Optimal Defaults with Normative Ambiguity

Jacob Goldin Daniel Reck[*]

December 14, 2018

Abstract

Default effects are pervasive, but the reason they arise is often unclear. We study optimal policy when it is ambiguous whether an observed default effect reflects a welfare-relevant preference or a mistake by decision-makers. Within a broad class of models, determining optimal policy is impossible without resolving this normative ambiguity. Depending on the resolution, optimal policy tends in opposite directions: either minimizing the number of non-default choices or promoting active choices. We illustrate our results using data on pension contribution defaults. When selecting a non-default option reduces employee welfare by less than $160, the optimal policy promotes active choices.

JEL Codes: D60, D90

[*] Goldin: Stanford Law School and NBER; email: jsgoldin@law.stanford.edu. Reck: Department of Economics, London School of Economics; email: d.h.reck@lse.ac.uk. We thank Ian Ayres, Doug Bernheim, Stefano DellaVigna, Alex Gelber, Wojciech Kopczuk, Claus Kreiner, Alex Rees-Jones, Emmanuel Saez, Dmitry Taubinsky, Eric Talley, and seminar participants at Berkeley, the University of Copenhagen, Vanderbilt, the National Tax Association Annual Conference, and the ASSA annual meeting for helpful discussion and comments.

A fundamental challenge in behavioral economics is determining whether some observed behavior reflects a mistake by decision-makers. If the behavior violates the predictions of a neoclassical model of decision-making, one can typically modify the preferences of the agents in the model to rationalize the observed decisions. Doing so can realign the model's predictions with the data, but introduces normative ambiguity; that is, do the modified preferences actually reflect the decision-maker's welfare or do they instead reflect a divergence between welfare and choice? Conducting welfare analysis in behavioral settings requires confronting this issue.

We study this problem for the case of defaults effects.[1] Decision-makers may act as if there is some cost to choosing an option other than the default, but this apparent cost may or may not actually affect decision-makers' welfare. In theory, this normative ambiguity might be resolved by identifying the precise behavioral model that generates default effects. In practice, distinguishing between alternative behavioral models is difficult in most settings in which default effects are observed. Prior research that studies optimal policy in this area has addressed the issue either by assuming a specific behavioral model (Carroll et al., 2009) or by considering the robustness of optimal policy conclusions to several alternative models (Bernheim, Fradkin and Popov, 2015). However, these approaches do not shed light on which features of the behavioral model drive the optimal policy results, making it difficult to extrapolate the results to settings in which other behavioral models may be at play, or settings in which additional policy tools are available.

The starting point for our approach is the fact that for a broad class of decision-making models, the effect of defaults on behavior can be characterized in terms of two ingredients: (1) decision-makers' utility over the menu of available options, and (2) an "as-if" cost to selecting an option that is not the default. This implied cost to opting out of the default is defined so as to rationalize decision-makers' observed behavior; decision-makers behave *as if* they face a cost of this magnitude to opting out of the default. Unlike standard revealed preference models, we do not impose that as-if costs actually reduce the welfare of the decision-makers who choose to incur them. Instead, we parameterize the degree to which as-if costs are normative (that is, the degree to which they enter into decision-makers' welfare). Concretely, then, the *normative ambiguity* we are referring to corresponds to uncertainty in the degree to which as-if costs are normative.

Alternative candidate models of default effects imply different conclusions about the degree to which as-if costs are normative. For example, one possible explanation for default effects is that decision-makers rationally seek to avoid exerting the mental effort required to choose between non-default options. In this model, all as-if costs will be normative. Alternatively, decision-makers might seek to avoid exerting mental effort, but systematically over-estimate the amount of effort that will be required to choose between the

[1] A large empirical literature documents the effect of defaults on behavior. Prominent examples include Madrian and Shea (2001); Johnson and Goldstein (2003); Haggag and Paci (2014); Handel (2013); Beshears et al. (2017).

non-default options. In this model, some – but not all – as-if costs will be normative. A third possibility is that decision-makers might inadvertently fail to consider making an active decision in the first place, in which case none of the observed as-if costs will be normative. Although falsifying specific candidate models might be possible with the right data, it is difficult to conceive of a convincing empirical test for determining the share of as-if costs that enter into decision-makers' welfare. This dilemma is worsened by the fact that decision-makers may be heterogeneous with respect to the model of decision-making that explains their behavior, and hence, in the degree to which their as-if costs are normative. Because it is difficult for outside observers to determine the share of as-if costs that are welfare-relevant, normative ambiguity is likely to arise in most settings in which default effects are observed.

We use our framework to characterize the optimal default in terms of three components: the distribution of (1) decision-makers' preferences over the available options; (2) as-if costs, and (3) the share of as-if costs that are normative. When these components are known, the optimal default can be determined without additional knowledge of the underlying positive model of behavior (at least within the class of models we study). Standard revealed preferences techniques can be used to recover the first two components, but not the third. Hence, our proposed approach is to identify (1) and (2) from observed choice data, and then to characterize the optimal default as a function of (3), based on the plausible range of decision-making models in the setting at hand.

We show that when as-if costs are mostly non-normative, the optimal policy induces decision-makers to make an active choice. Depending on the setting, the planner can implement this policy directly, by eliminating the presence of any default option from the decision, or indirectly, by setting as the default an option that decision-makers will find sufficiently undesirable. In contrast, when as-if costs are sufficiently normative, forcing active choice is not only undesirable, doing so actually *minimizes* social welfare. Instead, we show that a better approach in such settings is to set a default that leads relatively few decision-makers to opt out; doing so results in many people receiving an option that is close to the option they most prefer and few people incurring the (normative) costs of opting out. Optimal policy in this case resembles a rule thumb that has been proposed in the literature for setting a default by minimizing the number of opt-outs (Thaler and Sunstein, 2003); we clarify the conditions under which this rule of thumb achieves the social optimum.

Following the presentation of our main results, we relax the assumptions that individuals make no mistakes other than potentially behaving as if opt-out costs were higher than they are. When the choices made by active decision-makers are sub-optimal, the effect of defaults on welfare is complicated by the fact that those who do not opt out of the default may end up with a better option than if they were to have opted out. All else equal, we show that the presence of internalities affects the choice of optimal default through two mechanisms: (1) shifting the optimal policy so that decision-makers who follow the default achieve a welfare outcome that is higher than what they would have selected on their own, and (2) balancing this effect against the tendency of decision-makers to opt-out of the default when it is too far from the outcome they perceive to be their most-preferred. Following this line of thought, we show as well that internalities reduce the desirability of a policy that promotes active choice relative to one that induces some or all decision-makers to select the default option.

We illustrate our approach by applying it to data on employee contribution decisions to a 401(k) retirement plan. We characterize the optimal default as a function of the degree to which the as-if costs implied by employees' observed default sensitivity is normative. For the firm we study, we illustrate quantitatively our theoretical argument that one cannot identify the optimal policy without taking a stance on this question. The critical threshold in our data is whether the normative share of as-if costs is sufficiently low– i.e., less than about 8 percent of total as-if costs, which corresponds to about \$160 for the median employee. When the normative component of as-if costs is below this threshold, the optimal plan design is one that induces employees to make an active contribution decision. In contrast, when the normative component of as-if costs exceeds this threshold, the optimal policy is to set the default at the contribution rate that minimizes employee opt-outs, which, in this context, corresponds to the contribution rate that maximizes the employer match.

The optimal default in our empirical application therefore turns on the question of whether normative opt-out costs are higher or lower than the estimated threshold. We discuss additional analyses that shed light on this question. In the 401(k) setting we study, these analyses suggest that normative opt-out costs are so low that active choice is likely to be optimal. However, uncertainty as to the normative component of opt-out costs tilts the optimal policy toward minimizing opt-outs, since the desirability of active choice in this setting relies on normative ambiguity being resolved within a fairly narrow range. In addition, the presence of employee mistakes that are unrelated to the default tilts the results away from active choice and toward setting the default to a relatively high contribution rate.

Our results contribute to a growing literature on the welfare economics of default options. A closely-related paper to ours is Bernheim, Fradkin and Popov (2015) ("BFP"), which studies the optimal contribution rate default for an employer-provided 401(k) retirement plan. BFP consider a range of potential behavioral models for default effects and estimate the optimal policy within each model using employee data on 401(k) contribution decisions. They find that the optimal default is quite stable across models for the firms in their data, suggesting that normative ambiguity does not pose a major challenge to identifying the optimal default in this setting.

Our results build on BFP in several important respects. First, we use our model to clarify the settings in which policymakers should promote active choices, either by eliminating any default effect or by setting the default to a sufficiently undesirable option that all decision-makers will be induced to opt-out. In such cases, we show that determining the optimal policy is impossible without resolving normative ambiguity to at least some degree. This result has practical importance: in the 401(k) context, we find that the optimal policy plausibly takes the form of promoting active choice rather than setting the default to the contribution rate that minimizes employee opt-outs (as BFP conclude). In fact, this result is guaranteed if one imposes the same assumptions for resolving normative ambiguity that are relied on by BFP in parts of their analysis. A related advantage of our approach over the prior literature is that it allows one to quantify the dollar value of welfare-relevant opt-outs costs at which point the optimal policy switches from a default to active choice.[2]

The second way in which we build on BFP is by generalizing the level of abstraction to highlight which features of a behavioral model matter for shaping optimal policy. BFP's approach to welfare analysis is first, to select a positive model, and second, to identify the optimal policy within that model (holding fixed the positive model itself).[3] In contrast, we model default effects at a high level of generality so that our results are consistent with a broad class of positive models that might generate default effects (including the specific models that BFP consider).[4] This approach yields three payoffs: First, the lessons we draw for optimal policy emerge from general features of the problem rather than the specific models one happens to consider. For example, we show the generic reason why BFP find such stability in the optimal default across models: normative ambiguity tends not to affect the optimal default when active choice policies are ruled out and preferences over contribution rates are sufficiently well-behaved. Second, because our framework is not tied to a specific positive model, it can easily incorporate heterogeneity in the model that explains an observed default effect. Third, the generality of our approach dramatically simplifies some key features of the problem, making our framework transparent and easy to apply.

Our third contribution is considering optimal policy in settings in which decision-makers make mistakes for reasons that are unrelated to the presence of a default effect. Such mistakes are frequently invoked as a justification for using defaults to shape behavior in the first place (e.g., Thaler and Sunstein, 2008; Camerer et al., 2003). For example, in the retirement savings context, advocates often point to under-saving by employees as a justification for automatically enrolling employees into employer 401(k) savings plans. To

[2] BFP do consider one policy that has the effect of inducing active choice, which is to impose a tax on the option that is the default. However, they do not consider how the desirability of this policy varies based on the underlying model. Instead, they point out that a tax on the default can yield large welfare losses if some decision-makers remain passive no matter what. That concern applies to some, but not all, ways of encouraging active choice; for example, it would not apply to policies that simply remove the default, as in Carroll et al. (2009).

[3] BFP do vary the assumed welfare criterion within a given positive model, which generates some degree of normative ambiguity. However, because their approach to welfare analysis holds fixed the specific positive model under consideration, their results do not capture normative ambiguity stemming from differences between alternative positive models.

[4] As described in Section 1, below, the as-if cost representation we study captures all but the anchoring model considered by BFP. With respect to the anchoring model, we show in the online appendix that our key results extend to that setting as well.

our knowledge, however, the prior theoretical work on optimal defaults rules out "internalities" of this form. This omission is surprising, given that some of the proposed mechanisms by which default effects operate – such as present-bias – would also imply biased decision-making by decision-makers who do not select the default option, depending on the decision being observed.

Two other papers are also closely related to our own. Carroll et al. (2009) was the first to study when policies that force active choice are preferable to setting a default. Within a model in which present bias magnifies the cost of opting out of the default, they show that the desirability of active choice depends on the degree of time inconsistency that decision-makers exhibit.[5] We extend this result to settings in which decision-makers are sensitive to the default for reasons unrelated to present bias and to models in which active choosers make mistakes. More recently, Chesterley (2017) also studies the welfare effect of default options but focuses on a different set of policies than the ones we consider, such as policies that vary the cost of selecting a non-default option. In contrast, Chesterley's setup is not designed for studying normative ambiguity: his analysis assumes the social planner has perfect information about the extent to which decision-makers' sensitivity to the default reflects a welfare-relevant cost, and the only positive model he considers is one in which default effects are magnified because of present bias.

Another related strand of the literature attempts to disentangle various mechanisms for default effects. A few recent papers examine the implications of inattention for estimates of switching costs, often in the context of choosing a health insurance plan (Abaluck and Adams, 2017; Heiss et al., 2016). Using different strategies, both of these papers estimate that as-if switching costs are in the thousands of dollars, which is consistent with the estimates from pension plans we discuss below, but that once inattention is accounted for, the estimated as-if switching costs are only in the hundreds of dollars. Another recent paper, Blumenstock, Callen and Ghani (2017), conducts an experiment to test the mechanisms for default effects in a savings context, finding that the cognitive costs of choosing a plan appear to be the largest driver of default effects. These results can inform the normative judgments policy-makers must make, but they do not resolve normative ambiguity. For instance, believing that large estimates of as-if costs from default effects are entirely driven by inattention would imply that the planner must then determine whether and to what extent paying attention incurs a normative cost.[6] Understanding the role of normative judgments in behavioral welfare analysis is therefore complementary to understanding the mechanisms by which default effects operate.

The remainder of the paper is organized as follows: Section 1 sets out our model. Section 2 develops tools for comparing welfare between two defaults and derives a formula for the optimal policy. Section 3

[5] The focus of Carroll et al. (2009) is an empirical comparison of active choice to opt-in 401(k) plan design; they do not attempt to apply their theoretical results to data.

[6] For helpful discussions of issues relating to attention, information frictions, and policymaking, see Handel (2013) and Handel and Schwartzstein (2018).

considers policies promoting active choice. Section 4 considers the rule of thumb to select the default that minimizes the number of opt-outs. Section 5 extends the model to the case in which active decision-makers make mistakes. Section 6 illustrates our results using data on 401(k) plan contribution defaults. Section 7 concludes.

1 Model

1.1 Notation and Assumptions

Consider a population of decision-makers of measure 1. Decision-makers choose from a fixed menu X, where $x_i \in X$ denotes the option chosen by individual i. One option from the available menu is presented to decision-makers as the default, which we label by $d \in X$. Decision-makers have well-behaved preferences over the elements of X, represented by utility function $u_i(\cdot)$.[7] To facilitate welfare analysis, we assume that $u_i(\cdot)$ is cardinal and comparable across individuals. Preferences over X do not depend on the default.

Individual behavior is characterized by the solution to the following optimization problem:

$$x_i(d) = \arg\max_{x \in X} \; u_i(x) - \gamma_i \, 1_{\{x \neq d\}} \tag{1}$$

where $\gamma_i \geq 0$ for all i.[8] We will refer to γ_i as the *as-if cost* to selecting an option that is not the default. Let $x_i^* = \arg\max_{x \in X} u_i(x)$ denote the choice that maximizes (1) when $\gamma_i = 0$. We assume that decision-makers indifferent between selecting the default and a different option will select the default. Under these assumptions, behavior is given by

$$x_i(d) = \begin{cases} x_i^* & u_i(x_i^*) - u_i(d) > \gamma_i \\ d & u_i(x_i^*) - u_i(d) \leq \gamma_i \end{cases} \tag{2}$$

It will be useful to define an index of the degree to which an individual prefers opting out of the default, $a_i(d) = u_i(x_i^*) - u_i(d) - \gamma_i$. We will refer to a decision-maker with $a_i(d) > 0$ as *active* at default d and a decision-maker with $a_i(d) \leq 0$ as *passive* at default d.[9] We denote the cumulative distribution of $a_i(d)$ over

[7] Note that the differentiability of $u_i(x)$ may fail in some relevant applications. For example, in the context of default contribution rates to 401(k) plans with an employer match, $u(x)$ might exhibit an interior kink point at the contribution rate at which the match kicks in or at which the match is maximized. We discuss this further below.

[8] For simplicity, we focus on the case in which γ_i is fixed for each individual. We show in the online appendix that the intuition of our main results extends to the case in which γ_i depends on which option is the default.

[9] Note that a decision-maker who "actively" considers each option in the choice set before settling on the option that happens to be the default would still be referred to as "passive" in our terminology. In addition, we assume in equation (3), below, that such a decision-maker's welfare would be the same as that of a decision-maker who selected the default option without considering any alternatives. This assumption is innocuous for purposes of deriving the optimal default under our model, since a decision-maker who considers each option even when his most-preferred option is the default would also consider each option

the population of decision-makers at a given default by $F_{a;d}$.

Our main results will apply to the class of models generating behavior that can be represented by (2). Masatlioglu and Ok (2005) derives necessary and sufficient restrictions on behavior that a model with this representation must satisfy.[10] For example, (2) requires that if a decision-maker would choose x over y when y is the default, she must also choose x over y when x is the default. Similarly, if a decision-maker is active when x is the default and passive when y is the default, she must select y over x when choosing between the two when no default is available.

Equation (2) describes individual behavior. The following equation characterizes individual welfare in our model:

$$w_i(x,d) = u_i(x) - \pi_i\,\gamma_i\,\mathbf{1}_{\{x\neq d\}} \tag{3}$$

where $\mathbf{1}_{\{x\neq d\}}$ indicates whether the decision-maker selects an option other than the default, and $\pi_i \in [0,1]$ reflects the degree to which the as-if costs are normative – that is, the extent to which they affect the decision-maker's welfare.[11] Using the language of Kahneman, Wakker and Sarin (1997), one can think of the maximand in (1) as "decision utility" and the utility function in (3) as "experienced utility." When $\pi_i = 1$, a decision-maker's sensitivity to the default is rational. When $\pi_i = 0$, default sensitivity represents a complete mistake; the decision-maker behaves as if selecting a non-default option would reduce his welfare, but if he were to actually select a non-default option, his welfare would not decrease. When $\pi_i \in (0,1)$, the decision-maker exhibits too much sensitivity to the default; it would be rational for him to exhibit at least some sensitivity, but his behavior implies that the welfare reduction from opting out is greater than it would actually be.[12]

under alternative defaults.

[10] This representation is slightly less general than the one implied by the Masatlioglu and Ok (2005) axioms. Our main results extend to the more general representation as well (see the online appendix).

[11] One might extend our approach to settings in which $\pi_i > 1$, which may occur, for example, when opt-out costs are not fully salient.

[12] Close readers of BFP may wonder about the difference between the role of π in our model and the role of "frame-dependent weights" in theirs. The idea behind frame-dependent weights is that within certain positive models, the extent to which a decision-maker accounts for the welfare-relevant portion of opt-out costs will vary based on the choice environment (i.e., the "frame"). For example, if we assume that default effects are generated by present-bias, we would observe a different degree of the welfare-relevant opt-out costs reflected in the decision-maker's behavior depending on whether the opt-out decision was made during the same time period in which the opt-out costs were to be incurred (as opposed to during a prior period). If an observer wished to remain agnostic about whether behavior in one frame or another frame better represented decision-makers' preferences, one approach for doing so would be to use the framework developed by Bernheim and Rangel (2009) to construct bounds on welfare that reflect uncertainty about which set of observed choices should be used to infer preferences. The use of frame-dependent weights by BFP reflects this idea – it captures uncertainty about the proper perspective on welfare for a decision-maker when a given positive model implies that the decision-maker's behavior will vary based on some condition in the decision-making environment. Mechanically, frame-dependent weights enter into BFP's analysis in a similar way that π enters into our analysis, but the interpretation and use of the two concepts is quite different. In particular, frame-dependent weights reflect uncertainty in welfare stemming from uncertainty about the proper perspective on welfare within a given positive model. In contrast, we use π to capture different implications for welfare stemming from variation not only within a positive model, but also *between* alternative positive models. More importantly, in their empirical application, BFP assume a particular value for their frame-dependent weights (roughly equivalent to $\pi = 0.01$) that strikes them as ex ante reasonable; our approach is to remain agnostic about the share of as-if costs that are normative to highlight how assumptions of this type shape the welfare conclusions that emerge.

We denote a decision-maker's indirect utility by $v_i(d) \equiv w_i(x_i(d), d)$. Aggregate social welfare under default d is given by

$$W(d) \equiv \int\limits_i v_i(d)\, di.$$

An *optimal default* $d^* \in X$ is an option that yields the highest social welfare when presented as the default, $W(d^*) \geq W(d)\ \forall d \in X$.

To summarize, the decision-maker behaves as if selecting the non-default option incurs utility cost γ_i. However, selecting the non-default option in fact reduces the decision-maker's welfare by only $\pi_i \gamma_i$. Because the social welfare function incorporates γ_i only to the extent of π_i, the model generates a wedge between behavior and welfare whenever $\pi_i \neq 1$. For this reason, we label $\pi_i \gamma_i$ the *normative opt-out cost* and $(1-\pi_i)\gamma_i$ the *behavioral opt-out cost.*

1.2 Relationship to Positive Models of Default Effects

In this sub-section we briefly review alternative behavioral models that have been proposed to explain default effects and discuss the extent to which they do or do not map into our framework. The main insight is that although many behavioral models are consistent with our representation, each implies a different conclusion regarding the share of the as-if costs that are normative (π).

1.2.1 Real Opt-Out Costs

The real opt-out costs model is defined by $\pi_i = 1$. Decision-makers select from among the available options according to their preferences over the available items (u_i), while rationally accounting for the welfare-relevant costs associated with selecting an option that is not the default. These costs might include monetary costs, such as administrative fees for selecting a non-default option, or non-monetary costs such as the hassle or mental effort required to determine one's most-preferred option from the available menu. Although the latter category of costs are not present in neoclassical models, to the extent they are welfare-relevant, it is rational for decision-makers to account for them when determining whether to opt out of the default. Because $\pi_i = 1$, this positive model implies that decision utility (2) and experienced utility (3) are identical.

1.2.2 Status Quo Bias

Another proposed explanation for default effects is that decision-makers are biased towards following the status quo, and interpret the default option to be a continuation of the status quo (Masatlioglu and Ok, 2005). Decision-makers in this model follow a psychological heuristic in which they behave as if following the status quo is associated with some additional benefit $b_i \geq 0$:

$$x_i(d) = \arg\max_{x \in X} \; u_i(x) + b_i 1_{\{x=d\}} \tag{4}$$

Calling the status quo effect a "bias" suggests that this propensity to avoid deviating from the status quo option does not actually increase decision-makers' welfare:

$$w(x_i, d) = u_i(x) \tag{5}$$

The fact that b_i affects behavior but not welfare is what differentiates this positive model from the real opt-out costs model described above. It is easy to see that status quo bias maps into our framework with $\gamma_i = b_i$ and $\pi_i = 0$.

1.2.3 Endowment Effect

A related possibility is that default effects may be driven by an endowment effect, in which decision-makers perceive themselves as endowed with the default option and exhibit reluctance to exchange that endowment for other options Tversky and Kahneman (1991). Whether this additional reluctance enters into decision-makers' welfare is controversial (see Zeiler, 2017, for a discussion of this point). Behaviorally, default effects driven by the endowment effect can be modeled in the same way as default effects driven by status quo bias. When the endowment effect is fully normative, $\pi_i = 1$; when it is fully a bias, $\pi_i = 0$. It is also easy to imagine the endowment effect operates partly as a non-standard preference and partly as a bias, in which case $\pi_i \in (0, 1)$.

1.2.4 Quasi-Hyperbolic Discounting

In many cases, the as-if costs implied by observed default effects appear implausibly large. Consequently, a number of papers have considered behavioral models in which decision-makers behave as if the normative opt-out costs associated with a decision were magnified (i.e., $\pi_i < 1$). One way in which researchers have done this is by incorporating present-bias into a model of default effects (Carroll et al., 2009; Bernheim, Fradkin and Popov, 2015).

To illustrate how present bias fits into our framework, suppose that the decision-maker decides whether to opt-out from the default in the first period. In the second and all future periods, the decision-maker receives flow utility from the option she selected in the previous period, and decides again whether to opt out from the default. We assume for simplicity that opt-out costs and flow utility functions are fixed across periods; allowing individuals to realize a new, potentially lower, opt-out cost in future periods is a straightforward extension (see Carroll et al., 2009).[13]. Because opt-out costs and flow utility are fixed, the individual faces the same decision problem in each period and will make the same choice in each period. In this framework, we can therefore think of $u_i(x)$ utility for some option x received in perpetuity. As in Laibson (1997), $\delta_i \in (0,1]$ denotes the discount rate and $\beta_i \in (0,1)$ denotes the degree of present-bias. The contemporaneous cost of opting out is denoted by c_i.

Suppose first that the agent is *sophisticated,* so that she correctly anticipates her future opt-out decisions. In this case, choices are described by:

$$x_i(d) = \arg\max_{x \in X} \; \delta_i \beta_i u_i(x) - c_i 1_{\{x \neq d\}} \tag{6}$$

and welfare is described by:

$$w(x_i, d) = \delta_i \, u_i(x) - c_i 1_{\{x \neq d\}} \tag{7}$$

Normalizing these preferences shows that (6) and (7) are equivalent to (1) and (3), with $\gamma_i = \frac{c_i}{\delta_i \beta_i}$ and $\pi_i = \beta_i$.

Next suppose that the individual is *naive*, so that she may choose not to opt out today but expect to opt out at some point in the future. As in BFP, we consider the case of *partial naïveté*, with the degree of naïveté summarized by $\kappa_i \in [0,1]$. To evaluate her utility in the next period (after a negligible delay) if she opts out, the agent places weight κ_i on the case in which she decides whether to opt out in that period according to her long-run preferences ($\beta = 1$), and weight $(1 - \kappa_i)$ on the case in which she continues to be present-biased indefinitely (and thus continues to not opt-out). The perceived payoff to selecting the default is is derived by BFP and given by:

$$\beta_i \kappa_i \max\{\delta_i u_i(x^*) - c_i, \; \delta_i u_i(d)\} + \beta_i (1 - \kappa_i) \delta_i u_i(d) \tag{8}$$

As before, the agent believes that if she opts out, she receives $\beta_i \delta_i u_i(x^*) - c_i$. Comparing these two and simplifying, the agent opts out if and only if

$$u_i(x^*) - u_i(d) < \frac{1 - \beta_i \kappa_i}{\beta_i - \beta \kappa_i} \frac{c_i}{\delta_i}. \tag{9}$$

This model therefore simplifies to our costly opt-out model with $\gamma_i = \frac{1 - \beta_i \kappa_i}{\beta_i - \beta_i \kappa_i} \frac{c_i}{\delta_i}$ and $\pi_i = \frac{\beta_i - \beta_i \kappa}{1 - \beta_i \kappa}$. Note that when the agent is fully naive, i.e. $\kappa_i = 1$, the agent will procrastinate indefinitely and never opt out. In this

[13] The main difference in this extension is that there is potentially an option value to waiting for a lower cost in order to opt out in a later period.

case one would estimate empirically that the as-if costs were arbitrarily large for such an agent (or whatever fraction of such agents there are in the population), $\gamma_i \to \infty$, and, though they are never incurred, such costs would be totally irrelevant for welfare, $\pi_i = 0$.

1.2.5 Inattention

Another potential explanation for default effects is that some decision-makers neglect to make an active choice, and therefore fail to consider either the utility of the available options or the (real or perceived) opt-out costs associated with selecting the non-default option (Chetty, 2012; Goldin and Lawson, 2016). Following Masatlioglu, Nakajima and Ozbay (2012) we model inattention by supposing that decision-makers maximize utility over some subset of the available options, $\Gamma_i(X, d) \subseteq X$, where Γ_i represents what Masatlioglu, Nakajima and Ozbay refer to as an *attention filter:*

$$x_i(d) = \arg \max_{\Gamma_i(X,d)} \; u_i(x)$$

The following intuitive restriction on the possibilities for Γ_i permits us to import this model into our framework:

$$\forall i, \; \Gamma_i(X, \, d) \in \{\{d\}, \, X\}$$

In words, the individual either pays attention only to the default (passive choice) or she pays attention to the full menu (active choice).

Closing the model requires specifying a process by which $\Gamma_i(X, d)$ is determined. There are two intuitive possibilities. One is a heuristic model of attention, in which Γ_i is exogenous to the utility stakes of the decision being considered. In this model there are simply two (exogenously determined) types of agents: attentive choosers ($i \in \mathbf{A}$) and inattentive choosers ($i \notin \mathbf{A}$):

$$\Gamma_i(X, d) = \begin{cases} X, & i \in \mathbf{A} \\ \\ \{d\}, & i \notin \mathbf{A} \end{cases}$$

This behavior maps into our model with $\gamma_i \in \{0, \infty\}$ and $\pi_i = 0$.

Alternatively, the set of options to which a decision-maker is attentive may depend on the utility gain from choosing actively. Let $\tilde{\gamma}_i$ denote the perceived utility cost to making an active choice (e.g., mental effort) and let $\tilde{\pi}_i\tilde{\gamma}_i$ denote the actual utility costs to doing so. Welfare is given by:

$$w_i(X,d)=\begin{cases} u_i(x_i^*) - \tilde{\pi}_i\tilde{\gamma}_i, & \Gamma_i = X \\ u_i(d), & \Gamma_i = \{d\} \end{cases}$$

Individual i chooses to be active if the perceived utility gains from doing so exceed the associated costs:

$$\Gamma_i(X,d) = \begin{cases} X, & u_i(x^*) - u(d) > \tilde{\gamma}_i \\ \{d\}, & \text{otherwise} \end{cases}$$

It is apparent that this model is equivalent to the general model of default sensitivity laid out above, where the ambiguity over the welfare consequences of following the default is simply pushed back a level to the welfare consequences of choosing actively or passively: $\gamma_i = \tilde{\gamma}_i$ and $\pi_i = \tilde{\pi}_i$.

1.2.6 Combinations of the Above Models

In practice, default effects may be generated by combinations of the above models, in which decision-makers are more likely to select the default option partly because doing so avoids a normatively relevant cost (e.g., mental effort) and partly due to a bias or heuristic. In such cases, the as-if costs are neither fully normative nor fully behavioral, $\pi_i \in (0, 1)$. Similarly, decision-makers may be heterogeneous with respect to the decision-making model that explains the source of their default sensitivity.

1.2.7 Anchoring Effects

A possible mechanism by which defaults shape behavior is through a psychological anchoring effect, in which the default induces decision-makers to select an option closer to the default than they would otherwise choose (Tversky and Kahneman, 1974). Models of defaults as anchors cannot be represented using the opt-out cost representation that is our focus because they imply that the default potentially affects the behavior of all decision-makers, not only those who ultimately select it.[14]

Luckily, it is sometimes possible to distinguish anchoring models of defaults from opt-out cost models of default effects by investigating whether defaults induce peaks or troughs in the options near to them (Bernheim, Fradkin and Popov, 2015).[15] Although it is possible that defaults operate through anchoring effects in certain contexts, the empirical evidence reviewed in section 1.3 suggests that there are many contexts

[14] Technically, models of anchoring violate the axiom that Masatlioglu and Ok (2005) label Status Quo Independence (SQI*).

[15] Bernheim, Fradkin and Popov (2015) estimate a model incorporating anchoring and a fixed cost of opting out of the default, and note that the fit of the model improves somewhat when allowing for anchoring. However, some of this improvement in fit is mechanical, since adding anchoring adds to the model a free parameter that can explain why individuals choose the default.

in which the opt-out cost models appear to better fit the data. We nevertheless consider an extension to our framework that can incorporate anchoring effects in Online Appendix Section C. We show that our key findings regarding the desirability of active choices or minimizing opt-outs are largely unchanged in this extension to the model.

1.2.8 Defaults as Advice

Decision-makers might select the default option if they themselves are uncertain over which option is most consistent with their preferences and they believe that the planner's choice of default provides an informative signal as to which option is best for them. The optimal policy prescriptions we consider are geared towards a world in which the planner lacks ex ante information as to which option is most consistent with decision-makers' preferences, suggesting that rational (well-informed) decision-makers would not treat the default signal as having any informational content. Nonetheless, decision-makers might mistakenly construe the default as a suggestion by the planner and treat it as containing some informational content. One possibility is that decision-makers treat the suggestion as "take it or leave it" advice – i.e., they either follow the suggestion exactly or ignore it altogether, perhaps by gathering so much information on their own that the original suggestion has negligible signal value. Such a model is isomorphic to the status quo bias model when the default has no true signal value. Alternatively, decision-makers may take the suggested option into account, even if they do not accept it, and choose something closer to the default than what they otherwise would have chosen. In this case, the default affects decision-making like an anchor, where the effect of the default on a decision-maker's behavior depends on the strength of the decision-maker's prior and the perceived reliability by the decision-maker of the informational signal embodied in the choice of default.

1.3 Empirical Plausibility

In practice, it is often difficult to directly test the axiomatic foundations of particular behavioral models. With respect to models of default effects, for example, difficulties may arise because individuals have heterogeneous preferences and opt-out costs, or it may be impossible to observe the same individual choosing under alternative defaults, holding everything else fixed.

One prediction of our model that, with modest additional structure, does lend itself to testing is the idea that fewer individuals will select any given option when the default is close to that option than when the default is far from that option. Intuitively, decision-makers that prefer the option in question will be more likely to settle for the default – thus avoiding the opt-out costs – when the utility gains from selecting the non-default are relatively low. More formally, this prediction can be stated as follows:

Suppose that the menu X is ordered, and $u_i(\cdot)$ is single-peaked. Then (2) implies that for any two defaults $d^{'}$ and $d \in X$ such that $d^{'} > d$, it follows that $P(x_i(d) = x) \geq P(x_i(d^{'}) = x)$ for $x > d^{'}$, and $P(x_i(d) = x) \leq P(x_i(d^{'}) = x)$ for $x < d$.

Evidence consistent with this prediction has been documented across a range of settings, including: 401(k) contributions (e.g., Madrian and Shea 2001, Figure IIc; Choi et al., 2006, Figure 2), charitable contributions (Altmann et al., 2016); taxi ride tips (Haggag and Paci, 2014); and even thermostat temperature settings in office buildings (Brown et al., 2013). These findings support the empirical relevance of the class of behavioral models we study. Notably, the anchoring model of defaults discussed in Section 1.2.7 makes the opposite prediction, suggesting for example that we should observe $P(x_i(d) = x) < P(x_i(d') = x)$ for $x > d' > d$, at least at values of x that are sufficiently close to d'.[16]

2 Characterizing the Optimal Default

In this section we characterize the optimal default in terms of the components of the model described in Section 1.1.

Our first result highlights that the welfare achieved under a default can be decomposed between active and passive choosers as follows:

Lemma 1:

$$W(d) = E[u_i(x_i^*) - \pi_i\gamma_i \,|\, a_i(d) > 0]\ (1 - F_{a;d}(0)) + E[u_i(d) \,|\, a_i(d) \leq 0)\, F_{a;d}(0), \tag{10}$$

Lemma 1 simplifies the evaluation of welfare by showing that we can think of the welfare effect of a given default d in terms of two groups: (1) active choosers selecting x_i^* and incurring normative costs $\pi_i\gamma_i$, and (2) passive choosers selecting d.

Consider a change in the default from d_0 to d_1. From Lemma 1, it is apparent that such a change affects welfare directly for passive choosers, for whom it changes the option they select, and may in addition affect the composition of active and passive decision-makers (see Chesterley, 2017, for a discussion of this point). To study the welfare effects of this change, it will be useful to partition the population into four groups of decision-makers based on their behavior under the old default (d_0) and the new default (d_1):

[16] Note that research such as Choi et al. (2012) that reports evidence consistent with anchoring effects does not do so in the case of default options.

Group	**Behavior when default is:**		**Characterization**	**Fraction of**
	d_0	d_1		**Population**
Always Active	$a_i(d_0) > 0$	$a_i(d_1) > 0$	$u_i(x_i^*) - \max\{u_i(d_0),\ u_i(d_1)\} > \gamma_i$	$p(AA)$
Always Passive	$a_i(d_0) \leq 0$	$a_i(d_1) \leq 0$	$u_i(x_i^*) - \min\{u_i(d_0),\ u_i(d_1)\} \leq \gamma_i$	$p(PP)$
Active-to-Passive	$a_i(d_0) > 0$	$a_i(d_1) \leq 0$	$u_i(x_i^*) - u_i(d_0) > \gamma_i \geq u_i(x_i^*) - u_i(d_1)$	$p(AP)$
Passive-to-Active	$a_i(d_0) \leq 0$	$a_i(d_1) > 0$	$u_i(x_i^*) - u_i(d_1) > \gamma_i \geq u_i(x_i^*) - u_i(d_0)$	$p(PA)$

The table describes how the composition of these four groups is determined in terms of the behavioral parameters from equation (2). We denote the fraction of the population in each of these groups by $p(j)$ for $j \in \{AA,\ PP,\ PA,\ AP\}$. Intuitively, the passive-to-active group is composed of decision-makers for whom the original default is close enough to their preferred option to acquiesce to, but the new default is not, $i \in PA \implies u_i(d_0) > u_i(d_1)$. Similarly, decision-makers in the active-to-passive group are sufficiently dissatisfied with the old default to make an active choice, but content to choose passively under the new default, $i \in AP \implies u_i(d_1) > u_i(d_0)$.

The following proposition uses this decomposition to characterize the welfare effect of a change in default policy.

Proposition 1 *For any two defaults d_0, $d_1 \in X$:*

$$W(d_1) - W(d_0) = E\left[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i \,|\, PA\right] p(PA) - E\left[u_i(x^*) - u_i(d_1) - \pi_i\gamma_i \,|\, AP\right] p(AP)$$

$$+E\left[u_i(d_1) - u_i(d_0) \,|\, PP\right] p(PP) \tag{11}$$

Several features of (11) are notable. First, the always-active choosers, group AA, do not enter into the welfare effect of the default change. These individuals incur the same normative cost ($\pi_i\gamma_i$) and make the same choice (x_i^*) under both defaults. Second, for those who are passive at d_0 and active at d_1 (group PA), the change induces a utility gain from choosing actively, $u_i(x_i^*) - u_i(d_0)$, but also causes them to incur normative cost $\pi_i\gamma_i$. The first term in equation (11) reflects the change in social welfare from these individuals. The second term is the analogous contribution from individuals who are active at d_0 but not at d_1 (group PA). The third term reflects individuals who are passive under both defaults (group PP). The overall effect on this group's welfare depends on whether they (on average) prefer the new default or the original default.

One instructive special case concerns the situation when all individuals prefer the same option, $x_i^* = x^*$ for all i. Not surprisingly, the optimal policy is such settings is to set the default equal to decision-makers' most-preferred option, regardless of the $\pi_i' s$. Intuitively, complete homogeneity in preferences eliminates normative ambiguity because it eliminates the need to compare the welfare of active choosers with the welfare of passive choosers (see e.g. equation (10)); this is because no one incurs (potentially normatively relevant) opt-out costs.

Note that Proposition 1 holds regardless of the nature of the menu X – it might be discrete, continuous, or of multiple dimensionality. The next result considers situations where X is a real interval, which occurs in many applied contexts.

Proposition 2 *Let X be any interval in $\mathbb{R}$, and suppose $u_i(x)$ is everywhere differentiable for all i. If d^* represents an interior solution to the optimal default problem, the following first-order condition is satisfied:*

$$\begin{aligned} 0 = W^{'}(d^*) \quad &= \quad E[(1-\pi_i)\gamma_i \,|\, a_i(d^*) = 0,\ u_i'(d^*) < 0]\, f_{a|u'<0}(0)\, F_{u'}(0) \\ &- \quad E[(1-\pi_i)\gamma_i \,|\, a_i(d^*) = 0,\ u_i'(d^*) > 0]\, f_{a|u'>0}(0)\ (1 - F_{u'}(0)) \\ &+ \quad E\left[u'(d^*) \,|\, a_i(d^*) < 0\right] F_{a;d^*}(0) \end{aligned} \tag{12}$$

where $f_{a|u'>0}$ is the probability density function of $a_i(d^*)$ conditional on $u_i'(d^*) > 0$; $F_{u'}$ is the cumulative density function of $u_i'(d^*)$; and, as above, $F_{a;d^*}$ is the cumulative density function of $a_i(d^*)$.

As in Proposition 1, the three terms represent the welfare effects of the default change on decision-makers in the AP, PA, and PP groups. The first term represents the PA group; a decision-maker for whom $a_i(d) = 0$ and $u_i^{'}(d) < 0$ will be passive at the original default and active following a marginal increase in the default (which they prefer slightly less than the original default). Similarly, the second term represents decision-makers in the AP group, who are slightly better off after the marginal increase in the default, and therefore more willing to acquiesce to it. Decision-makers in the third group, with $a_i(d) < 0$, remain passive even after a small change in the desirability of the default.

How does the normative share of as-if costs affect the optimal default? Proposition 2 highlights that π matters for weighting the relative welfare effects of a change in the default for decision-makers in the PA and AP groups against the welfare effects for decision-makers in the PP group. When $\pi_i = 1$, the welfare effect depends only on decision-makers in the PP group, who experience a marginal change in welfare from moving to a slightly better or slightly worse default. The reason why is that decision-makers in the PA and AP groups behave as though they are indifferent between following the default and making an active choice ($a_i(d) = 0$). When $\pi_i = 1$ for decision-makers in these groups, that behavior fully reflects their welfare, and the envelope theorem implies that their welfare is not affected by a policy change that makes them active or passive.

In contrast, when $\pi_i < 1$, the welfare of the PA and AP groups will be weighted more heavily in determining the optimal default. The reason why is that decision-makers in the PA group were choosing to remain passive when their welfare would have been higher had they become active, and are better off after being induced to become active by the change in default. Conversely, those in the AP group would have higher welfare from being active, even after the change in the default induces them to become passive. The further π is from 1, the larger are these effects. In addition, although the fraction of the population in the AP and PA groups will generally be smaller than the fraction of the population in the PP group for marginal changes in the default, decision-makers in the former groups experience a *discrete* welfare change from the change in the default, whereas those in the PP group experience only a marginal change in their welfare from ending up with a slightly better or slightly worse default. We explore further how the optimal policy depends on π_i in the next two Sections.

3 Forcing Active Choices

This Section applies our framework to evaluate the desirability of policies that induce decision-makers to make active choices. In practice, such policies might take the form of (1) a "penalty default" (Ayres and Gertner, 1989) set to an option so undesirable that the vast majority of decision-makers are likely to opt out, or (2) restricting the opportunity set so that decision-makers are forced to make an active choice (e.g. Carroll et al., 2009). As an example of the former approach, one could imagine setting intestacy law – law governing inheritances in the absence of a will – so that individuals who die without leaving a will would have all of their assets taxed at a 100% rate. An example of the latter approach would be requiring new employees to make an active decision about how much to contribute to their 401(k) plans as a condition of employment.[17] For convenience, we will model both types of policies as penalty defaults, since their effects are the same under our assumptions.

Formally, we define a *penalty default* as some option $d_p \in X$ for which $a_i(d) > 0$ for all i. It is straightforward to show that whenever $u_i(d_p)$ is sufficiently low for all individuals, d_p will be a penalty default. Comparing a change in the default to a penalty default d_p from an arbitrary alternative d using Proposition 1, we have that

$$W(d_p) - W(d) = E[u_i(x^*) - u_i(d) - \pi_i \gamma_i | PA]\, p(PA) \tag{13}$$

Because individuals are never passive at d_p, only the first term of (11) matters for welfare.

[17] Another possible way to promote active choices is to reduce the costs of opting out of a default, considered by Chesterley (2017), or by taxing decision-makers who select the default option, considered by BFP.

The following proposition, which stems from (13), highlights the importance of resolving normative ambiguity when policies that promote active choice are available:

Proposition 3 Suppose that X is any menu and there exists a penalty default $d_p \in X$.

(3.1) *There exists a threshold* $\underline{\pi} \in [0, 1)$ *such that* $\pi_i \leq \underline{\pi}$ *for all* i *implies* d_p *maximizes social welfare over all* $d \in X$.

(3.2) *There exists a threshold* $\overline{\pi} \in (0, 1]$ *such that* $\pi_i \geq \overline{\pi}$ *for all* i *implies* d_p minimizes *social welfare over all* $d \in X$.

Proposition 3 shows that when forcing active choice is a feasible policy, it is never possible to identify the optimal default without taking a stance on whether or to what degree opt-out costs are normative.[18] Moreover, the stakes are high: forcing active choices can be either the best or the worst possible outcome for social welfare, depending on what π turns out to be.

To interpret (3.1), start from the benchmark case where $\pi_i = 0$ for everyone. In that case, forcing active choice results in everyone receiving the option they prefer and no one incurring any normative opt-out costs. The result in (3.1) generalizes this idea to the case where π is small but not necessarily zero.[19]

We can see from (13) that when the π_i's are large and as-if costs of opting out are normatively relevant, the active choice policy considered in (3.1) may not be desirable. The implication of (3.2) and is that requiring active choices may be *extremely* undesirable for high values of π. Note that when $\pi_i = 1$ for all i, the right-had side of (13) must be negative; this is because individuals who are passive at default d have $u_i(x^*) - u_i(d) < \gamma_i$. Such individuals reveal a preference for choosing passively. Hence, when $\pi_i = 1$ for all i, forcing active choice is not only dominated by other potential defaults that allow for some passive choice, it is dominated by *every other potential default.* The result in (3.2) generalizes the same reasoning to sufficiently high values of π that may nevertheless be less than 1.

4 Minimizing Opt-Outs

A commonly discussed rule of thumb for setting defaults, first proposed by Thaler and Sunstein (2003), is to select as the default whichever option minimizes the number of decision-makers who opt-out (i.e., who

[18] The case in which x_i^* is homogeneous is a knife's edge exception to this statement. In that case, setting $d = x^*$ achieves the highest possible social welfare for any value of π. If, however, $\pi_i = 0\ \forall i$ and there is any heterogeneity in x_i^*, forcing active choice becomes socially preferable.

[19] If some individuals remain passive at arbitrarily undesirable defaults, there is a stronger case for forcing active choice by eliminating any default option, as compared to forcing active choice by imposing a penalty default.

select any non-default option as their choice). Translated into our notation, the opt-out minimizing default, d^m, is defined as the value of d that maximizes: $W^m(d) \equiv F_{a;d}(0)$, where, as above, $F_{a;d}(\cdot)$ is the cumulative density function of $a_i(d)$.

Evaluating this expression at two possible defaults, d_0 and d_1, it is straightforward to derive that under W^m, social welfare is improved by changing the default from d_0 to d_1 if and only if $p(PA) < p(AP)$. That is, the default change must cause more decision-makers to become passive than it causes to become passive. To illustrate how this condition relates to welfare in our model, note that we may decompose (11) as:

$$\begin{aligned} W(d_1) - W(d_0) = & \qquad \underbrace{\left(p(AP) - p(PA)\right)\overline{\pi\gamma}}_{1} \\ & + \underbrace{p(AP)\, E\left[\pi_i\gamma_i - \overline{\pi\gamma}\,|\, AP\right] - p(PA)\, E\left[\pi_i\gamma_i - \overline{\pi\gamma}\,|\, PA\right]}_{2} \\ & + \underbrace{E[u_i(x^*) - u_i(d_0)|PA]\, p(PA) - E[u_i(x^*) - u_i(d_1)|AP]\, p(AP)}_{3} \\ & + \underbrace{E[u_i(d_1) - u_i(d_0)|PP]\, p(PP)}_{4} \end{aligned} \tag{14}$$

where $\overline{\pi\gamma} = E[\pi_i\gamma_i]$.

As an initial matter, note that term 1 compares $p(AP)$ and $p(PA)$ exactly as in W^m. Individuals who are active at d_1 but not d_0 (group PA) will incur opt-out costs under d_1 valued at $\pi\gamma$, which has a negative effect on their welfare. The opposite is true for the AP group, who incur costs under d_0 but not d_1. Term 1 therefore favors whichever default minimizes opt-outs. Therefore, when all of the other terms in 14 are negligible or have the same sign as the first term, the opt-out minimizing default coincides with the optimal default.

The other terms in 14 represent factors that may cause the optimal default to diverge from the default that minimizes opt-outs. Term 2 reflects the fact that even when the size of the AP and PA groups are the same, the magnitude of the normative opt-out costs of each may differ. Similarly, term 3 reflects that, aside from whatever cost they incur from being active, the AP group receives a utility gain from choosing x^* under d_1 instead of the default under d_0, and similarly for the PA with respect to d_0. The magnitude of these utility gains and losses from changes in whether decision-makers opt-out may be different between the AP and PA group. Finally, the fourth term captures how the change in the default affects welfare for the decision-makers who remain passive. Notably, the preferences of this group are completely neglected by the minimizing opt-outs rule, even though the choices of this group are directly affected by what the default is. When the preferences of group PP differ systematically from those of the PA and AP groups, the default selected by W^m may be sub-optimal because it fails to reflect the preferences of the decision-makers who remain passive under both defaults.

The relative importance of the terms in equation 14 depend on the magnitude of π and the relative sizes of the various groups. When the PP group is small, the overall welfare effect will be dominated by the other terms. If, in addition, $\pi_i = 1$, we can conclude that the net welfare effect for the PA group is negative, i.e., $E[u_i(x^*) - u_i(d_0) - \gamma_i | PA] < 0$. The opposite is true for the AP group, for whom we will have a net positive welfare effect when $\pi = 1$. Thus, when π is large, the size of the PP group is small, and when the magnitude of the welfare effects on the AP and PA groups is similar, the minimizing opt-out rule will tend to approximate the optimal default. In contrast, when π is small, or when the PP group is large and tends to prefer defaults that induce many decision-makers to opt-out, the minimizing opt-outs rule of thumb may perform poorly.

The following proposition provides sufficient conditions under which minimizing opt-outs yields the optimal default:

Proposition 4 *Suppose that $X = [x_{min}, x_{max}] \subseteq \mathbb{R}$ and that:*

(A4.1) *As-if costs γ_i are distributed independently of x_i^*.*

(A4.2) *Preferences are given by $u_i(x) = u(x - x_i^*)$ for some map $u : \mathbb{R} \to \mathbb{R}$, with $u'(0) = 0$, $u'' < 0$ and $u(c) = u(-c)$ for any c.*

(A4.3) *x_i^* follows a single-peaked and symmetric distribution about some mode x^m.*

Under these conditions, there exists a threshold $\overline{\pi} \in (0, 1]$ such that $\pi_i \geq \overline{\pi}$ for all i implies that the optimal default is the default that minimizes opt-outs.

Proposition 4 provides conditions under which minimizing opt-outs yields the optimal policy. Loosely speaking, these conditions occur when as-if costs are sufficiently normative, the distributions of the underlying behavioral parameters are independent, and decision-makers' preferences are well-behaved. We can understand the sufficient conditions in terms of their implications for the various terms in Equation (14). In particular, (A4.1) rules out a relationship between as-if costs γ_i and preferences that could cause the sign of term 2 in Equation (14) to have the opposite sign of term 1. Next, (A4.2) makes the comparison of the utility differences in the last two terms of Equation (14) straightforward, as all heterogeneity in $u_i(\cdot)$ derives from heterogeneity in the distribution of optimal choices x_i^*. Third, (A4.3) rules out features of the distribution of x_i^* that could pull the optimal default away from the opt-out-minimizing default via the third and fourth terms in Equation (14). For example, suppose x_i^* were distributed according to a single peaked distribution around some x^m, but with an extra point mass at $x' < x^m$. In that case, it would be possible that opt-outs were minimized at x_m, but that switching the default from x^m to some $d' < x^m$ would increase welfare by giving the point-mass of individuals (who are assumed to be passive under $d = x^m$ and d') an option closer to their preferred option x'. Together, (A4.2) and (A4.3) guarantee that the effect of a change in the default in opt-outs among the PA and AP groups in (14) is a strong signal about the change in welfare of the PP group. The symmetry assumptions in (A4.2) and (A4.3) ensure that, when $d = x^m$, (1) the effect of a marginal change in the default on the AP and PA groups cancel each other out, and (2) for every member of the PP group made better off by a marginal change in the default, there is another member of that group made worse off by the same amount. These cancellations ensure that the first-order condition for x^m to be a local optimum is satisfied. Given these assumptions, all that remains is to examine when $d = d^m$ is indeed the global optimum, which is guaranteed for sufficiently high π.[20]

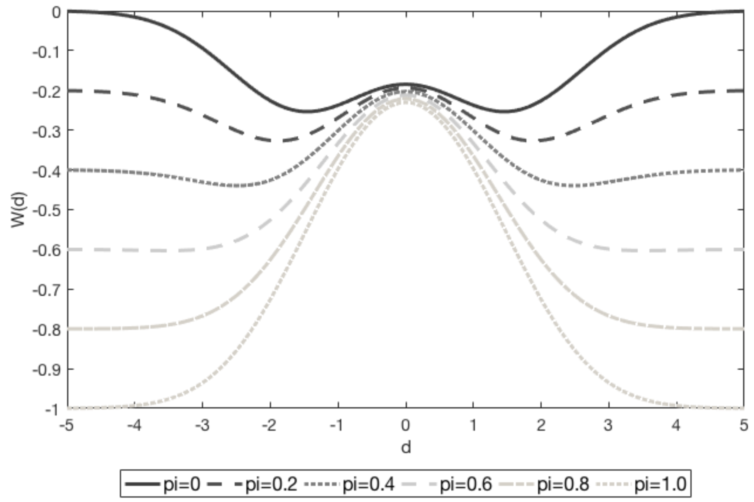
Together, Propositions 3 and 4 imply that, under the regularity conditions suggested by (A3.1) - (A3.3), the optimal policy rule is relatively simple. When π is sufficiently large, the social planner should minimize opt-outs. When π is sufficiently small and a penalty default is available, the social planner should force active choices. For intermediate values of π, other policies may be optimal. When the regularity conditions in (A3.1)-(A3.3) are not satisfied, minimizing opt-outs may not be the optimal policy for large π, but one can use the expression in Equation 14 to correct for asymmetries in $u(\cdot)$ or the distribution of x^*, or a correlation between as-if costs γ_i and optimal choices x_i^*.

Figure (1) plots social welfare for a stylized model that satisfies (A4.1)-(A4.3). To fill out the model, we assume that π_i is uniform across decision-makers, x_i^* follows a Gaussian distribution in the population, and $u(x - x_i^*)$ is quadratic – i.e., $u(x - x_i^*) = -\alpha(x - x_i^*)^2$ for suitably chosen $\alpha > 0$. To interpret the figure, recall that forcing active choice is equivalent in the model to selecting a default sufficiently extreme that all decision-makers choose to opt out, and, because we plot $W(d)$ assuming that $x^m = 0$, setting a default of zero will minimize opt-outs. The figure show shows that as π varies, the default that minimizes opt-outs remains a local optimum; the feature that varies with π is the relative attractiveness of forcing active choice. As suggested by the figure, the optimal policy in this stylized setting takes the form of a threshold rule around some threshold $\overline{\pi} \approx 0.2$. When $\pi > \overline{\pi}$, setting the default to minimize opt-outs is optimal. Instead, when $\pi < \overline{\pi}$, the optimal policy is to force active choice.

Our results in this section also shed additional light on previous results from the literature. Specifically, Carroll et al. (2009) consider the optimal policy within a model of default effects similar to the one we describe in Section 1.2.4, where present bias magnifies opt-out costs relative to true opt-out costs and the individual is a sophisticated quasi-hyperbolic discounter according to some (homogeneous) factor β. The authors add some additional structure to the model, namely a uniform distribution of costs over some finite

[20] BFP show that a distinct factor that tends to favor minimizing opt-outs when as if costs are normative is the presence of agglomerations in the distribution of x^*, such as might occur in a 401(k) savings context with an employer match (see BFP Theorem 2).

Figure 1: Social Welfare with Quadratic Preferences and Gaussian x_i^*



Note: This figure plots the function $W(d)$ for a simulated model in which $u(x - x_i^*) = -\alpha(x - x_i)^2$ with $\alpha = 0.25$, x_i^* follows a Gaussian distribution with mean zero and standard deviation one, and $\gamma_i = 1$ for all individuals. We assume π_i is homogeneous across decision-makers, and we plot $W(d)$ for several values of π. Setting $d = E[x_i^*] = 0$ will minimize opt-outs in this model, and setting an extreme default will force active choices. The simulation thus shows that the optimal policy follows a threshold rule over π, so that minimizing opt-outs is optimal for high values of π, and forcing active choices is optimal for low values of π.

interval (c.f. our Assumption A3.1), quadratic loss preferences (c.f. our assumption A3.2), and a uniform density of optimal choices x_i^* over some finite interval (c.f. our assumption A3.3). Within this model, they show that active choices are optimal when (1) β is sufficiently low and (2) optimal choices (x_i^*) are sufficiently heterogeneous. Conversely, when (1) β is sufficiently high and (2) preferences are less heterogeneous, the optimal choice will tend to be a "center default" that minimizes opt-outs (c.f. x^m in Proposition 4). Recall that in the sophisticated present bias model with the long-run view of welfare, $\pi = \beta$. Our Proposition 3 therefore illuminates the generic reason why active choices are optimal when β is low: *these are the situations when as-if costs are deemed normatively irrelevant.* Similarly, our Proposition 4 illuminates the generic reason why minimizing opt-outs is optimal when β is close to 1.

5 Mistaken Active Choices

Thus far we have restricted our focus to optimization frictions that arise because a default is present. Specifically, our analysis allows decision-makers to err with respect to the decision of whether to opt-out of a default, but assumes that those decision-makers who do opt out go on to choose optimally from the available options. However, in some settings, policymakers may believe that choices are distorted by biases unrelated to defaults. For example, in the retirement pension participation decision, present bias may cause employees to under-save even when they make active choices (see the introduction).

To set up this extension, we first assume that individual behavior is described as above by Equation (1). Hence, whether an individual is active or passive for a given default is the same as before. Now, however, we assume that individual welfare is given by

$$w_i(x) = u_i(x) + m_i(x) + \pi_i \gamma_i \{x \neq d\}, \tag{15}$$

where $m_i(x)$ is the internality imposed on the individual by his or her choice of x, i.e., the component of the welfare effect of x that is not taken into account by the decision-maker. Here, $a_i(d) > 0 \implies x_i(d) = x_i^a$. The active choice x_i^a maximizes $u_i(x)$ but not $u_i(x) + m_i(x)$ due to the internality.

For simplicity, we will focus on the case in which X is a real interval, and assume that both $u_i(x)$ and $m_i(x)$ are differentiable. In the pension contributions example, for an active chooser who under-saves, we would have $u_i'(x_i^a) = 0$ and $m_i'(x_i^a) > 0$. In this case, adopting a default that increases savings for that active chooser would increase his or her welfare on the margin, holding all else constant. Indirect utility and social welfare are defined as above.

We will describe how the optimal policy changes with the addition of internalities, first in general and then under some restrictions that provide additional intuition. Let $W_0^{'}(d)$ be the effect of a marginal change in the default on welfare in our original model, as derived in Proposition 2. With internalities, the analogue to this expression is given by:

$$\begin{array}{rcl} W^{'}(d) & = & W_0^{'}(d) + E\left[m_i(x_i^a) - m_i(d)\,|\,PA\right]\ P(PA) \\ & - & E\left[m_i(x_i^a) - m_i(d)\,|\,AP\right]\ P(AP) \\ & + & E\left[m_i'(d)\,|\,PP\right]\ P(PP). \end{array} \tag{16}$$

There are two changes in this expression relative to the one in Proposition 2. First, the PA and AP groups experience a discrete change in the internality from becoming active or becoming passive as the default changes. Second, the always-passive (PP) group in the third term of Equations (16) experience an additional marginal welfare effect, m_i', from the change in the default. As before, the welfare of always-active choosers does not enter into the evaluation of the welfare effect of a change in the default.

In order to compare optimal policy with and without internalities, it is instructive to place additional simplifying restrictions on $m_i(x)$. The following proposition illustrates how the presence of internalities affects the determination of the optimal default derived above:

Proposition 5 *In the model with internalities, suppose that*

(A5.1) *Preferences determining behavior are given by* $u_i(x) = u(x - x_i^a)$*, where* $u(x - x_i^a)$ *is quadratic and* $u'(0) = 0$

(A5.2) *Normative preferences are given by* $u_i(x) + m_i(x) = u(x - x_i^*)$ *for the same map* $u(\cdot)$

(A5.3) *The error in active choice* $x_i^a - x_i^*$ *is independent of* x_i^a *and* γ_i

Then the marginal social welfare effect of a change in the default is approximately given by $W_0^{'}(d) + \mu\, X^{'}(d)$*, where* $W_0(d)$ *denotes social welfare without internalities (see Equation (12)),* $\mu = E[m_i'(x_i(d))]$*, and* $X(d) = E[x_i(d)]$*.*

Proposition 5 highlights that optimal policy considerations here balance the concerns of the previous model, summarized by $W_0^{'}(d)$, with a new goal, which is to correct the internality generated by the decisions of the active choosers. For example, if $\mu > 0$ represents the average degree of under-saving among a population of decision-makers, the optimal savings contribution default would induce more saving than when the social planner assumed no internalities were present. The larger the mean marginal internality, μ, the further the deviation from the no-internality optimum.

In addition, it is straightforward to show that

$$\frac{\partial E[x_i(d)]}{\partial d} = E\left[x_i^a - d|PA\right]\ P(PA) + E\left[d - x_i^a|AP\right]\ P(AP) + P(PP). \tag{17}$$

Thus, the effect of a change in the default on total activity (x) has two components. First, both of the marginally active groups reduce their activity discretely – recall that $x_i^a < d$ for the PA group, and $x_i^a > d$ for the AP group. Second, the always-passive group increases their activity by a marginal amount.

To understand how optimal policy differs in this model relative to the model without internalities, suppose we initially have a default $\hat{d}$ that is optimal when no internalities are present, so $W_0^{'}(\hat{d}) = 0$. If, as in the under-saving example, $\mu > 0$, then a deviation from this default in whichever direction *increases* $X(d)$ would constitute an improvement in social welfare. Importantly, whether total activity increases with an increase in the default or with a decrease in the default is an empirical question. For example, one might assume that the presence of positive internalities from saving would lead the planner to prefer a higher default savings rate. Our results in this section show, however, that this intuition is only correct when there are relatively few marginally active choosers. In the case where total saving is relatively unaffected by the default, which may in fact be approximately correct in the oft-studied 401(k) setting (see e.g. Choi et al., 2004; Beshears et al., 2017), the presence of internalities from under-saving is irrelevant for the optimal default.

The assumptions under which Proposition 5 holds are instructive but not guaranteed. Assumptions (A5.1) and (A5.2) make the problem more tractable by ensuring that the internality is approximately linear.

Assumption (5.3) ensures that the marginal internality $\mu_i \equiv m_i'(x)$ is independent of other structural parameters governing individual behavior (though not necessarily of the *other* parameter summarizing mistakes, π_i). Relaxing (A5.3) but maintaining (A5.1) and (A5.2), it is straightforward to derive that:

$$\begin{aligned} W^{'}(d) &= W_0^{'}(d) + \mu X'(d) \\ &+ E[(\mu_i - \mu)(x_i^a - d)|PA]P(PA) \\ &- E[(\mu_i - \mu)(x_i^a - d)|AP]P(AP) \\ &+ E[\mu_i - \mu|PP]P(PP) \end{aligned} \tag{18}$$

Equation (18) shows how to modify the expression in Proposition 5 to account for the fact that the mean marginal internality may be different across the three groups of decision-makers we are interested in. For example, individuals with high values of γ_i, so that they are thus particularly sensitive to defaults, might be particularly bad under-savers.[21] In this case, the third term of Equation (18) would become more important, as the PP group have higher values of γ_i than the other groups. Such a modification would make higher defaults more attractive on the margin, relative to Proposition 5. Alternatively, we might suppose that individuals with low values of x_i^a are especially bad under-savers. As individuals with low x_i^a tend to be in the PA group, this modification would make the first term in Equation (18) larger. The fact that an increase in the default decreases the saving of the PA group would matter more for welfare than suggested by Proposition 5, which would make an increase in the default less attractive on the margin.

In addition to affecting the optimal default, the presence of internalities is also relevant for assessing the desirability of forcing active choice. As in Section 3, consider the comparison between a penalty default d^p (under which all individuals choose actively) and a generic default d (under which at least some individuals choose passively). Adapting Equation (13) to the case of internalities yields:

$$W(d^p) - W(d_0) = E\left[u_i(x^a) - u_i(d) - \pi_i\gamma_i + m_i(x_i^a) - m_i(d) \,|\, PA\right] \, p(PA) \tag{19}$$

Unlike in Proposition 3.1, it is no longer the case that a sufficiently small value of π guarantees that d^p is the optimal default. To see why, suppose that $\pi_i = 0$ for all individuals. As before, we know that $u_i(x_i^a) > u_i(d)$. However, it may be the case that $m_i(x_i^a) < m_i(d)$. When the difference $m_i(d) - m_i(x_i^a)$ is sufficiently large for enough individuals, d will be a better default than d^p. In the under-saving example, it is possible – though not assured – that a default under which some individuals choose passively will increase total saving relative to an active choice regime to such an extent that the active choice regime is not optimal,

[21] The notion that passive savers are more likely to under-save is discussed informally in Chetty et al. (2014). More generally, see Camerer et al. (2003).

even when as-if costs are completely irrelevant for welfare.

To summarize, incorporating mistakes by active choosers into the model means that the planner may be able to raise social welfare by choosing a default that makes those mistakes less likely to occur, and reduces the benefits of penalty defaults that cause decision-makers to choose actively. Uncertainty by the planner over the distribution of $m_i(x)$ thus creates a new difficulty for determining the optimal default that parallels the normative ambiguity caused by uncertainty over π.[22]

6 Empirical Illustration

This section illustrates our results using data on 401(k) plan contribution decisions. We choose to focus on this setting for two reasons. The first is that it is a setting in which defaults have been shown to affect behavior and in which the choice of default is of significant practical importance. The second is that it has been the focus of a recent and influential literature on optimal default policy; holding the setting constant in our analysis relative to this prior literature helps clarify the value added by our approach.

To preview our results, we draw two substantive contributions from this analysis. First, we generalize the result from BFP that the uncertainty over optimal defaults is small when the range of policies considered does not include policies that promote active choice. Specifically, we estimate the mapping between values of π and the optimal default. This allows us to conclude that the optimal policy BFP identifies applies not only for the illustrative models they consider, but rather for all behavioral models within a more general class (i.e., any model that is consistent with the opt-out cost representation). Our second contribution is to show that, in contrast to the results in BFP, normative ambiguity *does* generate meaningful uncertainty as to optimal policy once the policy space is expanded to include policies that promote active choice. When the as-if costs associated with default effects are mostly irrelevant from a welfare perspective, the optimal policy is to adopt a penalty default (e.g., setting a very high default contribution rate) or to require active choice as a condition of employment. In contrast, these policies are dominated when even a modest fraction of the observed as-if opt-out costs are normatively relevant. Finally, we discuss additional analyses that help to resolve the normative ambiguity over π, and we discuss how the planner's uncertainty over π matters for welfare. The illustration also highlights two of the chief benefits of our approach, namely the simplicity with which it can be applied and the transparency between our assumptions and the welfare conclusions that emerge.

[22] In some cases it is possible to identify potential internalities by varying framing, just as one can identify candidate values of π_i by varying framing. For example, with under-saving due to present bias, one can infer the size of the internality by varying the timing of (active) decisions relative to payoffs to identify the present bias parameter (β_i). Even in this example, however, using such a technique to infer the size of the internality requires adopting the long-run view of welfare, which is a normative judgment (Bernheim, 2009). See Lockwood and Taubisnky (2017) for more examples.

Setup The data we use consists of 401(k) contribution rates for newly eligible employees of several firms first analyzed by Choi et al. (2004, 2006) and Beshears et al. (2008).[23] We describe the relevant features of these data here, and refer readers to the earlier studies for additional detail. The first step in our approach is to estimate the distribution of parameters in the opt-out cost representation of behavior. To do so, we follow BFP and rely on a fitted structural as-if costs model of behavior to contribution rate data for each employer.[24] This model assumes preferences over individual contribution rates take the following form;

$$u_i(x) = \rho_i \ln(x + M(x) + \alpha) + \ln(z) \tag{20}$$

where ρ_i and α are preference parameters governing the overall preference for contributing and the price sensitivity of contributions respectively, $M(x)$ is the employer match as a function of x, and $z = 1 - (1-t)x$ is residual income pinned down by the budget constraint. We assume a marginal tax rate t of 20 percent. The firm matches 50 percent of employee contributions up to 6 percent of earnings, so we have $M(x) = 0.5x$ for $x \leq 0.06$, and $M(x) = 0.5 * 0.06$ for $x > 0.06$.

Estimating this model structurally requires assumptions about the distribution of ρ_i and γ_i. First, we assume these two variables are independent.[25] Second, we assume that γ follows an exponential distribution, modified to have a point mass λ_1 at zero, with cdf

$$\Phi(\gamma) = \begin{cases} \lambda_1 + (1-\lambda_1)(1 - e^{-\lambda_2 \gamma}), & \gamma \geq 0 \\ 0 & \gamma < 0 \end{cases} \tag{21}$$

Thus, a fraction λ_1 of individuals are assumed to have zero as-if costs and choose actively under all defaults. Third, we assume ρ_i follows a censored normal distribution: $\rho_i = \max\{0, \tilde{\rho}\}$ where $\tilde{\rho}$ is distributed normally with mean μ_ρ and variance σ^2.

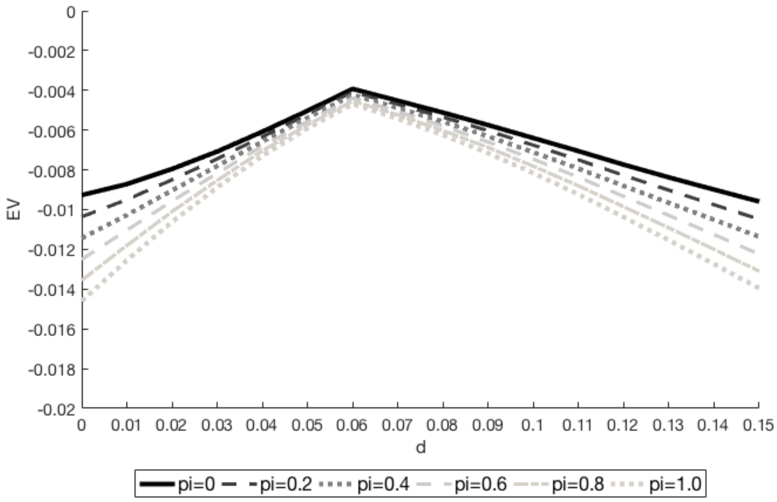
The cap on the employer match creates a large kink in the budget constraint at 6 percent of earnings, which induces bunching in the optimal contribution rate at 6 percent. The degree of bunching implicitly identifies the price sensitivity preference parameter, α. The model also predicts bunching at the corner solutions of 0 and the maximum contribution rate of 15 percent. Finally, we evaluate $W(d)$ using equivalent variation relative to a benchmark in which all individuals receive their most-preferred option x_i^* without incurring any

[23] For simplicity, we will present results for a single firm only (the one labeled "Company 3" in BFP). The results for the other firms are very similar and are presented in the online appendix.

[24] Specifically, we use the estimated structural model from the BFP publicly available replication files, as the underlying contribution rate data are proprietary and not available to us.

[25] This assumption allows us to extrapolate from the decisions of active choosers at some defaults, whose decisions are informative about ρ, to the population of decision-makers. As the active decision-makers will tend to have low γ_i, covariance between ρ_i and γ_i poses a significant challenges to identifying the preferences of individuals choose passively Goldin and Reck (2018).

Figure 2: Equivalent variation for different default contributions, by the welfare relevance of as-if costs

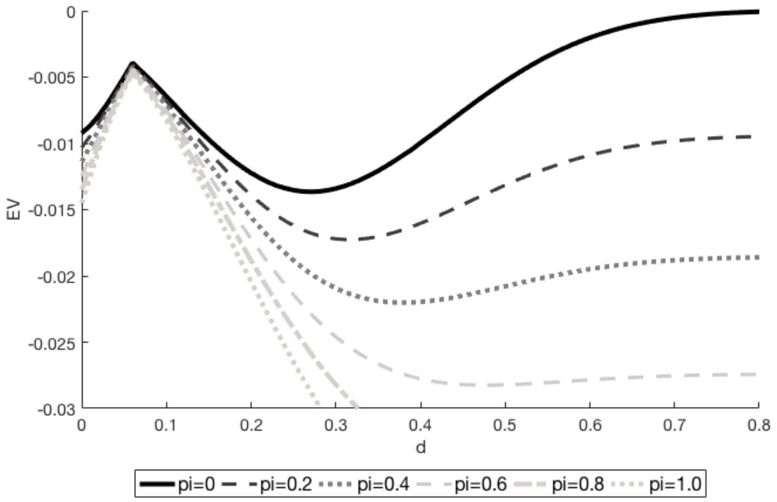


Note: This figure depicts welfare in equivalent variation units for five values of π (pi) ranging from zero, in which as-if costs are not at all welfare-relevant, to one, in which costs are fully welfare relevant, for Company 3 in BFP. Results for other companies are similar. We observe that regardless of the value of π, welfare is maximized by setting the default contribution at 6 percent of earnings, where employer matching contributions are maximized.

costs of opting out of a default, so that indirect utility is $v_i(d) = u_i(x_i(d)) - \pi_i\gamma 1\{x_i(d) \neq d\} - u_i(x_i^*)$. As the units of x_i are in percentages of annual salaries contributed to a 401(k) plan, the units of welfare thus correspond to the percentage of annual salary that would make individuals receiving x_i^* (without any costs) willing to switch to a default d. Because welfare is lower under a given default d than under the benchmark, equivalent variation is typically negative. Throughout the analysis, we assume a uniform value of π_i for the population, denoted π for simplicity. The calculation of welfare for any distribution of π_i, including those in which it covaries with other heterogeneous parameters, is straightforward.

Results Our first analysis is to solve for the optimal default as a function of the normative relevance of the as-if costs, π. Figure 2 depicts equivalent variation for alternative default contribution rates between 0 and 15 percent of earnings, for values of π ranging from zero to one. We find that regardless of π, the optimal policy is to set a default of 6 percent of earnings, which is where, for all three firms, employer matching contributions are maximized. This analysis generalizes the main finding in BFP to any positive model of default effects consistent with the opt-out cost representation, and to any view of welfare within such a model. We next extend the analysis to consider policies that promote active choice. We add this policy choice by extending the space of contribution rate defaults we consider. Proposition 3 suggests that extending the set of feasible policies in this way will introduce uncertainty into the optimal policy. Indeed, extending the policy space from that considered in Figure 2 yields exactly this result. As shown in Figure

Figure 3: Equivalent variation for different default contributions for an expanded menu of contribution rates



Note: This figure depicts welfare in equivalent variation units for five values of π (pi) ranging from zero, in which as-if costs are not at all welfare-relevant, to one, in which costs are fully welfare relevant, for "Company 3" in BFP. Results for the other companies are similar. We extend the range of possible options all the way to 80 percent. We observe that, unlike in Figure 1, normative ambiguity influences the optimal default. This occurs because the extreme defaults effectively force active choices for the full population, which is desirable for sufficiently low π.

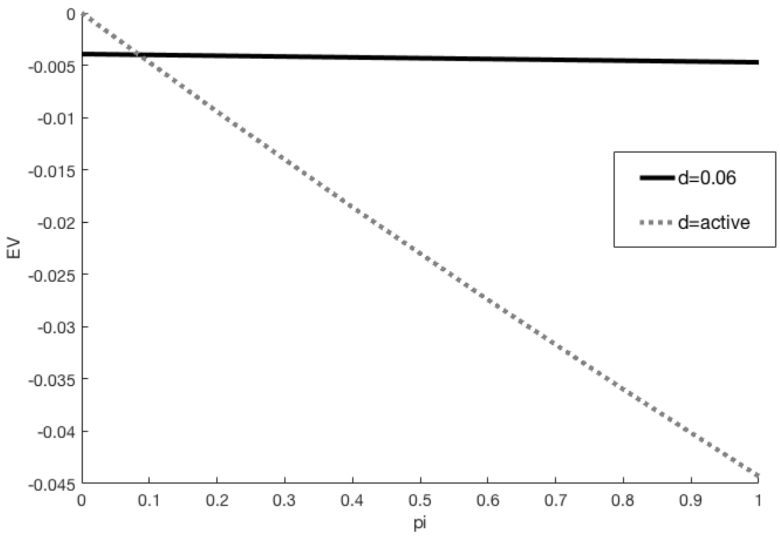
3, extremely high defaults dominate when π is low, but the 6 percent default dominates when π is moderate or large. As described above, the intuition for this result is that high defaults prompt many decision-makers to make an active choice, and when π is sufficiently low, the welfare cost of making this choice is low as well.[26] There is a strong qualitative similarity between the stylized model in Figure 1 and the estimated model in Figure 3; the main difference between these is the non-differentiable spike in the latter at the 6 percent default, which is caused by the kink in the budget constraint from employer matching contributions.

Figure 4 compares welfare under the 6 percent default to an active choice regime, for values of π ranging from zero to one. First, when $\pi = 0$, the active choice regime leads to exactly the same outcome as our benchmark in which all individuals costlessly receive x_i^*, so its equivalent variation is zero. Consistent with Propositions 3 and 4, we find that at higher values of π, the 6 percent, employer-contribution-maximizing default dominates the active choice default. The 6 percent default is also the default that minimizes opt-outs.[27] The optimal policy thus takes the form of a threshold rule: active choices dominate below the

[26] A similar phenomenon is evident in Figures A.7 to A.12 in the Online Appendix of BFP, who do not focus on it because they state that the extreme contribution rates exceed statutory limits on 401(k) contributions. However, setting an extremely high default contribution rate is not the only way to force employees to make active choices – the employer could simply require it as a condition of employment, as in Carroll et al. (2009). In addition, legal constraints in this policy area may be modified, as illustrated by the Pension Protection Act of 2006, which allowed employers to automatically enroll employees in 401(k) plans.

[27] The conditions imposed on preferences by the structural model are similar to the regularity conditions laid out in Proposition

Figure 4: Welfare Under Active Choice versus Minimizing Opt-Outs in 401(k) Plans



Note: This figure compares welfare under the 6 percent contribution default (d=0.06) with welfare under an active choice regime (d=active) for Company 3 in BFP. Results for other companies are similar. We observe that at low values of π, the active choice regime leads to higher welfare.

threshold, the 6 percent default dominates above the threshold. The threshold below which active choice dominates is about $\pi = 0.08$ for this firm.

Discussion What is the optimal policy with respect to default 401(k) contribution rates? As described in the prior section, the answer depends on the degree to which employees' observed default sensitivity reflects costs that are normative. Recall that γ reflects the magnitude of costs that would be required to rationalize employees' observed sensitivity to the default. For an employee with the median salary in the data ($40,000), the estimated distribution of γ_i for the 60 percent of employees estimated to have strictly positive costs has a mean of $3,386. The mean cost for *all* employees is about 5.07 percent of their salary, or $2,028 at the median salary. Ten percent of all employees are estimated to have an as-if cost that exceeds 15 percent of their salary, i.e. $6,000 for an employee with median salary. Because the threshold value of π for which active choice is optimal is $\pi = 0.08$, determining optimal policy in this setting requires determining whether the mean reduction in welfare that an employee would incur by opting out of the default is at least 8% of $2,028, or $162, which corresponds to about 0.4 percent of the median employee salary. When opting out of the default reduces welfare by at least $162 on average, the optimal policy is to set the default to the contribution rate that minimizes opt-outs, which here corresponds to the contribution rate that maximizes

4, notably including the independence of the determinant of optimal savings rates and as-if opt-out costs. The main difference derives from the presence of a large kink in the budget constraint due to the employer match, and the other corner solutions at zero and the maximum contribution.

the employer match, or 6 percent. In contrast, if the normative component of opt-out costs is below $162 on average, the optimal policy takes the form of setting the default to an extremely high contribution rate (perhaps as high as 80%), or, more realistically, requiring employees to make an active choice about how much they wish to contribute.[28]

Identifying the normative component of opt-out costs poses a methodological challenge, since, as described above, it cannot be directly inferred from observed choice data. Nonetheless, some conclusions can be drawn in cases in which other relevant information is known. For example, one might imagine surveying employees about the amount they would be willing to pay to to avoid the process of opting out of a 401(k) default and actively choosing their own contribution rate. Although stated preferences may be biased for other reasons, they may shed light on welfare when decision-makers' observed choices do not. Alternatively, an observer might estimate the normative component of opt-out costs from choices made in other contexts, which do not exhibit the same biases from default effects. For example, one might look to the price decision-makers are willing to pay to avoid filling out other forms or making other decisions of similar complexity, using data on the price of paid tax preparers or financial planners. Finally, an observer might estimate the number of non-work hours required to actively choose one's contribution rate and fill out the form, and price that time according to the employee's implied wage rate.

In the 401(k) setting we focus on, this type of analysis tends to suggest that normative component of as-if costs is smaller than 8 percent, which would imply that active choice is optimal. Suppose that employees value their time at $19 per hour, which is the equivalent hourly wage rate of an employee in our sample with median salary and a 40-hour work week. The threshold value of $162 would imply that, on average, the process of opting out and making an active choice would need to take more than 8 hours of the employee's time in order for the 6 percent default to dominate an active choice policy. This amount of time is on the high side of what most casual observers would likely consider plausible. The actual process of filling out forms and selecting an option typically takes less than an hour, though of course researching the available options to determine which one will be best takes more time. Relatedly, given the cognitive difficulty of making pension choices, the individual may value the time spent completing it at a higher rate than the $19/hour benchmark implied by wage rates.[29] Still, if we suppose the process takes 2 hours, employees would need to

[28] Our focus is on the optimal default option but we acknowledge that other policy variables may have at least as large an effect on participant welfare, such as the match rate. Relatedly, given the cost of employer matching contributions, an employer's goal may not be solely to maximize employees private welfare (Bubb and Warren, 2018). A higher default could increase employer's matching contributions, and the cost of this might lead employers to either 1) set a default below the level that would otherwise be welfare-maximizing given the match, and/or 2) set a lower match rate under a high default contribution than the match rates we observe at relatively low default contributions. Such issues would not be present for national automatic pension enrollment policies like those recently adopted in the United Kingdom, where the employee's pension contribution has no direct effect on the employer's contribution.

[29] In addition, this calculation uses the employee's average wage and is therefore likely to underestimate the employee's reservation wage for marginal hours worked. On the other hand, it uses the pre-tax wage rate, which likely overestimates the true opportunity cost of the employee's time.

value the time it takes to opt out at about \$80/hour, which seems high for an individual making \$40,000 per year. These conclusions match the intuition of others as well – BFP calibrate some of their models using parameters that effectively impose a value of $\pi = 0.01$.

Our results also imply, however that while active choice may be optimal under plausible restrictions on normative costs, it is in a sense a riskier policy than minimizing opt-outs. For company 3, Figure 4 shows that minimizing opt-outs leads to equivalent variation ranging from 0.4 to 0.5 percent of annual earnings (about \$160 to \$200 at the median earnings) relative the first-best benchmark, for values of π from zero to one. The active choice policy leads to equivalent variation ranging from 0 to 4.4 percent of annual earnings (\$0 to \$1,760 at the median earnings). One could directly incorporate this idea into optimal policy analysis by formalizing the planner's uncertainty over π and calculating expected social welfare.

Suppose the planner has some subjective probability distribution over the value of π. Given a uniform distribution over some range of π deemed plausible by the planner, we can compare the expected equivalent variation between active choices and minimizing opt-outs by integrating the difference between $W(d)$ for these two policies in Figure 4, over the relevant range of π.[30] With a uniform distribution over $[0, 1]$, the expected equivalent variation from active choice is -2.3 percent of annual earnings and the expected equivalent variation of minimizing opt-outs is -0.4 percent of earnings for company 3. The expected equivalent variation of minimizing opt-outs is much higher partly as a consequence of Propositions 3 and 4: when π is small, minimizing opt-outs is sub-optimal, but remains a local optimum; in contrast, when π is large, active choice minimizes social welfare. More fundamentally, when minimizing opt-outs, π only matters for the welfare of active choosers, who tend to have low as-if costs already; when setting an active choice policy, π matters for the welfare of all decision-makers, even those with very high as-if costs. One can of course impose restrictions on the distribution of π deemed subjectively plausible using similar arguments to those above. For example, if an observer were willing to impose that the normative component of opt-out costs was no more than 0.5% of an employee's annual salary, this would correspond to a maximum value for π of approximately 0.1. With π distributed uniformly between [0, 0.1], the expected equivalent variation from active choice is -0.24 percent of earnings, compared to -0.40 percent of earnings from setting a default contribution rate of 6%.

Adding Internalities We next examine the importance of internalities for the optimal 401(k) contribution default. If individuals are systematically biased in their assessment of how much they should save, it is possible that employees' active choices will be subject to a positive internality like the one studied in Section 5. To model the internality, we assume that for each individual, $m_i(x) = \mu\ x_i$, for some uniform constant

[30] With a subjective probability distribution that is not uniform, one naturally takes a weighted integral of this difference where the weights are the planners subjective likelihood for each value of π.

μ.[31] We can interpret μ in terms of a lump-sum-equivalent increase in salary in percentage terms. For example, when $\mu = 0.1$, the increase in the individuals' welfare of switching from contributing 0 percent of one's salary to the pension to contributing 10 percent is equivalent to a lump-sum transfer of 1 percent of the worker's salary.

The optimal default in this model is now a function of two unknown parameters: μ and π. Figure 5 describes the optimal policy. Figure 5a plots the optimal default as a function of the marginal internality μ for three values of π : 0, 0.1 and 1. Our earlier results are nested in the case where $m = 0$: the optimal policy is active choice when $\pi < 0.08$ and otherwise a 6 percent employee contribution. This threshold continues to describe optimal policy for small m. As m increases, active choice is no longer optimal for even low values of π.The primary advantage to active choice is that it allows each individual to select an individually optimal contribution rate; this advantage is eliminated when internalities are sufficiently large. As m continues to rise, the optimal default increases up to the maximum contribution of 15 percent. This occurs because when m is very large, the effect of an increase in the default on the saving of (always) passive savers dominates everything else from a welfare perspective.[32] This increase happens for any value of π, but it happens fastest when $\pi = 0$. When π is larger, opt-outs are costly, so the fact that increasing the default above 6 percent also leads to an increase in opt-outs makes higher defaults less desirable than when $\pi = 0$. Relatedly, when π is larger, one needs a smaller internality μ to make the active policy undesirable relative to the next-best option, which is always the 6 percent default over the relevant range of μ.

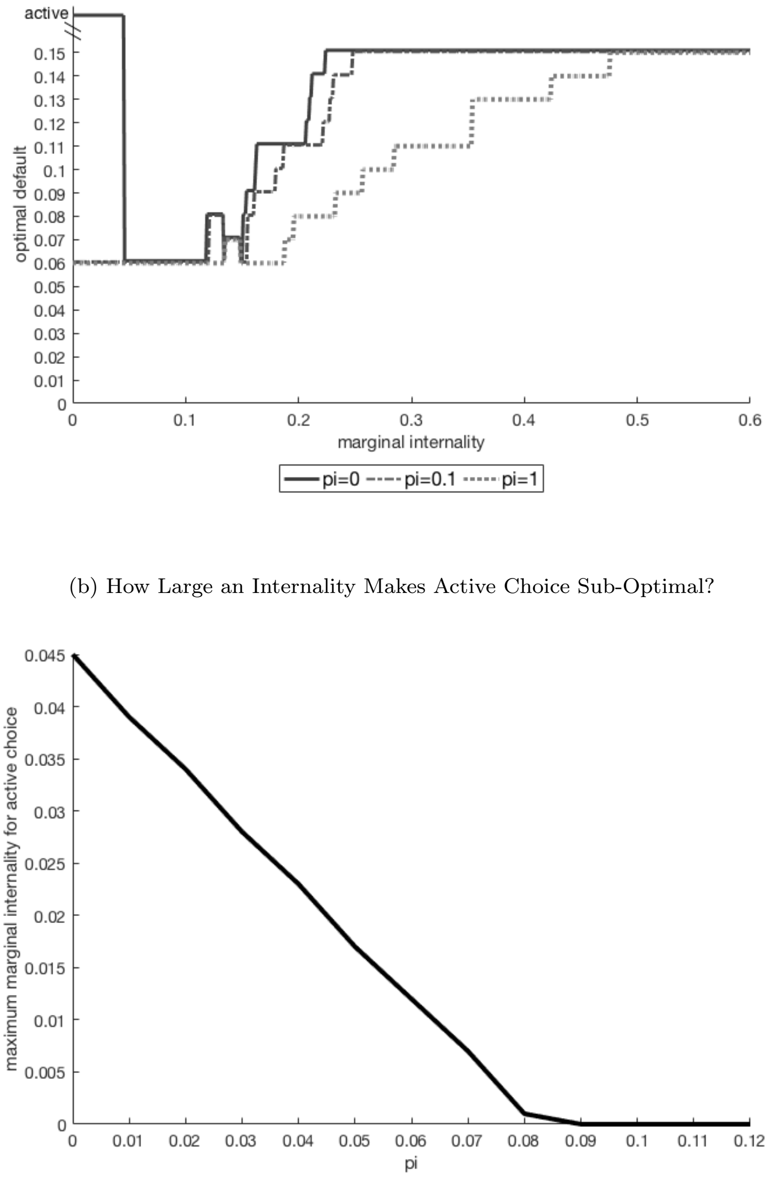
Figure 5b illustrates this phenomenon, plotting the largest possible value of μ for which the active choice policy is optimal over different values of π. When $\mu > 0.45$, the active choice policy cannot be optimal. Recall that a value of $\mu = 0.45$ would imply that a 10 percent increase in contributions (potentially including the employer match) increases the total internality by the equivalent of lump-sum transfer of 4.5 percent of earnings. Similarly, once $\pi > 0.08$, we know from our previous results that the active choice policy is always sub-optimal regardless of the internality μ. In contrast, for low values of both π and μ, active choice continues to remain the optimal policy.

We take three main conclusions from this analysis. First, the introduction of internalities introduces another caveat to the use of policies that promote active choices: if the active choosers are systematically biased toward under-saving, a default that induces higher saving (and relatively few opt-outs) can be preferable. Second, when internalities are sufficiently large, they can dominate all other considerations with respect to

[31] From the proof of Proposition 5, we can interpret $\mu_i \approx -u''(x_i^* - x_i^a)$ under (A5.1) and (A5.2). Hence, the assumption that μ is constant basically requires that the amount of under-saving under active choice, $x_i^* - x_i^a$, is relatively homogeneous among employees.

[32] This finding is not mechanical: if the model suggested that an interior default less than the maximum contribution maximized total saving, then that would be the optimal default for arbitrarily large μ. That such an internal default does not exist here is driven by the fact that the as-if costs are quite large, so that a large number of employees remain passive for even relatively high defaults.

Figure 5: The Influence of Internalities from Under-Saving on the Optimal Default



(a) Optimal Default by the Marginal Internality

(b) How Large an Internality Makes Active Choice Sub-Optimal?

Note: This figure considers the implications of internalities for the optimal default policy. Figure (5a) depicts the optimal default for variable values of μ and three values of π. We observe that active choices are dominated by the 6 percent default when μ is sufficiently large. As μ continues to increase, so that increasing saving is more and more valuable from a social welfare perspective, the optimal default increases up to the maximum contribution. Figure (5b) plots the highest value of μ under which the active choice policy is optimal, for variable values of π. As π becomes large, the active choice policy becomes less desirable independent of μ because of the importance of opt-out costs, so for larger π, a smaller internality is necessary to rule out that the active choice policy could be optimal.

setting an optimal default. Finally, if the main lesson of our previous analysis was that the optimal default contribution into 401(k) plans will depend on difficulties normative judgments, the possibility of internalities only makes the difficulties worse. Policymakers must decide not only whether to respect the validity of decision-makers' revealed opt-out costs, but also whether to respect the preferences decision-makers reveal when they make active choices.

7 Conclusion

Uncertainty over the decision-making model that generates an observed behavior is a pervasive source of difficulty in behavioral economics. Under a range of positive models of default effects, decision-makers' behavior can be described using "as-if" preferences over opt-out costs revealed by their observed choices. Revealed preference analysis can recover information about these as-if preferences, but cannot answer whether these as-if preferences accurately reflect individuals' welfare. Our analysis of the optimal default problem clarifies the conditions in which optimal policy determinations do and do not depend on the degree to which these as-if opt-out costs are normatively relevant.

In general, we find that uncertainty as to the fraction of opt-out costs that are normative – as would be associated with uncertainty over the underlying behavioral model that generates default effects – poses a serious problem for determining the optimal default. In a limited number of cases, robustness criteria like those proposed by Hansen and Sargent (2008) and Bernheim and Rangel (2009) will allow us to examine cases where optimal policy is invariant to the share of opt-out costs that are normatively relevant. However, these special cases are the exception and not the rule. They obtain only when there is little heterogeneity in decision-makers' preferences or when the policy space is restricted to rule out policies that promote active choice. Conversely, when most opt-out costs are known to be normatively relevant and preferences are sufficiently well-behaved, we show that minimizing the fraction of decision-makers' opting out of the default provides a good rule of thumb for achieving the optimal policy. Once internalities are considered, the case for active choice policies hinges on decision-makers making enough of one type of mistake (i.e., over-weighting the costs of opting out of a default) but not making too much of a different type of mistake (i.e., making a biased choice when selecting among non-default options).

When policies that promote active choice are feasible, determining the optimal default requires making a normative judgment. Two kinds of empirical evidence can help with these judgments. The first is to use various interventions attempting to reduce or enhance default effects to doing so shed light on the positive mechanisms driving default effects, as in Blumenstock, Callen and Ghani (2017). Even with evidence regarding the mechanism driving default effects, however, one must make a normative judgment determining whether that mechanism acts by imposing normative costs or by driving a wedge between choices and welfare. The second potential strategy is to gather external evidence to make an informed judgment. This approach requires assumptions about what choices in other settings, such as those that reveal the monetary value of workers' time, or those in which workers state hypothetically how much they would accept to make an active choice, tell us about the normative relevance of as-if costs. We give examples of this type of reasoning in our examination of pension plan defaults. While both of these strategies can help the planner make an informed choice, neither can resolve the problem without external judgments about which other choices are relevant and which are not.

Finally, the particular difficulty in determining optimal policy we identify is new but not unprecedented. The modern theory of optimal redistributive taxation also incorporates normative judgments, in this case judgments about the social value of equity (Mirrlees, 1971; Saez, 2001). Determining the optimal redistributive tax policy requires normative judgments that cannot be resolved via revealed preference alone. Often these judgments are parameterized, using welfare weights or the curvature of a social welfare function. Indeed the whole of classic optimal policy problems may be divided into those where revealed preferences alone yield the optimal policy via Pareto comparisons (Ramsey, 1927; Kaldor, 1939; Hicks, 1939), and those such as the optimal redistributive tax problem where normative judgments are necessary. Our results here suggest that a similar division, between welfare comparisons that require normative judgments and those that do not, will be fruitful for "behavioral" optimal policy problems.

References

Abaluck, Jason, and Abi Adams. 2017. "What do consumers consider before they choose? Identification from Asymmetric Demand Responses." Working Paper.

Altmann, Steffen, Armin Falk, Paul Heidhues, and Rajshri Jayaraman. 2016. "Defaults and donations: Evidence from a field experiment." Working Paper.

Ayres, Ian, and Robert Gertner. 1989. "Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules." *The Yale Law Journal*, 99(1): 87–130.

Bernheim, B Douglas. 2009. "Behavioral Welfare Economics." *Journal of the European Economic Association*, 7(2-3): 267–319.

Bernheim, B Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *The Quarterly Journal of Economics*, 124(1): 51–104.

Bernheim, B Douglas, Andrey Fradkin, and Igor Popov. 2015. "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review*, 105(9): 2798–2837.

Beshears, John, James J Choi, David Laibson, and Brigitte C Madrian. 2008. "The Importance of Default Options for Retirement Savings Outcomes: Evidence from the United States." In *Lessons from Pension Re-form in the Americas.* , ed. Stephen J Kay and Tapen Sinha, 59–87. Oxford University Press.

Beshears, John, Schlomo Benartzi, Richard Mason, and Katherine L Milkman. 2017. "How Do Consumers Respond When Default Options Push the Envelope?" Working Paper.

Blumenstock, Joshua, Michael Callen, and Tarek Ghani. 2017. "Why do Defaults Affect Behavior? Experimental Evidence from Afghanistan." Working Paper.

Brown, Zachary, Nick Johnstone, Ivan Haščič, Laura Vong, and Francis Barascud. 2013. "Testing the effect of defaults on the thermostat settings of OECD employees." *Energy Economics*, 39: 128–134.

Bubb, Ryan, and Patrick L Warren. 2018. "An Equilibrium Theory of Retirement Plan Design." Working Paper.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'donoghue, and Matthew Rabin. 2003. "Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism"." *University of Pennsylvania Law Review*, 1211–1254.

Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *The Quarterly Journal of Economics*, 124(4): 1639–1674.

Chesterley, Nicholas. 2017. "Defaults, Decision Costs and Welfare in Behavioural Policy Design." *Economica*, 84(333): 16–33.

Chetty, Raj. 2012. "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica*, 80: 969–1018.

Chetty, Raj, John N Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. 2014. "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark." *The Quarterly Journal of Economics*, 129(3): 1141–1219.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2004. "For Better or for Worse: Default Effects and 401 (k) Savings Behavior." In *Perspectives on the Economics of Aging.* , ed. David A Wise, 81–126. University of Chicago Press.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2006. "Saving for Retirement on the Path of Least Resistance." In *Behavioral Public Finance: Toward a New Agenda.* , ed. Edward J McCaffery and Joel Slemrod. Russell Sage Foundation.

Goldin, Jacob, and Daniel Reck. 2018. "Revealed Preference Analysis with Framing Effects."

Goldin, Jacob, and Nicholas Lawson. 2016. "Defaults, Mandates, and Taxes: Policy Design with Active and Passive Decision-Makers." *American Law and Economic Review.*

Haggag, Kareem, and Giovanni Paci. 2014. "Default Tips." *American Economic Journal: Applied Economics*, 6(3): 1–19.

Handel, Benjamin, and Joshua Schwartzstein. 2018. "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives*, 32(1): 155–178.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *The American Economic Review*, 103(7): 2643–2682.

Hansen, Lars Peter, and Thomas J Sargent. 2008. *Robustness.* Princeton university press.

Heiss, Florian, Daniel McFadden, Joahim Winter, Amelie Wupperman, and Bo Zhou. 2016. "Inattention and Switching Costs as Sources of Inertia in Medicare Part D." Working Paper.

Hicks, John R. 1939. "The Foundations of Welfare Economics." *The Economic Journal*, 49(196): 696–712.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science*, 302(5649): 1338–1339.

Kahneman, Daniel, Peter P Wakker, and Rakesh Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility." *The Quarterly Journal of Economics*, 112(2): 375–406.

Kaldor, Nicholas. 1939. "Welfare Propositions in Economics and Interpersonal Comparisons of Utility." *The Economic Journal*, 49(195): 549–552.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics*, 443–477.

Lockwood, Benjamin, and Dmitry Taubisnky. 2017. "Regressive Sin Taxes." Working Paper.

Madrian, Brigitte C, and Dennis F Shea. 2001. "The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior." *The Quarterly Journal of Economics*, 116(4): 1149–1187.

Masatlioglu, Yusufcan, and Efe A Ok. 2005. "Rational choice with status quo bias." *Journal of Economic Theory*, 121(1): 1–29.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y Ozbay. 2012. "Revealed Attention." *The American Economic Review*, 102(5): 2183–2205.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *The Review of Economic Studies*, 38(2): 175–208.

Ramsey, Frank P. 1927. "A Contribution to the Theory of Taxation." *The Economic Journal*, 37(145): 47–61.

Saez, Emmanuel. 2001. "Using Elasticicities to Derive Optimal Income Tax Rates." *The Review of Economic Studies*, 68(1): 205–229.

Thaler, Richard H, and Cass R Sunstein. 2003. "Libertarian Paternalism." *American Economic Review*, 175–179.

Thaler, Richard H, and Cass R Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science.*

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics.*

Zeiler, Kathryn. 2017. "Mistaken About Mistakes." *European Journal of Law and Economics.*

Appendix (For Online Publication Only)

A Appendix: Proofs

Lemma 1:

$$W(d) = E[u_i(x_i^*) - \pi_i\gamma_i | a_i(d) > 0]\ (1 - F_{a;d}(0)) + E[u_i(d)|a_i(d) \leq 0)\, F_{a;d}(0),$$

where $F_{a;d}(\cdot)$ denotes the cumulative density function of $a_i(d)$.

Proof: From the definition of the social welfare function we know that $W(d) = E[v_i(d)]$. By the law of iterated expectations,

$$W(d) = E[v_i(d)|a_i(d) > 0]P(a_i(d) > 0) + E[v_i(d)|a_i(d) \leq 0]P(a_i(d) \leq 0)$$

We know from the consumer's problem and the definition of $a_i(d)$ that 1) $a_i(d) \leq 0 \implies x_i(d) = d$ and 2) $a_i(d) > 0 \implies x_i(d) = x_i^* = \arg\max u_i(x)$. Substituting these into $v_i(d) = w_i(x_i(d), d) = u_i(x_i(d)) - \pi_i\gamma_i 1\{x_i(d) \neq d\}$ gives the result. ■

Proposition 1: *For any two defaults $d_0, d_1 \in X$:*

$$W(d_1) - W(d_0) = E[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i|PA]\, p(PA) - E[u_i(x^*) - u_i(d_1) - \pi_i\gamma_i|AP]\, p(AP) + E[u_i(d_1) - u_i(d_0)|PP]\, p(PP).$$

Proof: We know that $W(d_1) - W(d_0) = E[v_i(d_1) - v_i(d_0)]$. We partition individuals into the four groups $(PA, AP, PP$ and $AA)$ and apply the law of iterated expectations to express the change in welfare as a probability-weighted sum over these four groups. As before, $a_i(d) \leq 0 \implies x_i(d) = d$ and 2) $a_i(d) > 0 \implies x_i(d) = x_i^* = \arg\max u_i(x)$. In the PA group, $a_i(d_1) > 0$ so $v_i(d_1) = u_i(x_i^*) - \pi_i\gamma_i$,and $a_i(d_0) \leq 0$, so $v_i(d_0) = u_i(d_0)$. Thus $E[v_i(d_1) - v_i(d_0)|PA] = E[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i|PA]$. Proceeding similarly for the other four groups and substituting in the resulting expressions yields the desired result. ■

Proposition 2: *Let X be any interval in $\mathbb{R}$. If d^* represents an interior solution to the optimal default problem, the following first-order condition is satisfied:*

$$\begin{aligned} 0 = W^{'}(d^*) &= & E[(1-\pi_i)\gamma_i | a_i(d^*) = 0,\ u_i'(d^*) < 0]\, f_{a|u'<0}(0)\, F_{u'}(0) \\ &- & E[(1-\pi_i)\gamma_i | a_i(d^*) = 0,\ u_i'(d^*) > 0]\, f_{a|u'<0}(0)\ (1 - F_{u'}(0)) \\ &+ & E\left[u'(d^*)\,|\, a_i(d^*) < 0\right]\, F_{a;d^*}(0) \end{aligned}$$

where $f_{a|u'>0}$ is the probability density function of $a_i(d^*)$ conditional on $u_i'(d^*) > 0$; $F_{u'}$ is the cumulative density function of $u_i'(d^*)$; and, as above, $F_{a;d^*}$ is the cumulative density function of $a_i(d^*)$.

Proof: One can obtain this result by direct calculation of the derivative of the welfare function, as divided into active and passive choosers in Lemma 1 (i.e. expressing the expectations as integrals and applying Leibniz rule). One can also obtain the result by plugging in $d_1 = d_0 + \Delta d$ in Proposition 1, taking the limit as Δd approaches zero, plugging in the definitions of the primitives, and noting that the PA and AP groups now both have $a_i(d) = 0$, which implies that $u_i(x^*) - u_i(d) = \gamma_i$ by construction. ■

Proposition 3 Suppose that there exists a penalty default $d_p \in X$.

(3.1) *There exists a threshold $\underline{\pi} \in [0, 1)$ such that $\pi_i \leq \underline{\pi}$ for all i implies d_p maximizes social welfare.*

Proof: We will prove the existence of a threshold $\underline{\pi} \in [0, 1)$ such that when $\pi_i \leq \underline{\pi}$, $W(d^p) \geq W(d)$ for any d.

Let $X^A \subset X$ be the subset of X such that for any $d \in X^A$, $P(a_i(d) \leq 0) > 0$.

Let $d \in X$ be an arbitrary default. We know $W(d^p) \geq W(d)$ is trivially true when d is also a penalty default, i.e. $d \notin X^A$ as then $W(d) = W(d_p)$ for any π. Next suppose $d \in X^A$, so $p(PA) > 0$. Let $\tilde{\pi}(d) = \inf_{i \in PA(d)} \pi_i$ be the smallest possible value of π_i for the PA group for default d. We know from Equation (13) that

$$W(d_p) - W(d) \geq p(PA)\{E[u_i(x^*) - u_i(d)|PA] - \tilde{\pi}(d)E[\gamma_i|PA]\} \tag{22}$$

The RHS of this expression is a continuous and strictly monotonically decreasing function of $\tilde{\pi}(d)$ (so long as $E[\gamma_i|PA] > 0$, which must be true because PA individuals choose passively). When $\tilde{\pi}(d) = 0$, the RHS of this expression is strictly positive because $u_i(x^*) \geq u_i(d)$ for all i. When $\tilde{\pi}(d) = 1$, the RHS is strictly negative because $u_i(x^*) - u_i(d) < \gamma_i$ for all individuals that are passive at d, which is the PA group in this situation. The Intermediate Value Theorem then implies there is a value of $\tilde{\pi}(d)$, such that we know that the expression on the RHS of (22) is 0. Denoting this threshold by $\underline{\pi}(d)$, we have that $W(d_p) - W(d) \geq 0$ when $\pi_i \leq \underline{\pi}(d)$ for all i. The result then follows from letting $\underline{\pi} = \inf_{d \in X^A} \underline{\pi}(d)$, so that $\pi_i < \underline{\pi}$ implies $W(d_p) - W(d) \geq 0$ for any d. ■

(3.2) *There exists a threshold $\overline{\pi} \in (0,1]$ such that $\pi_i \geq \overline{\pi}$ for all i implies d_p* minimizes *social welfare.*

Proof: The proof is analogous to the proof of (3.1). For any default d, let $\hat{\pi}(d) = \inf_{i \in PA(d)} \pi_i$. Using equation (13) and a similar Intermediate Value Theorem argument to the above we derive that there is a threshold $\overline{\pi}(d)$, such that $\pi_i \geq \overline{\pi}(d)$ implies $W(d_p) - W(d) \leq 0$. The result then follows from letting $\overline{\pi} = \sup_{d \in X^A} \overline{\pi}(d)$. ■.

Proposition 4 *Suppose that $X = [x_{min}, x_{max}] \subseteq \mathbb{R}$ and that:*

(A4.1) *As-if costs γ_i are distributed independently of x_i^*.*

(A4.2) *Preferences are given by $u_i(x) = u(x - x_i^*)$ for some map $u : \mathbb{R} \to \mathbb{R}$, with $u'(0) = 0$, $u'' < 0$ and $u(c) = u(-c)$ for any c.*

(A4.3) *x_i^* follows a single-peaked and symmetric distribution about some mode x^m.*

Under these conditions, there exists a threshold $\overline{\pi} \in (0,1]$ such that $\pi_i \geq \overline{\pi}$ for all i implies that the optimal default is the default that minimizes opt-outs.

Proof: We provide the proof the theorem for the case when $\pi_i = 1$ for all i. It is straightforward to show that if the theorem holds when $\pi_i = 1$ for all i, it must hold for sufficiently high π_i.

Starting from the case where $\pi_i = 1$ for all i, we first prove that $W'(x^m) = 0$, $W'(d) > 0$ for $d < x^m$, and $W'(d) < 0$ for $d > x^m$., which implies that W has a unique global maximum at x^m. We then prove that opt-outs are minimized under x^m. We start by letting $d \in X$ be some default.

Step 1: Characterizing the first and second derivative of $W(d)$.

By (A4.1) we know that $W_i(d) = \int_\gamma E[v_i(d)|\gamma_i = \gamma] f(\gamma) d\gamma$. Let $W_\gamma(d) = E[v_i(d)|\gamma_i = \gamma]$. Note that to prove our result, it suffices to prove that for any fixed γ, $W_\gamma^{'}(d) = 0$ if $d = x^m$, and $W_\gamma''(d) < 0$ always.

We first introduce some notation involving the function $u()$. Without loss of generality $u(0) = 0$. Taking γ as given, by (A4.2) there is some unique value ξ such that $u(\xi) = u(-\xi) = \gamma$. Note that when $x^* = d - \xi$, utility at the default is given by $u(d - x^*) = u(d - (d - \xi)) = u(\xi) = \gamma$, and similarly when $x^* = d + \xi$, $u(d - (d + \xi)) = \gamma$. By (4.2), an individual is active when $x_i^* \leq d - \xi$ or $x_i^* \geq d + \xi$.

We next characterize $W_\gamma'(d)$. Assuming $\pi_i = \pi$ is homogeneous for all i (which is true when $\pi_i = 1$ for all i), welfare at d is given by

$$W_\gamma(d) = \int\limits_{x^*=-\infty}^{d-\xi} \gamma f(x^*)dx^* + \int\limits_{x^*=d-\xi}^{d+\xi} u(d-x^*)f(x^*)dx^* + \int\limits_{x^*=d+\xi}^{\infty} \gamma f(x^*)dx^*,$$

Where $f(x^*)$ is the pdf of x_i^*. Note that $f(x^*)$ does not depend on γ by (A4.1). Differentiating the above with respect to d and applying $u(d-x_i^*) = \gamma$ at $x_i^* = d-\xi$ or $d+\xi$, we obtain

$$W'_\gamma(d) = \gamma(1-\pi)[f(d-\xi) - f(d+\xi)] + \int\limits_{x^*=d-\xi}^{d+\xi} u'(d-x^*)f(x^*)dx^* \tag{23}$$

This is an analogue of Proposition 2 for some fixed γ, with the added structure of (A4.2). When $\pi = 1$ (A4.4), the first term of this expression, which corresponds to the PA and AP groups, vanishes, leaving only the PP group, which we now split into those with $x^* < d$ and those with $x^* > d$:

$$W'_\gamma(d) = \int\limits_{x^*=d-\xi}^{d} u'(d-x^*)f(x^*)dx^* + \int\limits_{x^*=d}^{d+\xi} u'(d-x^*)f(x^*)dx^*. \tag{24}$$

Step 2: For any constant ζ, $f(d+\zeta) \geq f(d-\zeta) \iff x^m \geq d$.

Suppose $x^m \geq d$ and take a constant ζ. If $x^m > d+\zeta > d-\zeta$, the result immediately follows from the assumption in (A4.3) that $f()$ is single-peaked. If $d+\zeta \geq x^m \geq d > d-\zeta$ take a constant c such that $d+\xi - x_m = x_m - c$. By symmetry about x^m, $f(c) = f(d+\zeta)$. We know that $c < x_m$, because $x_m - (d+\zeta) \leq 0$. We also know that $c \geq d-\zeta$, because we presumed $x_m \geq d$. We then have $x^m \geq c \geq d-\zeta$. The single-peaked assumption then implies $f(d+\zeta) = f(c) \geq f(d-\zeta)$.

Supposing $x^m < d$ and proceeding analogously proves the converse.

Step 3: $x^m \geq d \iff W'_\gamma(d) \geq 0$.

Starting from equation (24), note that by (A4.2) the first term is positive ($u' > 0$ when $x^* < d$) and the second term is negative ($u' < 0$ when $x^* < d$). We can compare the signs of the two terms in the previous expression by re-writing this equation, using the symmetry of the utility function, as:

$$W'_\gamma(d) = \int\limits_{x^*=d-\xi}^{d} u'(d-x^*)[f(x^*) - f(\tilde{x})]dx^*$$

where $\tilde{x} = 2d - x^*$, so that $d - x^* = -(d-\tilde{x})$. We know from symmetry that when $d = x^m$, $f(x^*) = f(\tilde{x})$, so $W'(x^m) = 0$.

As $u'(d-x^*) > 0$ in the range of integration we use above. When $x^m > d$, the result in Step 3 implies that $f(x^*) \geq f(\tilde{x})$ for $x^* \in [d-\zeta, d]$, so we know that $W'_\gamma(d) \geq 0$. When $x^m < d$, the result in step 2 implies that $f(x^*) \leq f(\tilde{x})$ for $x^* \in [d-\zeta, d]$, and we know that $W_\gamma'(d) \leq 0$.

Step 3 proves that there is a unique global maximum of W at x^m.

Step 4: Setting $d = x^m$ minimizes opt-outs.

Let the frequency of opt-outs be given by $A(d) = P(a_i(d) > 0)$. Using $\xi = u^{-1}(\gamma)$ from before and letting F be the cdf of x_i^*, we know that

$$A(d) = F(d-\xi) + 1 - F(d+\xi)$$

Taking a derivative with respect to d, we have that

$$A'(d) = f(d-\xi) - f(d+\xi).$$

Setting $d = x^m$, it is straightforward to verify using (A4.3) that $A'(d) = 0$ if $d = x^m$, $A'(d) < 0$ if $d < x^m$, and $A'(d) > 0$ if $d > x^m$, which is sufficient to prove that x^m minimizes $A(d)$. ■

Proposition 5 *In the model with internalities, suppose that*

(A5.1) Preferences determining behavior are given by $u_i(x) = u(x - x_i^a)$, where $u(x - x_i^a)$ is approximately quadratic and $u'(0) = 0$.

(A5.2) Normative preferences are given by $u_i(x) + m_i(x) = u(x - x_i^*)$ for the same map $u(\cdot)$.

(A5.3) *The error in active choice $x_i^a - x_i^*$ is independent of x_i^a and γ_i.*

Then the marginal social welfare effect of a change in the default is approximately given by $W_0^{'}(d) + \mu_m X^{'}(d)$, where $W_0(d)$ denotes social welfare without internalities (see Equation (12)), $\mu_m = E[\mu_i]$, and $X(d) = E[x_i(d)]$.

Proof **Step 1:** (A5.1) and (A5.2) imply that the internality $m(x)$ is linear.

By (A5.1) we can write

$$u_i(x) = \frac{u''(0)}{2}(x - x_i^a)^2.$$

By (A5.2) we can write

$$u_i(x) + m_i(x) = \frac{u''(0)}{2}(x - x_i^*)^2.$$

Subtracting the previous expression from this one and simplifying we obtain

$$m_i(x) = -u''(x_i^* - x_i^a)x + \frac{u''}{2}(x_i^{*2} - x_i^{a2}). \tag{25}$$

The second term is a constant with respect to x, and may there fore be safely ignored.

Step 2: Proving the result.

The result essentially follows from Equations (16) and (17) from the text. Specifically, apply the linear internality to Equation (17) to obtain:

$$\begin{aligned} W^{'}(d) &= W_0^{'}(d) + E\left[\mu_i(x_i^a - d)\,|\,PA\right]\ P(PA) \\ - & \qquad E\left[\mu_i(x_i^a - d)\,|\,AP\right]\ P(AP) \\ + & \qquad\quad E\left[\mu_i\,|\,PP\right]\ P(PP). \end{aligned}$$

Next, note that $\mu_i = m_i'(x) = -u''(x_i^* - x_i^a)$ by (25). Applying (A5.3) then implies that we can pull out the $E[\mu_i]$ terms.

$$W^{'}(d) = W_0^{'}(d) + \mu\{E[x^a - d|PA]P(PA) - E[x^a - d)|AP]P(AP) + P(PP)\}$$

Noting that the term inside curly brackets is the expression for $X'(d)$ in Equation (17), we obtain the desired result. ■

B Relationship to the Axiomatization of Masatlioglu and Ok (2005)

Masatlioglu and Ok (2005) provides an axiomatic characterization of a model very similar to the fixed as-if cost model we use. Their paper seeks to rationalize status quo bias; recall that we showed in Section 1.2.2 that giving extra utility to the status quo is the same as having a fixed cost of not choosing the status quo (see equation (4)). The representation of choices used by Masatlioglu and Ok (see their equations (3) and (4)) is isomorphic to our own (see our equation (2), and also equation (4)), with one exception: the fixed as-if cost could depend on the default in their model. Whether and to what extent γ depends on d is difficult to test empirically, but we know of no evidence suggesting that it does. Nevertheless, here we discuss further the implications of our restriction that γ does not depend on d by relaxing it and examining welfare.

Consider a model that is identical to our baseline model except that the fixed cost is a function of d for each individual, denoted $\gamma_i(d)$. It is straightforward to show that the derivative from Proposition 2 becomes

$$\begin{aligned} 0 \quad = \quad W'(d) \quad = \quad & E\left[\pi_i \gamma_i'(d) | \, AA\right] P(AA) \\ + \quad & E\left[(1-\pi_i)\gamma_i(d) | \, PA\right] P(PA) \\ - \quad & E\left[(1-\pi_i)\gamma_i(d) | \, AP\right] P(AP) \\ + \quad & E[u'(d)|PP]P(PP). \end{aligned} \tag{26}$$

This expression is identical to the expression in Proposition 2 except for the first term. In our basic model, individuals that are always active for a change in the default do not experience any change in their welfare. When the fixed costs depend on d and $\pi_i > 0$, changing the default can affect the welfare of these decision-makers because . The analogue of equation (11) is also straightforward to derive for this model.

First, we note that the argument in Proposition 3 (see the proof above) for active choices being optimal for sufficiently low π is unaffected by this addition. When as-if costs are not normative, forcing active choices still leads all individuals to receive x_i^* without incurring any costs. Whether forcing active choices *minimizes* welfare for sufficiently high π is unclear. The difficulty is that the penalty default d^p could in principle have a lower fixed cost ($\gamma(d^m)$) than other defaults, which can make the penalty default relatively more attractive than some other defaults.

We know by the same logic as Proposition 4 (proof above) that the last three terms of (26) will all be zero under (A3.1)-(A3.3) when we minimize opt-outs, and that ignoring the changes in $\gamma(d)$ for active choosers we would get to a global optimum by minimizing opt-outs when π_i is sufficiently high for all individuals. The additional term in Equation (26) therefore implies that minimizing opt-outs will not be optimal in general when the change in $\gamma(d)$ for a marginal change in the default is zero. Intuitively, if increasing the default from the opt-out minimizing default would reduce the cost incurred by active decision-makers, we know the aggregate effect on all other decision-makers is zero (by Proposition 3), so such an increase in the default would be an improvement on minimizing opt-outs. For a more extreme example, suppose there is a default d^* such that $\gamma_i(d^*) = 0$ for all i. Such a default is obviously the optimal default regardless of the π_i's.[33]

To summarize, our result that active choices are desirable when default effects are purely driven by behavioral frictions survives the extension implied by the model of Masatlioglu and Ok (2005). Minimizing opt-outs will still be a good rule of thumb when default effects are real costs and the dependence between the costs and the default is not too strong, but if the costs vary strongly with the default it may be possible to improve on the opt-out minimization rule of thumb.

[33] When $\pi = 0$, both the active choice policy and the default d^* are optimal defaults.

C Variable Opt-Out Costs

Thus far we have assumed that as-if opt-out costs are constant (for a given individual) and do not depend on which non-default option the decision-maker selects. An alternative behavioral model is that defaults "pull" decision-makers towards options near the default in addition to making them more likely to select the default itself. For example, defaults may serve as an anchor (Example 1.2.7).

Ultimately, the question of whether defaults effects can be better described by including variable as-if costs in the model is an empirical question. Empirical evidence, reviewed in Section 1.3, regularly finds that increases in the default can affect choices far away from the default, suggesting that fixed costs are likely present. A variable costs model alone, such as a model of anchoring and adjustment where a higher default tends to lead to higher $x_i(d)$, would not predict, for example, that the fraction of individuals who contribute nothing to their pension would increase when the default rate of contribution is increased. Whether adding variable costs gives the model *additional* explanatory power relative to the fixed-cost-only model is more difficult to test. One possibility is to look closely at choices around the default. The fixed costs model with no variable cost predicts a "hole" in the observed distribution of choices around the default, whereas adding variable costs model predicts a "hill" around the default when fixed costs are sufficiently low. Still, given that both fixed and variable costs are plausibly heterogeneous, separately identifying these two components of decision-makers' revealed preferences without strong assumptions about distributions of the two costs is difficult. Here, we show how the inclusion of a variable costs affects the conclusions of our main analysis, especially the desirability of active choices versus minimizing opt-outs.

We focus on the case where X is a real interval. Suppose that instead of (1), individual behavior is given by

$$x_i(d) = \arg\max_{x \in X}\; u_i(x_i) - c_i(x_i - d) - \gamma 1\{x_i \neq d\}. \tag{27}$$

For simplicity, we will assume that u_i is single-peaked, with $u_i'(x_i^*) = 0$ and $u_i'' < 0$ everywhere. For this extension, we assume that the as-if cost associated with choosing a non-default option increases the further the chosen option is from d, so that $c_i'(x_i - d) \geq 0$ when $x_i - d > 0$, and $c_i'(x_i - d) \leq 0$ when $x_i - d < 0$. The as-if cost function is twice differentiable, with $c'' \geq 0$. We normalize $c(0)$ to zero. In this model individuals choose the default when passive, or $\tilde{x}(d) = \arg\max u_i(x_i) - c_i(x_i - d)$ when active. The individual is active if $\tilde{a}_i(d) \equiv [u_i(\tilde{x}_i(d)) - c_i(\tilde{x}_i(d) - d)] - u_i(d) - \gamma_i > 0$.

Similar to before, welfare is given by

$$w_i(x) = u_i(x) - \rho_i c_i(x - d) - \pi_i \gamma_i 1\{x_i \neq d\}, \tag{28}$$

where ρ_i denotes the normative relevance of variable costs $c_i(\cdot)$ and π_i the normative relevance of fixed costs as before. Indirect utility and social welfare are also defined similarly to before.

Given any change in the default, we can divide individuals into four groups as before, except now these groups are based on $\tilde{a}_i(d)$. Taking a derivative of the welfare function with respect to d, we have that the necessary condition from Proposition 2 becomes, with the addition of variable costs,

$$\begin{aligned} 0 \quad = \quad W'(d) \quad &= \quad E\left[\rho_i c_i' + (1-\rho_i)c_i' \tfrac{c_i''}{c_i''-u_i''}\middle| AA\right] P(AA) \\ &+ \quad E\left[(1-\rho)c + (1-\pi_i)\gamma_i\right| PA] \, P(PA) \\ &- \quad E\left[(1-\rho)c + (1-\pi_i)\gamma_i\right| AP] \, P(AP) \\ &+ \quad E[u'(d)|PP]P(PP). \end{aligned} \tag{29}$$

where all components involving $c_i(\cdot)$ are evaluated at $x = \tilde{x}(d)$.

Adding variable costs changes this expression in two ways. First, the always-active choosers (AA) are affected by a change in the default. The sign of the welfare effect on an always-active chooser is positive if and only if $x_i^* < d$. For an individual with $x_i^* < d$, we will have that $x_i^* < x_i(d) < d$, and an increase in the default makes it costlier to choose an option close to x_i^*. The $\rho_i c_i'$ term of the welfare effect for members of the AA group in Equation 28 corresponds to the direct welfare effect of increasing this cost. Such an individual also increases x_i in response to this change in costs: it is straightforward to show that $\tilde{x}_i'(d) = \frac{c_i''}{c_i''-u_i''} \in [0, 1)$. The second term of the welfare effect for the AA group corresponds to the welfare impact of this change in behavior.[34] As before, when as-if costs are fully normatively relevant for all individuals, $\rho_i = 1$, and the envelope theorem eliminates the indirect welfare effect from the behavioral response. However, when $\rho_i < 1$, the individual over-reacts to the increase in costs, reducing their welfare. The opposite intuition applies when $x_i^* > d$; such individuals in the AA group are made better off by an increase in the default. The second addition to the welfare calculation is the extra variable cost incurred by marginally active decision-makers in the PA and AP groups. As it changes welfare discretely when the individual switches between choosing actively and choosing passively, this component affects welfare in exactly the same fashion as the fixed cost.

Our key result that forcing active choice is optimal when default effects are driven purely by behavioral frictions will still be true in this model, but properly examining an active choice policy requires subtle reasoning here. In this model, setting an extreme default so that everyone opts out will not necessarily be equivalent to forcing active choices directly. One might naturally suppose that when forcing active choices, the planner sets no anchor, which eliminates the variable costs, whereas when a penalty default acts as an anchor, the variable costs will matter for behavior and welfare. Suppose there is a policy that forces decision-makers to make active choices and eliminates variable costs (i.e. it does not set an anchor). It

[34] Note that the behavioral response is $\tilde{x}'(d) = 0$ when costs are linear, i.e. $c'' = 0$.

Table 1: Model Parameters and Plan Characteristics

Parameter	Value
Mean savings utility weight, μ_ρ, company 1	0.2150
Mean savings utility weight, μ_ρ, company 2	0.1313
Mean savings utility weight, μ_ρ, company 3	0.1570
Standard deviation of savings utility weight, σ	0.0910
Savings shift parameter, α	0.1340
Fraction with zero as-if costs, λ_1	0.4011
As-if costs distribution parameter, λ_2	11.81
Maximum matched contribution (all companies)	0.06
Employer match rate, company 1	1.0
Employer match rate, company 2	0.5
Employer match rate, company 3	0.5

Note: this table reports the parameter values we use in our empirical illustration. The parameer values come from Table 2 of Bernheim, Fradkin and Popov (2015), for the "basic model."

is straightforward to show that such a policy will be globally optimal when π_i is sufficiently small for all individuals (regardless of ρ_i), exactly as in Proposition 3. However, whether such a policy becomes extremely undesirable when π_i and ρ_i are sufficiently high is not clear in this model, because the policy that forces active choices also eliminates the variable costs and this can improve welfare. Conversely, a penalty default will surely minimize welfare when π_i and ρ_i are sufficiently high, but due to the large distortions on active choices it may have through the variable costs, it may not be optimal when π_i and ρ_i are large.

By a very similar procedure to the one we use in Proposition 4, one can show that minimizing opt-outs is optimal when π_i and ρ_i are sufficiently large, under some regularity conditions. Specifically, we could maintain Assumptions (A4.1)-(A4.3), and add the assumption that the variable cost function is the same for all individuals, $c_i(x_i - d) = c(x_i - d)$. Under these assumptions minimizing opt-outs will still be globally optimal when default effects are driven by real components of individual welfare.

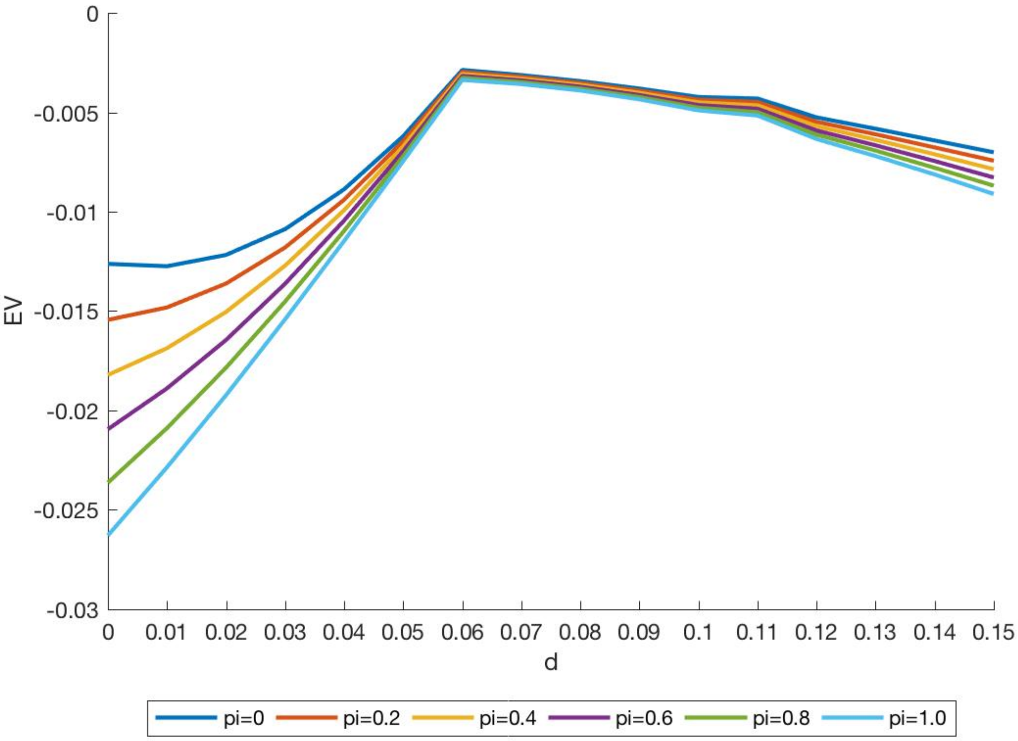
D Additional Details from Empirical Application

This appendix shows results from two other employers, labeled "Company 1" and "Company 2" in BFP, along with additional detail on the parameters used to calculate welfare in the model. In the model used here, the difference in distributions of contributions at different companies is used to identify differences in μ_ρ, which governs overall preferences over savings rates. Different companies also have different matching contribution rates. All other parameters are the same across companies. A complete table of parameter values is contained in Table 1.

Figure 6 repeats Figures 2 and 4 in the body of the text for Company 1. Figure 7 does the same for Company 2. We can see that apart from relatively minor differences, we obtain the same results for all three companies. The most noticeable difference is actually that the higher, 100 percent match rate at company 1 makes defaults lower than 6 percent much less desirable, which is intuitive.

Figure 6: Results for Company 1

(a) Equivalent variation over defaults, by π



(b) Active Choices versus Minimizing Opt-Outs

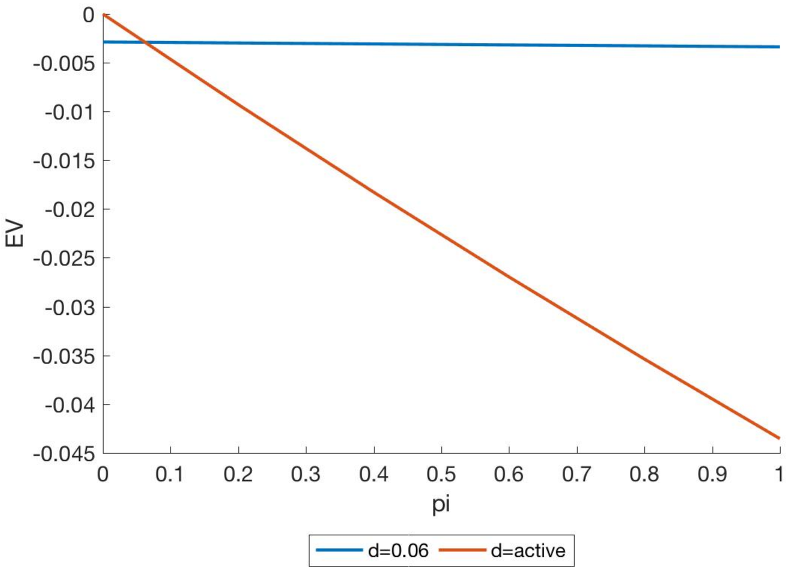
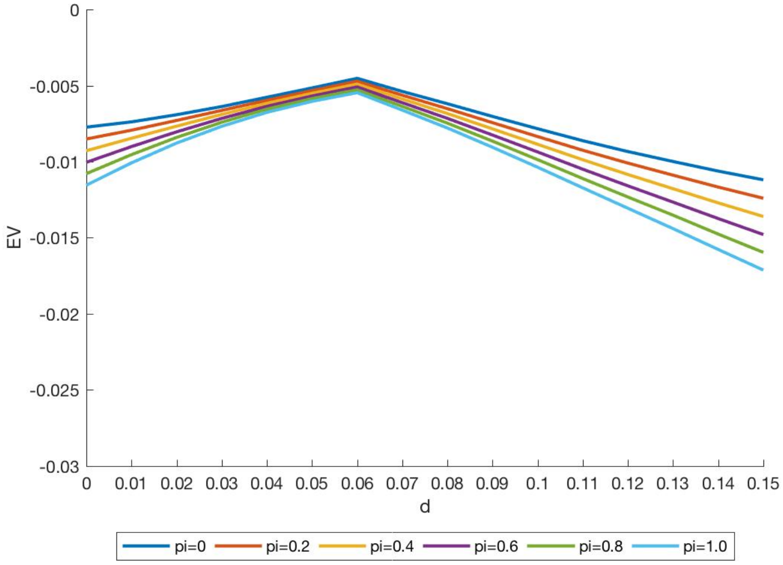
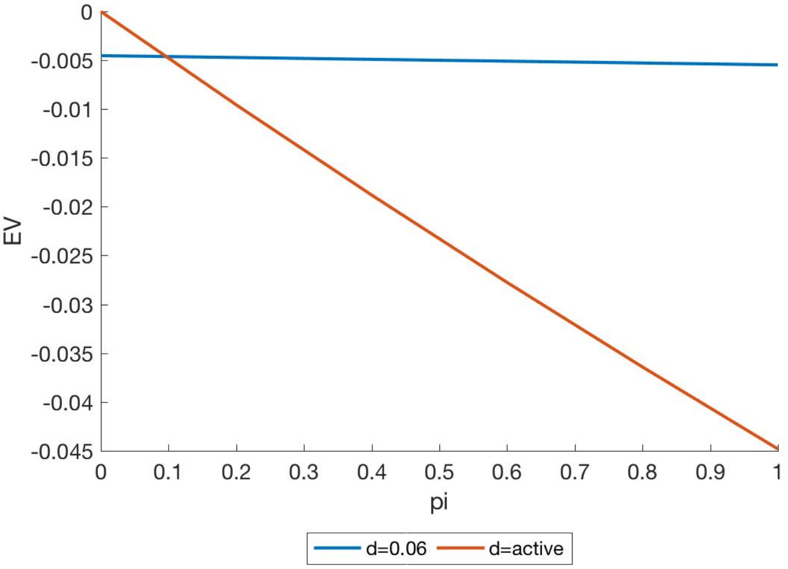


Figure 7: Results for Company 2

(a) Equivalent variation over defaults, by π



(b) Active Choices versus Minimizing Opt-Outs


EV
pi
d=0.06
d=active