(41)

From Equation 37 we have that

\[
\kappa'_\phi(\phi) = \frac{(p - \kappa_x(\phi)) \frac{\partial x}{\partial \phi}}{x}.
\]

41
From Equation 36, we have that

\[
\frac{(p - \kappa_\phi(\phi))}{x} = -\frac{1}{\partial x/\partial p}.
\]

Combining both equations, we obtain

\[
\kappa_\phi'(\phi) = -\frac{\partial x/\partial \phi}{\partial x/\partial p}.
\]  

(42)

Solving in Equations 40 and 41, and using the characterization of \( \kappa_\phi'(\phi) \) from Equation 42, we have that

\[
\frac{\partial p}{\partial \nu} = -\frac{\partial x}{\partial \nu} + \frac{\partial x}{\partial \phi} \frac{\partial \phi}{\partial \nu} + \left( p - \kappa_\phi(\phi) \right) \frac{\partial^2 x}{\partial \phi \partial \nu}.
\]  

(43)

\[
\frac{\partial \phi}{\partial \nu} = \frac{\partial x/\partial \phi}{\partial x/\partial p} \frac{\partial \phi}{\partial \nu} + \frac{\partial \phi}{\partial \phi} \frac{\partial x}{\partial \nu} - \left( p - \kappa_\phi(\phi) \right) \frac{\partial^2 x}{\partial \phi \partial \nu}.
\]  

(44)

From Equation 43 we have

\[
0 = \frac{\partial x}{\partial \nu} + \left( p - \kappa_\phi(\phi) \right) \frac{\partial^2 x}{\partial \phi \partial \nu}.
\]

therefore

\[
\frac{\partial x}{\partial \nu} = -(p - \kappa_\phi(\phi)) \frac{\partial^2 x}{\partial \phi \partial \nu},
\]

and when \((p - \kappa_\phi(\phi)) \rightarrow 0\), then \(\frac{\partial x}{\partial \nu} \rightarrow 0\).

Solving for \(\frac{\partial p}{\partial \nu}\) in the system of two unknowns and two equations (43 and 44), we obtain

\[
\frac{\partial p}{\partial \nu} = \frac{\kappa_\phi'(\phi) \cdot x \cdot \frac{\partial x}{\partial \nu} + \left( p - \kappa_\phi(\phi) \right) \cdot \left[ \frac{\partial^2 x}{\partial \phi \partial \nu} \left( \kappa_\phi'(\phi) \cdot x - \left( \frac{\partial x}{\partial \nu} \right)^2 \right) + \frac{\partial x}{\partial \phi} \frac{\partial^2 x}{\partial \phi \partial \nu} \right]}{-\frac{\partial x}{\partial \phi} \kappa_\phi'(\phi) \cdot x}
\]  

(45)

When \((p - \kappa_\phi(\phi)) \rightarrow 0\), then from Equation 45

\[
\frac{\partial p}{\partial \nu} \rightarrow -\frac{\partial x}{\partial \nu} < 0,
\]

42
as long as \( \frac{\partial \phi}{\partial \nu} < 0 \). Plugging the characterization of \( \frac{\partial \phi}{\partial \nu} \) from Equation 45 on Equation 44, we have:

\[
\frac{\partial \phi}{\partial \nu} = \left( p - \kappa_\phi(\phi) \right) \cdot \frac{-\frac{\partial x}{\partial \phi} \left( \frac{\partial^2 \phi}{\partial \nu \partial \phi} \right)}{\kappa''(\phi) \cdot x - \frac{\partial x/\partial \phi}{\partial x/\partial \phi} + \frac{\partial^2 x}{\partial \phi^2}} + \frac{\partial^2 x}{\partial \phi^2}
\]

When \( (p - \kappa_\phi(\phi)) \to 0 \), then \( \frac{\partial \phi}{\partial \nu} \to 0 \).

\( \square \)

**PROOF OF COROLLARY 13.** Suppose that we have the following conditions:

1. the nudge is effective: \( \frac{\partial E_\theta[x]}{\partial \nu} < 0 \),
2. for a small nudge, the implementation costs are small: \( \kappa_\nu(0) = 0 \), and \( \kappa'_\nu(0) = 0 \),
3. for a small nudge the psychological costs are small: \( \frac{\partial e}{\partial \nu} \bigg|_{\nu=0} = 0 \).

Because the nudge is effective, and Because of Proposition 12, in the limit of perfect competition the expression in Equation 39 from the proof of Proposition 11 becomes

\[
W'(\nu) \to E_\theta \left[ -\kappa'_\nu(\nu) \cdot x + \frac{\partial e}{\partial \nu} \frac{\partial \psi}{\partial e} - x \frac{\partial p}{\partial \nu} \right]
\]

From the conditions on the implementation and psychological costs of the nudge, we have that the last expression, evaluated at \( \nu = 0 \), becomes:

\[
W'(0) \to -E_\theta[x] \cdot \frac{\partial p}{\partial \nu} \bigg|_{\nu=0} > 0,
\]

since \( \frac{\partial p}{\partial \nu} \to -\frac{\partial E_\theta[x]/\partial \nu}{\partial E_\theta[x]/\partial p} < 0 \), by Proposition 12.

\( \square \)

**References**


49


