The Welfare Economics of Default Options in 401(k) Plans

Appendix

B. Douglas Bernheim, Stanford University and NBER
Andrey Fradkin, NBER Igor Popov, Stanford University

March 27, 2015

1 Foundations for the state evaluation function

Assume that, as of period 0, the worker cares only about his opt-out effort level e and his (possibly state-contingent) future consumption trajectory, c, which encompasses not only goods but also effort subsequently expended to change contribution rates.[1] Period 0 preferences correspond to a utility function $u(e, \omega) + U(c, \theta)$.

Choosing c to solve the dynamic optimization problem of maximizing U subject to future opportunity constraints (parameterized by a vector π) for fixed x and z yields an optimal continuation consumption correspondence $C(x, z, \theta, \pi)$. (Here we interpret θ as including not only preference parameters but also existing stocks of wealth, which for our purposes can be treated as preference shifters.) We assume that, given x, π does not depend on d.[2] Defining the state evaluation (or indirect utility) function $V(x, z, \theta, \pi) = U(c, \theta)$ for $c \in C(x, z, \theta, \pi)$, we can we treat the worker's short-term problem as one of solving the maximization problem described in Section 2.1 of the main text.

Notice that d enters this problem only through the period 0 opportunity constraint for

[1] The elements of c are potentially indexed by both time and states of nature. All consumption other than e takes place after period 0.

[2] Future default rates depend on the initial default rate only indirectly through the initial contribution rate (which in practice establishes a new default).

(e, x, z) bundles; any choice of x renders the initial d subsequently irrelevant.[3] That observation allows us to implement our framework empirically by estimating the reduced-form valuation function V rather than the primitive utility function U, and to simplify the analysis of optimal defaults by working with reduced-form preferences over (e, x, z) bundles rather than primitive preferences over (e, c) bundles.[4] In taking this approach, it is possible that we will either (a) impose structure on V that is inconsistent with the underlying optimization problem, or (b) fail to impose structure implied by that problem. With respect to (a), our assumptions concerning V are modest and largely innocuous.[5] With respect to (b), we are skeptical of the prospects for deriving helpful properties of sufficient generality; in any event, empirical analysis adds appropriate structure by fitting V to data, and our theoretical analysis yields useful insights without additional structure.

2 Opt-out conditions

In this section, we derive the opt-out conditions presented in the main text.

The basic model with costly opt-out. The opt-out decision is governed by a comparison of $V(d, 1-\tau(d))$ and $V(x^*, 1-\tau(x^*), \theta) - \gamma$. Consequently, the worker opts out iff $\Delta(d) \geq \gamma$.

Sophisticated time inconsistency. When the opt-out decision is made in the contemporaneous frame, it is governed by a comparison of $\beta V(d, 1-\tau(d))$ and $\beta V(x^*, 1-\tau(x^*)) - \gamma$. Accordingly, the worker opts out iff $\Delta(d) \geq \beta^{-1}\gamma$. When the opt-out decision is made in the forward-looking frame, it is governed by a comparison of $\beta V(d, 1-\tau(d))$ and $\beta V(x^*, 1-\tau(x^*)) - \beta\gamma$. Accordingly, the worker opts out iff $\Delta(d) \geq \gamma$.

Naive time inconsistency. When the opt-out decision is made in a contemporaneous

[3] This property hinges on the assumed absence of any relation between the future opportunity set (parameterized by π) and the initial default rate d, given the initial contribution rate x.

[4] Without knowing anything about the correspondence C, we can conclude that the bundle (e, c) for $c \in C(x, z, \theta)$ is chosen over (and hence revealed preferred to) (e', c') for $c' \in C(x', z', \theta)$ from the observation that (e, x, z) is chosen over (e', x', z'); hence formal welfare analysis is possible.

[5] We explicitly acknowledge a potential exception in Section 5.1.

frame f, it is governed by a comparison of

$$\beta\kappa(f)\max\{V(x^*, 1-\tau(x^*)) - \gamma, V(d, 1-\tau(d))\} + \beta(1-\kappa(f))V(d, 1-\tau(d))$$

(the worker's anticipated payoff if he does not opt out immediately) and

$$\beta V(x^*, 1-\tau(x^*)) - \gamma$$

(his anticipated payoff if he opts out immediately). Plainly, if the max term in the first expression equals its second element, he will not opt out. Accordingly, he opts out iff

$$\beta(1-\kappa(f))V(d, 1-\tau(d)) + (1-\beta\kappa(f))\gamma \leq \beta(1-\kappa(f))V(x^*, 1-\tau(x^*)),$$

or

$$\Delta(d) \geq \frac{\beta^{-1} - \kappa(f)}{1-\kappa(f)}\gamma.$$

When the opt-out decision is made in a forward-looking frame f, it is governed by a comparison of

$$\beta\kappa(f)\left[\max\{V(x^*, 1-\tau(x^*)) - \gamma, V(d, 1-\tau(d))\}\right] + \beta(1-\kappa(f))V(d, 1-\tau(d))$$

(the worker's anticipated payoff if he does not opt out immediately) and

$$\beta V(x^*, 1-\tau(x^*)) - \beta\gamma$$

(his anticipated payoff if he opts out immediately). Once again, if the max term in the first expression equals its second element, he will not opt out. Accordingly, he opts out iff

$$\beta(1-\kappa(f))V(d, 1-\tau(d)) + \beta(1-\kappa(f))\gamma \leq \beta(1-\kappa(f))V(x^*, 1-\tau(x^*)),$$

or

$$\Delta(d) \geq \gamma.$$

Inattentiveness. For our model of inattentiveness, each frame specifies not only factors that influence attention, but also a status quo bundle. For opt-out choices, the contribution

rate for the status quo bundle coincides with the default. When the opt-out decision is made in a frame f' (for which the status quo contribution rate is d), it is governed by a comparison of $V(d, 1-\tau(d))$ and $V(x^*, 1-\tau(x^*)) - (\gamma + \chi(f'))$. Consequently, the worker opts out iff $\Delta(d) \geq \gamma + \chi(f')$.

Anchoring. With anchoring, when the opt-out decision is made in the frame f', it is governed by a comparison of $V(d, 1-\tau(d), d)$ and $V(x^*(f'), 1-\tau(x^*(f')), f') - \gamma$. Consequently, the worker opts out iff $\Delta(d, f') \geq \gamma$. With naturally occurring institutions, $f' = d$.

3 Additional technical assumptions

For the purpose of stating our additional technical assumptions, we make the dependence of V (and hence of x^*) on preference parameters, θ, explicit. We assume that V is strictly quasiconcave in (x, z), strictly increasing in both x and z, with $\lim_{z\to 0} V(x, z, \theta) = -\infty$ and $\lim_{z\to\infty} V(x, z, \theta) = +\infty$, and continuously differentiable (except at $z = 0$).[6] We allow the preference parameters $\xi \equiv (\gamma, \theta) \in [0, \overline{\gamma}] \times \Theta \equiv \Omega$ to differ across workers and use H to denote their CDF.[7] Except where stated otherwise, we assume H has full support on Ω and $\overline{\gamma}$ is very large, so that the fraction of individuals opting out of any default lies strictly between 0 and unity. We take Θ (and hence Ω) to be compact. We assume τ is strictly increasing, piecewise linear, continuous, and convex. Under our assumptions, the ideal point $x^*(\theta)$ is unique and varies continuously with θ. We assume that the (induced) distribution of $x^*(\theta)$ has full support on $[0, \overline{x}]$, with atoms at 0, $\overline{x}$, and the kink points of τ (if any), but nowhere else,[8] and that the density is bounded at all other points.

For all models with frame-dependent weighting, we assume that the mapping D is the same for all workers. For the models of naive time inconsistency and attention, we posit the

[6] When extending the model to anchoring, we make the same assumptions conditional on each frame f.

[7] Notice that we treat γ rather than some previously suppressed argument of u as the preference parameter governing opt-out costs; this is valid as long as we take the opt-out technology as fixed.

[8] This reasonable property can be derived from more primitive assumptions about the distribution of θ and the properties of V, but the associated technical issues do not illuminate the problem of interest. For the anchoring model, we make the same assumption about $x^*(\theta, f)$ for each f.

existence of some frame $\overline{f}$ either most conducive to naivete or least conducive to attention, and assume that, for any default d, the set of workers opting out has positive measure even with $\overline{f}$.

For the model with anchoring, we assume for some purposes that an increase in f weakly shifts the individual's choices toward higher x (*monotonicity*). Formally, if $u(e,\theta) + V(x,z,\theta,f) \geq u(e',\theta) + V(x',z',\theta,f)$, where $x > x'$ and $z < z'$, then $u(e,\theta) + V(x,z,\theta,f') > u(e',\theta) + V(x',z',\theta,f')$ for $f' > f$.

4 Proofs

The proofs of the theorems stated in the main text make use of the following notation. As in BR, we define a *generalized choice situation* (abbreviated GCS), $G = (X, f)$ as a constraint set X paired with a psychological frame f.[9] A choice correspondence C maps GCSs to the available alternatives the individual is willing to choose. We use $\mathcal{G}^*$ to denote the domain of the choice correspondence, and $\mathcal{G} \subseteq \mathcal{G}^*$ to denote the welfare-relevant domain.

Proof of Theorem 1

Let m denote a monetary transfer, and let $X(m)$ and f denote the individual's opportunity set and decision frame, respectively. For any alternative bundle x,[10]

$$EV_A(x) = \inf \{m \mid yP^*x \text{ for all } m' \geq m \text{ and } y \in C(X(m'), f)\}$$

and

$$EV_B(x) = \sup \{m \mid xP^*y \text{ for all } m' \leq m \text{ and } y \in C(X(m'), f)\}$$

First we show that if P_i^* is transitive, then zP_i^*x implies $EV_{Ai}(z) \geq EV_{Ai}(x)$ and $EV_{Bi}(z) \geq EV_{Bi}(x)$. Choose any $\varepsilon > 0$. By definition, yP_i^*z for all $m' \geq EV_{Ai}(z) + \varepsilon$ and $y \in C(X(m'), f)$. Thus, by transitivity, yP_i^*x for all $m' \geq EV_{Ai}(z) + \varepsilon$ and $y \in C(X(m'), f)$, which implies $EV_{Ai}(x) \leq EV_{Ai}(z)$. Similarly, by definition, xP_i^*y for all $m' \leq EV_{Ai}(x) - \varepsilon$ and

[9] Bernheim and Rangel (2009) used the term "ancillary condition" rather than psychological frame.

[10] The definitions given here are special cases of the definitions in Bernheim and Rangel (2009), in that here the alternative to the status quo is a specific bundle x, rather than an alternative opportunity set.

$y \in C(X(m'), f)$. Thus, by transitivity, zP_i^*y for all $m' \leq EV_{Ai}(x) - \varepsilon$ and $y \in C(X(m'), f)$, which implies $EV_{Bi}(z) \geq EV_{Bi}(x)$.

Next choose any $x' \in X_M$. If x' is a weak generalized Pareto optimum we are done, so suppose it is not. Consider the (necessarily) non-empty set $U = \{y \in X \mid yP_i^*x' \text{ for all } i\}$. Choose any individual j and consider some z' and f such that $(U, f) \in \mathcal{G}$ and $z' \in C_j(U, f)$.[11] We claim that z' is a weak generalized Pareto optimum in X. If it were not, then there would be some w such that wP_i^*z' for all i. By the transitivity of P_i^*, we would then have $w \in U$, which contradicts $z' \in C_j(U, f)$ (because in particular wP_j^*z'). From our first step, we then have $EV_{Ai}(z') \geq EV_{Ai}(x')$ and $EV_{Bi}(z') \geq EV_{Bi}(x')$ for all i, from which it follows that

$$\sum_i (\lambda_{Ai} EV_{Ai}(z') + \lambda_{Bi} EV_{Bi}(z')) \geq \sum_i (\lambda_{Ai} EV_{Ai}(x') + \lambda_{Bi} EV_{Bi}(x'))$$

Consequently, $z' \in X_M$. $\square$

Proof of Theorem 2

For workers who choose the default, let $m^0(d, \theta)$ be the solution to:

$$V(x^*(\theta), 1 + m^0(d, \theta) - \tau(x^*), \theta) = V(d, 1 - \tau(d), \theta). \tag{1}$$

Also let $m^1(\theta, \gamma, f)$ be the solution to:

$$V(x^*(\theta), 1 + m^1(\theta, \gamma, f) - \tau(x^*(\theta)), \theta) = V(x^*(\theta), 1 - \tau(x^*(\theta)), \theta) - D(f)\gamma, \tag{2}$$

where $D(0) = \beta^{-1}$ and $D(-1) = 1$. Our assumptions on V guarantee existence and uniqueness of the solutions, as well as continuity of the resulting functions. Plainly, $m^1(\theta, \gamma, f) < 0$ for $\gamma > 0$.

Given the compactness of $[0, \overline{\gamma}] \times \Theta$, there exists (γ', θ') that minimizes $m^1(\theta, \gamma)$ on that domain; moreover, because $V(x^*(\theta'), 0, \theta') = -\infty$ while $V(x^*(\theta'), 1 - \tau(x^*(\theta')), \theta') - D(f)\gamma'$ is finite, we know that $m^1(\gamma', \theta') \equiv m_L \in (-1, 0)$. Trivially, $m^0(d, \theta)$ achieves a maximum of

[11] Here we are employing the assumptions, stated in BR, that (i) $C(G)$ is non-empty for all $G \in \mathcal{G}^*$, and (ii) for every set Z there exists a frame f such that $(Z, f) \in \mathcal{G}$.

0 (when $x^*(\theta) = d$) on its domain. Because V is continuously differentiable and $[m_L, 0] \times \Theta$ is compact, $V_z(x^*(\theta), 1 + m, \theta)$ has a minimum, $v_L > 0$ (recall that V is strictly increasing in z) and a maximum, v_H, on that domain.

Define $Q(d, f)$ as the set of values of (θ, γ) for which the worker elects the default; i.e., (θ, γ) such that

$$V(d, 1 - \tau(d), \theta) \geq V(x^*(\theta), 1 - \tau(x^*(\theta)), \theta) - D(f)\gamma,$$

or equivalently

$$m^0(d, \theta) \geq m^1(\theta, \gamma, f).$$

Aggregate worker surplus is given by:

$$\int_\Omega m^1(\theta, \gamma, f) dH(\xi) + \int_{Q(d,f)} \left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] dH(\xi).$$

Only the second term, which measures the incremental benefit received by workers who elect the default, varies with d. Thus the worker-surplus maximization problem is:

$$\max_d \int_{Q(d,f)} \left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] dH(\xi) \tag{3}$$

Let $\phi(x)$ denote the fraction of individuals for whom $x^*(\theta) = x$. Under our assumptions, $\phi(x)$ is strictly positive for $x \in \mathcal{A}$ and zero otherwise. Let $\phi^* \equiv \max_{d \in \mathcal{A}} \phi(d)$.

Consider any $d \in \mathcal{A}$. For any individual with $x^*(\theta) = d$, we have

$$V(x^*(\theta), 1 + m^0(d, \theta) - \tau(x^*(\theta)), \theta) - V(x^*(\theta), 1 + m^1(\theta, \gamma, f) - \tau(x^*(\theta)), \theta) = D(f)\gamma.$$

It follows that

$$\left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] v_H \geq D(f)\gamma.$$

Consequently, we have

$$\int_{Q(d,f)} \left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] dH^\theta(\theta) dH_k^\gamma(\gamma) \geq \frac{\phi(d) D(f) \gamma_k}{v_H}. \tag{4}$$

Now consider any $d \notin \mathcal{A}$. From equations (1) and (2), we see that, for all $(\gamma, \theta) \in Q(d, f)$,

$$V(x^*(\theta), 1 + m^0(d, \theta) - \tau\left(x^*(\theta)\right), \theta) - V(x^*(\theta), 1 + m^1(\theta, \gamma, f) - \tau\left(x^*(\theta)\right), \theta) \leq D(f)\gamma$$

(where we have used the fact that $V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta) \geq V(d, 1 - \tau\left(d\right), \theta))$. It follows that

$$\left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] v_L \leq D(f)\gamma.$$

Consequently,

$$\int_{Q(d,f)} \left[m^0(d, \theta) - m^1(\theta, \gamma, f)\right] dH^\theta(\theta) dH_k^\gamma(\gamma) \leq \frac{D(f)\overline{\gamma}_k}{v_L} \int_{\overline{Q}(d,f,\overline{\gamma}_k)} dH^\theta(\theta). \tag{5}$$

where $\overline{Q}(d, f, \gamma) \subset \Theta$ denotes the opt-in set for a fixed value of γ, and where we have used the fact that an increase in γ expands the set of opt-ins.

Now suppose the theorem is false. Then there is some sequence H_k^γ with $\overline{\gamma}_k \to 0$ and $\gamma_k/\overline{\gamma}_k > e^* > 0$, and an associated sequence of optimal defaults $d_k \notin \mathcal{A}$ with $d_k \to d^* \notin \mathcal{A}$. Plainly, from (4) and (5),we must have, for all k,

$$\int_{\overline{Q}(d_k,f,\overline{\gamma}_k)} dH^\theta(\theta) \geq \frac{v_L}{v_H} \phi^* e^* > 0.$$

Accordingly, we will introduce a contradiction by demonstrating that $\int_{\overline{Q}(d_k,f,\overline{\gamma}_k)} dH^\theta(\theta) \to 0$.

We claim that, with a fixed opt-out cost of $\overline{\gamma}_k$, if $d_k \to d^* \notin \mathcal{A}$, then for all $\varepsilon > 0$ there exists K^ε such that for $k > K^\varepsilon$ all those with ideal points outside $(d^* - \varepsilon, d^* + \varepsilon)$ opt out. We prove this claim in four steps.

Step 1: With a fixed opt-out cost of $\overline{\gamma}_k$ and a default of $d^* - \frac{\varepsilon}{2}$, there exists K_L^ε such that for $k > K_L^\varepsilon$, all workers for whom $x^*(\theta) \leq d^* - \varepsilon$ opt out.

Because $x^*(\theta)$ is continuous and Θ compact, we know that $\{\theta \mid x^*(\theta) \leq d^* - \varepsilon\}$ is compact. Thus, we can define

$$\vartheta_L = \max_{\theta \text{ s.t. } x^*(\theta) \leq d^* - \varepsilon} \left[V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta) - V(d^* - \frac{\varepsilon}{2}, 1 - \tau\left(d^* - \frac{\varepsilon}{2}\right), \theta)\right].$$

Furthermore, because $x^*(\theta)$ is unique, we necessarily have $\vartheta_L > 0$ (otherwise we would have $x^*(\theta) = d^* - \frac{\varepsilon}{2}$ for some θ s.t. $x^*(\theta) \leq d^* - \varepsilon$). Step 1 then follows from the fact that there exists K_L^ε such that $D(f)\overline{\gamma}_k < \vartheta_L$ for all $k > K_L^\varepsilon$.

Step 2: With a fixed opt-out cost of $\overline{\gamma}_k$ and a default of $d^* + \frac{\varepsilon}{2}$, there exists K_H^ε such that for $k > K_H^\varepsilon$, all workers for whom $x^*(\theta) \geq d^* + \varepsilon$ opt out.

The proof mirrors that of Step 1. The set $\{\theta \mid x^*(\theta) \geq d^* + \varepsilon\}$ is also compact, so we define

$$\vartheta_H = \max_{\theta \text{ s.t. } x^*(\theta) \geq d^* + \varepsilon} \left[V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta) - V(d^* + \frac{\varepsilon}{2}, 1 - \tau\left(d^* + \frac{\varepsilon}{2}\right), \theta) \right],$$

and observe that $\vartheta_H > 0$. Step 2 then follows from the fact that there exists K_H^ε such that $D(f)\overline{\gamma}_k < \vartheta_H$ for all $k > K_H^\varepsilon$.

Step 3: With a fixed opt-out cost of $\overline{\gamma}_k$, any default $d \in \left[d^* - \frac{\varepsilon}{2}, d^* + \frac{\varepsilon}{2}\right]$, and $k > \max\{K_L^\varepsilon, K_H^\varepsilon\}$, all workers for whom $x^*(\theta) \notin (d^* - \varepsilon, d^* + \varepsilon)$ opt out.

Consider a worker for whom $x^*(\theta) \leq d^* - \varepsilon$. By Step 1, for $k > K_L^\varepsilon$ we know that

$$V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta) - D(f)\overline{\gamma}_k > V(d^* - \frac{\varepsilon}{2}, 1 - \tau\left(d^* - \frac{\varepsilon}{2}\right), \theta) \tag{6}$$

With $d \in \left[d^* - \frac{\varepsilon}{2}, d^* + \frac{\varepsilon}{2}\right]$, we also have

$$V(d^* - \frac{\varepsilon}{2}, 1 - \tau\left(d^* - \frac{\varepsilon}{2}\right), \theta) \geq V(d, 1 - \tau\left(d\right), \theta) \tag{7}$$

To see why, let $q \in (0,1)$ satisfy $qx^*(\theta) + (1-q)d = d^* - \frac{\varepsilon}{2}$, and define $\widetilde{z} = 1 - q\tau\left(x^*(\theta)\right) - (1-q)\tau(d)$. Because V is quasiconcave,

$$V(d^* - \frac{\varepsilon}{2}, \widetilde{z}, \theta, 0) \geq \min\left\{V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta), V(d, 1 - \tau\left(d\right), \theta)\right\} = V(d, 1 - \tau\left(d\right), \theta)$$

Because τ is convex, $V(d^* - \frac{\varepsilon}{2}, 1 - \tau\left(d^* - \frac{\varepsilon}{2}\right), \theta) \geq V(d^* - \frac{\varepsilon}{2}, \widetilde{z}, \theta, 0)$. Combining these inequalities yields (7). Combining (6) and (7), we obtain

$$V(x^*(\theta), 1 - \tau\left(x^*(\theta)\right), \theta) - D(f)\overline{\gamma}_k > V(d, 1 - \tau\left(d\right), \theta),$$

which implies that the worker opts out of d, as desired.

The case of any worker for whom $x^*(\theta) \geq d^* + \varepsilon$ is completely analogous, but employs Step 2 instead of Step 1.

Step 4: Now we prove the claim. Because $d_k \to d^*$, there exists K_I^ε such that, for $k > K_I^\varepsilon$, we have $d_k \in \left[d^* - \frac{\varepsilon}{2}, d^* + \frac{\varepsilon}{2}\right]$. Defining $K^\varepsilon = \max\{K_L^\varepsilon, K_H^\varepsilon, K_I^\varepsilon\}$, we see that for $k > K^\varepsilon$, with a fixed opt-out cost of $\overline{\gamma}_k$ and a default rate of d_k, all workers for whom $x^*(\theta) \notin (d^* - \varepsilon, d^* + \varepsilon)$ opt out.

Having established the claim, we now complete the proof of the theorem. If $d^* \notin \mathcal{A}$, then the measure of workers with ideal points in $(d^* - \varepsilon, d^* + \varepsilon)$, call it $y(\varepsilon)$, converges to zero along with ε. But plainly $y(\varepsilon) \geq \int_{\overline{Q}(d,f,\overline{\gamma}_k)} dH^\theta(\theta)$ for $k > K^\varepsilon$. Consequently, we have $\int_{D(d_k,\overline{\gamma}_k)} dH^\theta(\theta) \to 0$, and thus the desired contradiction. $\square$

Proof of Theorem 3

Part (i): Opt-out choices are unaffected by the change in decision frame unless $\Delta(d) \in [\gamma, \beta^{-1}\gamma]$ (and those at the boundaries of this interval may or may not change their choices). So consider a worker whose value of $\Delta(d)$ falls in this range; if he is on one of the boundaries, assume his choice changes. With opt-out choices made in the contemporaneous frame, EV_A is the value of m_A^C that satisfies[12]

$$V(x^*, 1 + m_A^C - \tau(x^*)) = V(d, 1 - \tau(d)).$$

With opt-out choices made in the forward-looking frame, EV_A is the value of m_A^F that satisfies

$$V(x^*, 1 + m_A^F - \tau(x^*)) = V(x^*, 1 - \tau(x^*)) - \gamma.$$

Thus we have

$$\begin{aligned} V(x^*, 1 + m_A^F - \tau(x^*)) - V(x^*, 1 + m_A^C - \tau(x^*)) &= \Delta(d) - \gamma \\ &\geq 0, \end{aligned}$$

with strict inequality when $\Delta(d) \neq \gamma$. It follows that, for those who change their choices, $m_A^F \geq m_A^C$, with strict inequality when $\Delta(d) \neq \gamma$.

[12]Recall that for this model, EV_A is assessed in the forward-looking frame.

Likewise, with opt-out choices made in the contemporaneous frame, EV_B is the value of m_B^C that satisfies[13]

$$V(x^*, 1 + m_B^C - \tau(x^*)) = V(d, 1 - \tau(d)).$$

With opt-out choices made in the forward-looking frame, EV_B is the value of m_B^F that satisfies

$$V(x^*, 1 + m_B^F - \tau(x^*)) = V(x^*, 1 - \tau(x^*)) - \beta^{-1}\gamma.$$

Thus we have

$$\begin{aligned} V(x^*, 1 + m_B^C - \tau(x^*)) - V(x^*, 1 + m_B^F - \tau(x^*)) &= \beta^{-1}\gamma - \Delta(d) \\ &\geq 0, \end{aligned}$$

with strict inequality when $\Delta(d) \neq \beta^{-1}\gamma$. It follows that, for those who change their choices, $m_B^C \geq m_B^F$, with strict inequality when $\Delta(d) \neq \beta^{-1}\gamma$.

Part (ii): Opt-out choices are unaffected by the change in decision frame unless $\Delta(d) \in [\gamma, \frac{\beta^{-1}-\kappa(f^*)}{1-\kappa(f^*)}\gamma]$ (and those at the boundaries of this interval may or may not change their choices). So consider a worker whose value of $\Delta(d)$ falls in this range; if he is on one of the boundaries, assume his choice changes. With opt-out choices made in the naturally occurring (maximally naive contemporaneous) frame, EV_A is the value of m_A^C that satisfies[14]

$$\begin{aligned} V(x^*, 1 + m_A^C - \tau(x^*)) &= \kappa(f^*) \max\{V(x^*, 1 - \tau(x^*)) - \gamma, V(d, 1 - \tau(d))\} + (1 - \kappa(f^*))V(d, 1 - \tau(d)) \\ &= \kappa(f^*)\left(V(x^*, 1 - \tau(x^*)) - \gamma\right) + (1 - \kappa(f^*))V(d, 1 - \tau(d)), \end{aligned}$$

where the second equality follows from $\Delta(d) \geq \gamma$. With opt-out choices made in any forward-looking frame, EV_A is the value of m_A^F that satisfies

$$V(x^*, 1 + m_A^F - \tau(x^*)) = V(x^*, 1 - \tau(x^*)) - \gamma.$$

Thus we have

$$\begin{aligned} V(x^*, 1 + m_A^F - \tau(x^*)) - V(x^*, 1 + m_A^C - \tau(x^*)) &= (1 - \kappa(f^*))\left(\Delta(d) - \gamma\right) \\ &\geq 0, \end{aligned}$$

[13]Recall that for this model, EV_B is evaluated in the contemporaneous frame.

[14]Recall that for this model, EV_A is assessed in a maximally naive forward-looking frame.

with strict inequality when $\Delta(d) \neq \gamma$. It follows that, for those who change their choices, $m_A^F \geq m_A^C$, with strict inequality when $\Delta(d) \neq \gamma$.

Likewise, with opt-out choices made in the contemporaneous frame, EV_B is the value of m_B^C that satisfies[15]

$$V(x^*, 1 + m_B^C - \tau(x^*)) = V(d, 1 - \tau(d)).$$

With opt-out choices made in the forward-looking frame, EV_B is the value of m_B^F that satisfies

$$V(x^*, 1 + m_B^F - \tau(x^*)) = V(x^*, 1 - \tau(x^*)) - \beta^{-1}\gamma.$$

Thus we have

$$V(x^*, 1 + m_B^C - \tau(x^*)) - V(x^*, 1 + m_B^F - \tau(x^*)) = \beta^{-1}\gamma - \Delta(d).$$

It follows that, for those who change their choices, $m_B^C < m_B^F$ for $\Delta(d) \in [\gamma, \beta^{-1}\gamma)$, $m_B^C > m_B^F$ for $\Delta(d) \in (\beta^{-1}\gamma, \frac{\beta^{-1}-\kappa(f^*)}{1-\kappa(f^*)}\gamma]$, and $m_B^C = m_B^F$ for $\Delta(d) = \beta^{-1}\gamma$.

Part (iii): Opt-out choices are unaffected by the change in decision frame unless $\Delta(d) \in [\gamma, \gamma + \chi(f^*)]$ (and those at the boundaries of this interval may or may not change their choices). So consider a worker whose value of $\Delta(d)$ falls in this range; if he is on one of the boundaries, assume his choice changes. With opt-out choices made in the naturally occurring frame (one with maximal inattentiveness in which the default is the status quo), EV_A is the value of m_A^I that satisfies[16]

$$V(x^*, 1 + m_A^I - \tau(x^*)) - \chi(f^*) = V(d, 1 - \tau(d)).$$

With opt-out choices choices made in maximally attentive frames, EV_A is the value of m_A^A that satisfies

$$V(x^*, 1 + m_A^A - \tau(x^*)) - \chi(f^*) = V(x^*, 1 - \tau(x^*)) - \gamma.$$

[15]Recall that for this model, EV_B is assessed in the minimally naive contemporaneous frame.

[16]Recall that for this model, EV_A is assessed in maximally inattentives frame for which the alternative to the baseline is the status quo.

Thus we have

$$V(x^*, 1 + m_A^A - \tau(x^*)) - V(x^*, 1 + m_A^I - \tau(x^*)) = \Delta(d) - \gamma$$
$$\geq 0,$$

with strict inequality when $\Delta(d) \neq \gamma$. It follows that, for those who change their choices, $m_A^A \geq m_A^I$, with strict inequality when $\Delta(d) \neq \gamma$.

Likewise, with opt-out choices made in the naturally occurring frame, EV_B is the value of m_B^I that satisfies[17]

$$V(x^*, 1 + m_B^I - \tau(x^*)) = V(d, 1 - \tau(d)) - \chi(f^*).$$

With opt-out choices made in maximally attentive frames, EV_B is the value of m_B^A that satisfies

$$V(x^*, 1 + m_B^A - \tau(x^*)) = V(x^*, 1 - \tau(x^*)) - \gamma - \chi(f^*).$$

Thus we have

$$V(x^*, 1 + m_B^A - \tau(x^*)) - V(x^*, 1 + m_B^I - \tau(x^*)) = \Delta(d) - \gamma$$
$$\geq 0,$$

with strict inequality when $\Delta(d) \neq \gamma$. It follows that, for those who change their choices, $m_B^A \geq m_B^I$, with strict inequality when $\Delta(d) \neq \gamma$. $\square$

Proof of Theorem 4

With zero opt-out costs, EV evaluated in frame f is given by the value of m satisfying

$$V(x_0, 1 + m - \tau(x_0), f) = V\left(x^*(d), 1 - \tau\left(x^*(d)\right), f\right),$$

where x_0 is the baseline contribution rate. Because V is strictly increasing in z, the value of d that maximizes the RHS also maximizes EV evaluated in frame f. By definition, the

[17] Recall that for this model, EV_B is assessed in maximally inattentive frames for which the baseline is the status quo.

solution to $\max_{x \in X} V(x, 1-\tau(x), f)$ is $x = x^*(f)$. It follows immediately that the solution to $\max_{d \in X} V(x^*(d), 1-\tau(x^*(d)), f)$ is $d = f$.

Next we show that EV, evaluated from the perspective of frame f, is non-decreasing for $d < f$ and non-increasing for $d > f$. First observe that, as a consequence of our monotonicity assumption, $x^*(d)$ is non-decreasing in d. Second, note that $V(x, 1-\tau(x), f)$ is non-decreasing in x for $x < x^*(f)$ and non-increasing in x for $x > x^*(f)$. To see why, consider any x', x'' with $x'' > x' \geq x^*(f)$. Let $z' = 1-\tau(x')$, $z'' = 1-\tau(x'')$, and

$$\widetilde{z} = (1-\tau(x'')) \frac{x' - x^*(f)}{x'' - x^*(f)} + (1-\tau(x^*(\theta))) \frac{x'' - x'}{x'' - x^*(f)}.$$

Because V is quasiconcave,

$$V(x', \widetilde{z}, f) \geq \min \{V(x^*(f), 1-\tau(x^*(f)), f), V(x'', 1-\tau(x''), f)\} = V(x'', 1-\tau(x''), f).$$

Because τ is convex,

$$V(x', 1-\tau(x'), f) \geq V(x', \widetilde{z}, f).$$

Combining these inequalities, we have

$$V(x', 1-\tau(x'), f) \geq V(x'', 1-\tau(x''), f),$$

as desired. An analogous argument establishes the same inequality for $x'' < x' \leq x^*(f)$. Third, it follows as a consequence of the first two steps that $V(x^*(d), 1-\tau(x^*(d)), f)$ is non-decreasing in d for $d < f$ and non-increasing in d for $d > f$. The desired properties then follow from the fact that $V(x_0, 1+m-\tau(x_0), f)$ is non-decreasing in m.

Now suppose all choices are deemed welfare-relevant. Consider part (i). From the first part of the proof, we know that, for every worker, the best outcome from the perspective of frame f is achieved by setting $d = f$. Because this model satisfies the multi-self conditions, the best option for worker i from the perspective of any frame f is unimprovable according to P_i^*. It follows immediately that every default rate d is a weak generalized Pareto optimum.

Now we turn to part (ii). Consider two frames, f and f', with $f' > f$. Suppose $x_0 < x^*(d)$. Equivalent variation assessed from the perspective of f, call it m_f, satisfies

$$V(x_0, 1+m_f-\tau(x_0), f) = V\left(x^*(d), 1-\tau\left(x^*(d,\theta)\right), f\right).$$

By monotonicity, we have

$$V(x_0, 1 + m_f - \tau(x_0), f') < V\left(x^*(d), 1 - \tau\left(x^*(d)\right), f'\right).$$

Defining $m_{f'}$ as equivalent variation from the perspective of frame f', it follows immediately that $m_{f'} > m_f$. Therefore, with $x_0 < x^*(d)$, EV_A is assessed from the perspective of frame $f = \overline{x}$, and EV_B is assessed from the perspective of the frame $f = 0$. An analogous argument implies that, with $x_0 > x^*(d)$, EV_A is assessed from the perspective of frame $f = 0$, and EV_B is assessed from the perspective of the frame $f = \overline{x}$. Part (ii) then follows directly from the first portion of the theorem. □

Proof of Theorem 5

Throughout this proof, we use i to denote a particular worker. BR define the relation R_i^* as follows: xR_i^*y iff $y \in C_i(X, f)$ implies $x \in C_i(X, f)$ for all $(X, f) \in \mathcal{G}$. Also, x is a weak generalized Pareto improvement over y iff xR_i^*y for every individual and xP_i^*y for some individual.

Part 1: Regardless of whether the welfare-relevant domain is restricted or unrestricted, offering a plan with $d = 0$, where choices are made in frame f^c such that $D(f^c) \geq D_M$ for cases of frame-dependent weighting, yields a weak generalized Pareto improvement over no plan.

Partition the set of employees into two groups, those who opt out and those who do not (both of which have positive measure under our assumptions). Those who do not opt out receive the bundle $(e, x, z) = (0, 0, 1)$ both with and without the plan. By definition, $(0, 0, 1)R_i^*(0, 0, 1)$. A worker who opts out chooses some bundle (e', x', z'), where $x' > 0$ and $z' < 1$, over the bundle $(0, 0, 1)$. With anchoring, the choice is made in frame $f^c = d = 0$, and our monotonicity assumption implies that the same worker would choose (e', x', z') over $(0, 0, 1)$ in any frame $f > 0$. With frame-dependent weighting, the choice is made in some f^c with $D(f^c) \geq D_M$, from which it follows that the same worker would choose (e', x', z') over $(0, 0, 1)$ in any welfare-relevant frame f. Thus, we have $(e', x', z')P_i^*(0, 0, 1)$ for those

who opt out.[18] The desired conclusion follows directly.

Part 2: Regardless of whether the welfare-relevant domain is restricted or unrestricted, and regardless of the prevailing choice frame for the cases of frame-dependent weighting, offering a plan with $d > 0$ does not yield a weak generalized Pareto improvement over no plan.

Consider the set of workers for whom $x^*(\overline{x}, \theta^i) = 0$ in the case of anchoring, and $x^*(\theta^i) = 0$ in the case frame-dependent weighting (both of which have positive measure under our assumptions). In the prevailing choice frame, f^c, such workers either opt out to $x = 0$ and receive the bundle $(e', 0, 1)$ (in the case of anchoring, any workers opting out would choose $x = 0$ because, by our monotonicity requirement, $x^*(\overline{x}, \theta^i) = 0$ implies $x^*(f, \theta^i) = 0$ for all f, including f^c), or fail to opt out and receive the bundle $(0, d, 1 - \tau(d))$. In the first case we do not have $(e', 0, 1)R_i^*(0, 0, 1)$, and in the second we do not have $(0, d, 1 - \tau(d))R_i^*(0, 0, 1)$ (in the cases of frame-dependent weighting because $x^*(\theta^i) = 0$, and in the case of anchoring because because $x^*(\overline{x}, \theta^i) = 0$ implies $x^*(f, \theta^i) = 0$ for all f). The desired conclusion follows directly.

Part 3: For models with frame-dependent weighting, a plan with $d = 0$ does not achieve a weak generalized Pareto improvement over no plan if choices are made in some frame f^c with $D(f^c) < D_M$.

Suppose $d = 0$ and that choices are made in such a frame. Consider the set of workers for whom $\gamma^i \in \left(\frac{1}{D_M}\Delta(d, \theta^i), \frac{1}{D(f^c)}\Delta(d, \theta^i)\right)$, which has positive measure (because the interval is open for all θ^i). Because the choice frame is f^c, any such worker opts out and receives the bundle $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$. But the same worker would choose the bundle $(0, 0, 1)$ over $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$ in a frame f such that $D(f) = D_M$. Thus, we do *not* have $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))\, R_i^*(0, 0, 1)$.

Part 4: For models with frame-dependent weighting, fixing $d = 0$, a plan with choices made in frame f^c such that $D(f^c) = D_M$ achieves a weak generalized Pareto improvement

[18] The same reasoning implies that, for those who are willing to either opt out or choose the default, we have $(e', x', z')R_i^*(0, 0, 1)$.

over any plan with choice made in frame f' with $D(f') > D_M$.

Suppose $d = 0$ and consider the choice frames f^c and f' with $D(f') > D(f^c) = D_M$. We partition the set of workers as follows: for group L, $\gamma^i < \frac{1}{D(f')}\Delta(d, \theta^i)$; for group I, $\gamma^i \in \left(\frac{1}{D(f')}\Delta(d, \theta^i), \frac{1}{D_M}\Delta(d, \theta^i)\right)$; and for group H, $\gamma^i > \frac{1}{D_M}\Delta(d, \theta^i)$. (We will consider workers at the boundaries between these groups separately below.) For the same reasons as in Part 3, each of these groups has positive measure. Those in group L opt out and receive the bundle $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$ in both frames, and by definition $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))\, R_i^*\, (e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$. Those in group H end up with $(0, d, 1 - \tau(d))$ in both frames because they do not opt out, and by definition $(0, d, 1 - \tau(d)) R_i^*(0, d, 1-\tau(d))$. Those in group I opt out in frame f^c, receiving bundle $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$, and do not opt out in frame f', receiving bundle $(0, d, 1 - \tau(d))$. Moreover, all such workers would choose $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$ over $(0, d, 1 - \tau(d))$ in all welfare-relevant frames. Thus, $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))\, P_i^*(0, d, 1-\tau(d))$. We treat workers at the boundary between groups L and I the same as members of group I if they opt out in frame f', and the same as members of group L if they do not opt out in frame f'. We treat workers at the boundary between groups I and H the same as members of group H if they do not opt out in frame f^c; if they do opt out in frame f^c, we still have $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))\, R_i^*(0, d, 1-\tau(d))$ because, in any welfare-relevant evaluation frame f'', they choose $(e', x^*(\theta^i), 1 - \tau(x^*(\theta^i)))$ over $(0, d, 1 - \tau(d))$ strictly if $D(f'') < D_M$, and weakly if $D(f'') = D_M$. The desired conclusion follows directly. □

5 Model fit

In this section, we present figures depicting the fitted and actual distributions of employee contribution rates under each default regime for each of the three companies. Figure A.1 is for the basic model, while Figure A.2 is for the anchoring model.

6 Additional simulation results

In this section we provide the following supplementary figures, all of which pertain to models of frame-dependent weighting. Figures A.3 through A.6 pertain to our model of inattention; they show EV_A^I and EV_B^I as functions of the default rate for, respectively, decisions made in the naturally occurring frame with an employee match, decisions made in the naturally occurring frame without an employee match, decisions made in the alternative frame with an employee match, and decisions made in the alternative frame without an employee match. Figures A.7 through A.10 contain the same information as figures 1, 2, 5, and 6 in the text, except that here we have extended the range of the default rates to 90%. Figure A.11 shows the overall opt-out frequencies as functions of the default rate for decisions made in the naturally occurring and alternative frames with an employee match; Figure A.12 shows the same opt-out frequencies without an employee match.