Soc Choice Welfare (2007) 28: 337–359
DOI 10.1007/s00355-006-0169-0

ORIGINAL PAPER

Nathan Berg · Gerd Gigerenzer

Psychology implies paternalism? Bounded rationality may reduce the rationale to regulate risk-taking

Received: 5 September 2005 / Accepted: 14 February 2006 / Published online: 23 June 2006

Abstract Behavioral economists increasingly argue that violations of rationality axioms provide a new rationale for paternalism – to "de-bias" individuals who exhibit errors, biases and other allegedly pathological psychological regularities associated with Tversky and Kahneman's (in Science 185:1124–1131, 1974) heuristics-and-biases program. The argument is flawed, however, in neglecting to distinguish aggregate from individual rationality. The aggregate consequences of departures from normative decision-making axioms may be Pareto-inferior or superior. Without a well-specified theory of aggregation, individual-level biases do not necessarily imply losses in efficiency. This paper considers the problem of using a social-welfare function to decide whether to regulate risk-taking behavior in a population whose individual-level behavior may or may not be consistent with expected utility maximization. According to the social-welfare objective, unregulated aggregate risk distributions resulting from non-maximizing behavior are often more acceptable (i.e., lead to a weaker rationale for paternalism) than population distributions generated by behavior that conforms to the standard axioms. Thus, psychological theories that depart from axiomatic decision-making norms do not necessarily strengthen the case for paternalism, and conformity with such norms is generally not an appropriate policy-making objective in itself.

N. Berg (✉)
School of Social Sciences, University of Texas at Dallas, GR 31 211300,
Box 830688, Richardson, TX 75083-0688, USA
E-mail: nberg@utdallas.edu

N. Berg · G. Gigerenzer
Center for Adaptive Behavior and Cognition,
Max Planck Institute for Human Development,
Lentzeallee 94, 14195 Berlin, Germany

1 Introduction

By now it is uncontroversial to assert that real-world decision makers frequently depart from the prescriptions of normative decision theory. What remains controversial is how that descriptive claim, which draws on psychological theory and a rich experimental literature, is to be interpreted and used in the analysis of economic policy. At the heart of the controversy surrounding psychology and economics, a field also known as behavioral economics, is the question: if one accepts challenges to the assumptive paradigm of economic rationality, does it follow that governments should pursue policies of benevolent paternalism?[1]

There are two main lines of argument concerning paternalism based on different interpretations of the experimental record. Some argue that behavioral economics has succeeded in identifying a body of well-established decision-making biases that justify government intervention, or at least minor institutional modifications to help individuals avoid systematically mistaken behavior resulting in ex post regret. There is a range of positions supportive of at least limited forms of paternalism or its subtle variant, opposition to anti-paternalism (i.e., anti-anti-paternalism).[2] For instance, Sunstein and Thaler (2003) and Thaler and Sunstein (2003) argue that because preferences have been shown to be unstable in a number of policy-relevant settings, neutral or *a*-paternalistic policies do not exist. According to Thaler and Sunstein, institutions necessarily establish defaults that function as anchors or frames and influence patterns of choice. Therefore, policy makers should be openly strategic in establishing institutions that are as helpful as possible, particularly with regard to defaults. According to the anti-anti-paternalism position, all policy impinges on choice to some extent and is therefore paternalistic, Thaler and Sunstein argue that anti-paternalism, because it fails to deal with the inevitability that institutions influence preferences, is willfully noncommital as a normative approach because it refuses to take a clear stand concerning better and worse distributions of property rights.[3]

Thaler and Sunstein see opportunities for policy makers to beneficially influence individual choice by wisely selecting institutional defaults (e.g., whether employees must opt-in or opt-out of savings programs; whether individuals are presumed to be organ donors or non-donors; and whether auto insurance policies are mandated to include or exclude the right to sue for punitive damages, allowing for opt-out and opt-in options, respectively). They argue that prescriptive guidelines for setting institutional defaults that include opt-out provisions do not coerce but rather frame choice in socially useful ways. Thus, opposition to anti-paternalism can, they claim, be libertarian in spirit.

Based on the theoretical behavioral economics literature and findings from psychological experiments, O'Donoghue and Rabin (2003) advocate what they call

[1] See, for example, Sunstein (1997), Ng (1999, 2003), Sheshinski (2002, 2003), Sunstein and Thaler (2003), Thaler and Sunstein (2003), Camerer et al. (2003), O'Donoghue and Rabin (2003), Caplin and Leahy (2003), Bernheim and Rangel (2004), Kopcke et al. (2004), and Frey and Stutzer (2004).

[2] Suber (1999) provides a definition of paternalism in historical context with discussion of related philosophical issues.

[3] Thaler and Benartzi (2004) follow Raiffa (1982) in distinguishing normative theory, which establishes an ideal benchmark based on optimization, from prescriptive theory, which attempts to advise decision makers how to move closer to that ideal.

asymmetric paternalism, making a stronger case for policies aimed at influencing individual choice than Thaler and Sunstein call for. The goal of asymmetric paternalism is to help those with self-control problems cut back on "sinful" consumption (i.e., consumption followed by regret) while imposing minimal burden on those who indulge rationally, that is, without regret. Bernheim and Rangel (2004) appear to take a similar view of behavioral economics as a subdiscipline whose primary focus is pathologically biased decision making, carrying with it the implication of an expanded scope (relative to other subfields in economics) for welfare-improving paternalistic regulation. The view is not unusual. Behavioral economists have argued that myopia among retirement savers legitimizes forced savings programs (Aaron 1999; Choi et al. 2005); that overconfidence justifies required cooling-off periods before finalizing large consumer purchases and financial decisions (Camerer et al. 2003); and that imprecise and costly decision making creates new reasons for governments to restrict consumer choice sets through regulation (Iyengar and Lepper 2000; Sheshinski 2003; Schwartz 2004). The medical ethics and behavioral law and economics literatures complement the abundant theoretical work in behavioral economics by providing numerous applications in which difficult questions of paternalism arise, for example Gruber (2002), Slovic (2000), Sunstein et al. (2000), Schneider (1998), Elster (1992) and VanDeVeer (1986).

In an opposing interpretation of the empirical failures of the standard rationality paradigm, economists and psychologists have argued that aggregates of limited-information decision makers using heuristics, or rules of thumb, can achieve high levels of economic performance and social coordination without complete descriptions of their environments or well-defined objective functions (Arthur 1994; Buchanan and Wagner 1977; Hayek 1945; Herrmann-Pillath 1996; Hildebrand 1994; Kirman 1983, 1993; Kysar et al. forthcoming; Lesourne 1992; Rossi 2004; Smith 2003; Tisdell 1996; Vriend 1995). Although it remains an open question as to which departures from standard normative benchmarks in economics can be rationalized as adaptive with respect to intuitively reasonable alternative performance criteria, the phenomenon of adaptiveness-enhancing departures from usual normative benchmarks has been documented in a variety of theoretical settings. For example, organisms whose decision rules are based on a few cues rather than high-dimensional constrained optimization may enjoy improved survival chances when the environment undergoes periodic cataclysmic change (Bookstaber and Langsam 1985). Simple heuristics such as mimicking one's peers can induce additional risk-taking and consequently improve average returns in a risky investment task (Berg and Lien 2003). Distortion of investors' beliefs away from rational expectations can move financial markets with asymmetric information to Pareto-superior equilibria (Berg and Lien 2005). By the criterion of accuracy for out-of-sample prediction (i.e., cross-validation rather than data fitting), information-frugal prediction rules have been shown to unambiguously outperform large regression models using various real-world data sets (Gigerenzer et al. 1999). And, Cosmides and Tooby (1994) describe the human mind's specialization for handling regularly occurring problems in our hunter–gatherer past as "better than rational" when rationality is defined according to the standard definition in economics, as a set of content-blind axioms or norms for general problem solving.

Accordingly, the psychological view of man, together with the observation that aggregates of boundedly rational individuals frequently find solutions to coordination problems that the best informed experts cannot (Simon 1978, 1982; Epstein 1995, 2004; Surowiecki 2004), may be interpreted as supportive of a decentralized or laissez faire approach to policy.[4] Asserting a positive link between psychology and anti-paternalism, however, is just as much of an over-generalization as asserting that there are links in the opposite direction. This paper argues instead that alleged links between psychological models of man and arguments concerning paternalism rest on selective emphases of the psychological literature and narrow definitions of rationality. Our claim is that the policy implications of theoretical and empirical departures from neoclassical rational choice are indeterminate. Without specifying more institutional detail, committing to specific social-welfare metrics, and imposing auxiliary assumptions about social interaction and the problem of aggregation, psychology implies no definitive position on paternalism.

This is not to say, however, that theorizing about paternalism should be avoided. On the contrary, political–economic debates about the role of government and institutional design generate a legitimate need for theoretical arguments, pro and con, concerning paternalism. The paternalism question is, in practice, substantive and deserves serious consideration rather than cursory judgments based on a priori ideological conviction. Hayek, for example, in a single publication, opposed redistribution in principle while endorsing the idea of a "safety net" in particular instances (Caldwell 2004).

The questions *How much choice?* and *Which distributions of rights?* are too important to allow narrow debates about axiomatic rationality to crowd out serious analysis of real-world policy problems. The intended contribution of the model presented in this paper is to provide a concrete case in which the rationale for paternalism can be quantified and shown to be indeterminate with respect to changing behavioral hypotheses ranging from psychological to neoclassical man. This indeterminacy demonstrates the absence of logical links between psychology and paternalism.

2 Quantifying paternalism

2.1 A policy maker with paternalism parameter ϵ

Consider a policy maker who wishes to limit what he or she regards as excessive risk-taking in a heterogeneous population. The meaning of "excessive" reflects the policy maker's beliefs about acceptable and unacceptable degrees of risky behavior, summarized by the parameter ϵ, which is interpreted as the policy maker's taste for paternalism. The policy maker's notion of excessive risk may be motivated by a theory of externalities, intergenerational distributional concerns, or a desire to protect individuals (whose preferences are time-inconsistent or beliefs are biased) from experiencing ex post regret. Given such a mechanism that motivates the policy maker to limit risk-taking, ϵ can in principle be endogenized. Section 2.5 provides an example of this in which ϵ is endogenized as a function of policy costs and social costs generated by an externality.

[4] A more traditional price-adjustment argument that reaches a similar conclusion is given by Caplan (1999), who argues that the price mechanism can achieve optimal levels of irrationality, concentrating irrationality in decision domains where its costs are low.

The model is a descriptive account of the policy maker's reasoning process rather than an efficiency argument in favor of any particular level of paternalism. To descriptively address the question of how behavioral hypotheses influence reasoning about paternalistic intervention, it is not necessary to take a stand on whether the policy maker is right or wrong in his or her definition of excessive risk. This is analogous to the optimal tax literature's convention of taking the tax authority's goal of raising a pre-specified quantity of revenue as exogenously given.

It is assumed that ϵ lies on the unit interval with $\epsilon = 0$ indicating zero desire for intervention and $\epsilon = 1$ indicating a preference for the absolute abolishment of risk-taking in the specified domain. The extremes of zero intervention and absolute abolishment are not uncommon among real-world policy responses to risky behavior. For example, restrictions limiting the amount of minutes individuals can spend jogging are virtually unheard of, while bans forbidding the use of cocaine in any amount are nearly universal. Intermediate restrictions ($0 < \epsilon < 1$) are probably even more common, examples of which are discussed subsequently.

2.2 The policy maker's preferences over population distributions of risk-taking

Corresponding to every fully-specified behavioral theory, θ is a pre-regulation population distribution of individual risk-taking represented by the pdf $f_\theta(x)$ where x denotes risk-taking actions ranging from minimum ($x = 0$) to maximum ($x = 1$) risk. The pdf $f_\theta(x)$ always denotes the pre-regulation distribution, and regulated risk-taking distributions are expressed in terms of it. The risk-taking action of a generic individual drawn at random from a population whose behavior is distributed according to theory θ is denoted x_θ.

The model assumes that the policy maker decides on whether the paternalistic policy $\epsilon > 0$ should be implemented by assigning to ϵ a social-welfare score and comparing it with the social-welfare score for the zero-regulation case.[5] Social-welfare scores carry the disadvantage, however, of depending on difficult-to-measure parameters needed to weight different choices of x, to quantify social costs, and to account for the costs of implementing and administering ϵ. The following section investigates conditions under which policy-specific parameters drop out of the policy maker's social-welfare comparison of behavioral hypotheses. The parameters drop out because variation in social welfare is generated by the policy maker's consideration of different behavioral hypotheses (i.e., variation in θ) rather than changes in the paternalism parameter (i.e., variation in ϵ).

Quantifying the policy maker's rationale for paternalism as the social-welfare gains from regulation provides the basic tool for modeling policy responses to changing behavioral assumptions. To see how the rationale for paternalism changes when the behavioral theory $f_{\theta_0}(x)$ is replaced by $f_{\theta_1}(x)$, a difference-in-differences is computed: the gains from regulation when $f_{\theta_1}(x)$ describes the unregulated population minus the gains from regulation when $f_{\theta_0}(x)$ is the unregulated distribution.

[5] Non-social-welfare metrics would also be possible, for example, if the policy maker were a rent seeker or acted on more complex political motives. The social-welfare quantification of the rationale for paternalism serves as a natural benchmark, uncomplicated by incentive problems and descriptively accurate in the event that those in power are benevolent and well-informed about economics.

There are actually four risk-taking distributions used in the difference-in-differences computation, two unregulated distributions under the behavioral hypotheses θ_1 and θ_0, and two regulated distributions (i.e., with policy ϵ in place) under those same hypotheses. As mentioned already, the symbol $f_\theta(x)$ always denotes the unregulated distribution. Using a simple theory of how regulation affects behavior, the regulated distribution can be expressed in terms of it.

Under a condition derived below, the policy maker's preferences over population distributions of risk-taking simplifies to the following rule: the policy maker has a stronger rationale for paternalism (i.e., inclination to intervene) under hypothesis θ_1 relative to hypothesis θ_0 whenever there is more weight in the upper tail of $f_{\theta_1}(x)$ defined by the cut-off point $x = 1 - \epsilon$ than in $f_{\theta_0}(x)$. The point $x = 1 - \epsilon$ is referred to as the *maximum allowable risk threshold.* For example, a laissez faire policy maker with no taste for paternalism chooses paternalism parameter $\epsilon = 0$, which implies a maximum allowable risk threshold of 1. Accordingly, the policy maker is equally satisfied with every continuous risk-taking distribution on the unit interval because they all have zero weight under the upper tail cut off by $1 - \epsilon = 1$. In contrast, a maximally paternalistic policy maker regards any positive level of risk-taking as undesirable. For intermediate levels of preference for paternalism (i.e., $0 < \epsilon < 1$), the simplified preference structure implies that the policy maker has a stronger rationale for ϵ-level paternalism under hypothesis θ_1 (relative to hypothesis θ_0) if and only if $\Pr(x_{\theta_1} > 1 - \epsilon) > \Pr(x_{\theta_0} > 1 - \epsilon)$. The simplified preference structure can be rationalized by showing that the difference-in-differences measure described above reduces to a monotonic function of the *rate of excessive risk-taking* defined by:

$$\rho_\theta \equiv \Pr(x_\theta > 1 - \epsilon) = \int_{1-\epsilon}^{1} f_\theta(x)\mathrm{d}x. \tag{1}$$

2.3 Welfare

This section provides conditions under which a policy preference structure that ranks risk-taking distributions according to the weight in their upper tails can be rationalized by a social-welfare function.

Suppose $w(x)$ is a weighting function that nets out social costs generated by externalities or hypothesized psychological biases. Let $k(\epsilon)$ be an increasing function that represents the costs of implementing and enforcing the policy ϵ. The following social-welfare function provides a conventional tool for ranking combinations of behavioral (θ), policy (ϵ), social-cost ($w(x)$) and policy-cost ($k(\epsilon)$) parameters:

$$\mathrm{SW}_\theta(\epsilon) = \int_{0}^{1-\epsilon} w(x) f_\theta(x)\mathrm{d}x + w(1-\epsilon)\int_{1-\epsilon}^{1} f_\theta(x)\mathrm{d}x - k(\epsilon). \tag{2}$$

An assumption built into the social-welfare function (2) is that the policy-enforcement technology is effective in obtaining compliance, so that individuals who choose excessive risk in the pre-regulation case (i.e., the fraction of the population such that $x_\theta > 1 - \epsilon$) cluster at the maximum allowable risk threshold $x = 1 - \epsilon$ once regulation is imposed. Non-compliance costs are folded into $k(\epsilon)$. The first term of the social-welfare function sums net benefits over individuals in the interior of the acceptable risk-taking spectrum. The second term accounts for net benefits generated by the fraction of the population $\int_{1-\epsilon}^{1} f_\theta(x)\mathrm{d}x$ that clusters at the maximum allowable risk point after regulation is imposed. The third term subtracts implementation and administration costs $k(\epsilon)$. Imposing the normalization $k(0) = 0$, the social-welfare function evaluated at the minimum and maximum levels of paternalism yields $\mathrm{SW}_\theta(0) = \int_0^1 w(x) f_\theta(x)\mathrm{d}x$ and $SW_\theta(1) = w(0) - k(1)$.

Because policy implementation is costly, policy makers who seek to maximize $\mathrm{SW}_\theta(\epsilon)$ will impose risk-limiting regulation (i.e., set $\epsilon > 0$) only when net benefits as measured by $w(x)$ are decreasing. To see why (assuming differentiability), note that $\mathrm{SW}_\theta(\epsilon)$ is decreasing whenever $w(x)$ is increasing:

$$\mathrm{SW}'_\theta(\epsilon) = -w'(1-\epsilon) \int\limits_{1-\epsilon}^{1} f_\theta(x)\mathrm{d}x - k'(\epsilon). \tag{3}$$

Because the costs of implementing paternalism are increasing in ϵ, the second term, $-k'(\epsilon)$, is negative. If $w'(1-\epsilon) \geq 0$, then $\mathrm{SW}'_\theta(\epsilon) < 0$ and ϵ cannot be a social-welfare maximizer. Thus, if $\epsilon > 0$ is a social-welfare maximizer, it must be the case that $w'(1-\epsilon) < 0$.

We quantify the policy maker's *rationale for paternalism* at ϵ, denoted $\Delta\mathrm{SW}_\theta(\epsilon)$, as the gains in social welfare achieved by regulation relative to the zero-intervention social-welfare reference point:

$$\Delta\mathrm{SW}_\theta(\epsilon) \equiv \mathrm{SW}_\theta(\epsilon) - \mathrm{SW}_\theta(0) = \int\limits_{1-\epsilon}^{1} [w(1-\epsilon) - w(x)] f_\theta(x)\mathrm{d}x - k(\epsilon). \tag{4}$$

The integral sums the social savings attributable to the policy, and the final term subtracts its costs. The integrand is guaranteed to be positive if $w(x)$ is decreasing to the right of $x = 1 - \epsilon$, although the sign of the rationale for paternalism is indeterminate.

If the policy maker had reliable knowledge of $f_\theta(x)$, $w(x)$ and $k(x)$, then a necessary condition for rationalizing paternalism would be $\Delta\mathrm{SW}_\theta(\epsilon) > 0$. As mentioned before, the disadvantage of $\Delta\mathrm{SW}_\theta(\epsilon)$ is that it depends on parameters that cannot be determined without specifying the benefits and costs of risk-taking and the available technology for enforcement and administration. However, we demonstrate now that, for fixed ϵ, $\Delta\mathrm{SW}_\theta(\epsilon)$ is monotonically increasing in the much simpler upper-tail probability ρ_θ, which is independent of $w(x)$ and $k(x)$.

2.4 When is the rationale for paternalism monotonic in ρ_θ?

To state a precise condition under which $\Delta\mathrm{SW}_\theta(x)$ is monotonic in ρ_θ, we examine the derivative $\frac{\mathrm{d}}{\mathrm{d}\rho_\theta}\Delta\mathrm{SW}_\theta(x)|_\epsilon$, which measures the response of the rationale for paternalism (as defined in (4)) to a small increase in ρ_θ caused by a shift in the policy maker's behavioral hypothesis. The difference-in-differences calculation described earlier plays a key role because it provides a finite approximation for the derivative's numerator and shares its sign.

To increase ρ_θ by a small increment requires a perturbation of $f_\theta(x)$ that re-distributes a small slice from the left of $x = 1 - \epsilon$ to the right-hand upper tail. There are many such perturbations. We consider a special class of them under which the slice is defined by arbitrarily chosen points and re-distributed uniformly to the upper tail (i.e., between $x = 1 - \epsilon$ and $x = 1$). Let δ_0 and δ_1 denote arbitrarily chosen numbers such that $0 < \delta_0 < \delta_1 < 1 - \epsilon$. The numbers define an interval $[\delta_0, \delta_1]$ strictly to the left of $1 - \epsilon$ from which probability mass is to be removed and redistributed to the upper tail. The fraction of the population that is re-distributed in the perturbation is $\int_{\delta_0}^{\delta_1} f_\theta(x)\mathrm{d}x \approx \mathrm{d}\rho_\theta$, the denominator in our approximation of $\frac{\mathrm{d}}{\mathrm{d}\rho_\theta}\Delta\mathrm{SW}_\theta(x)|_\epsilon$. The perturbed risk-taking pdf is

$$\tilde{f}_\theta(x) = \begin{cases} f_\theta(x) & 0 < x < \delta_0,\ \delta_1 < x < 1 - \epsilon, \\ f_\theta(x) + \left[\int_{\delta_0}^{\delta_1} f_\theta(z)\mathrm{d}z/\epsilon\right] & 1 - \epsilon < x < 1, \\ 0 & \text{elsewhere.} \end{cases} \tag{5}$$

The term in brackets is the normalizing constant that re-distributes the removed mass uniformly on the upper tail of length ϵ. Recall that the un-perturbed pre-regulation social-welfare function was given by $\mathrm{SW}_\theta(0) = \int_0^1 w(x) f_\theta(x)\mathrm{d}x$. In contrast, the perturbed pre-regulation social-welfare function is:

$$\tilde{\mathrm{SW}}_\theta(0) = \int_0^{\delta_0} w(x) f_\theta(x)\mathrm{d}x + \int_{\delta_1}^{1} w(x) f_\theta(x)\mathrm{d}x + \left[\int_{\delta_0}^{\delta_1} f_\theta(x)\mathrm{d}x/\epsilon\right] \int_{1-\epsilon}^{1} w(x)\mathrm{d}x. \tag{6}$$

After imposing the risk-limiting regulatory policy ϵ, all those whose pre-regulation actions were in the upper tail are, analogous to the un-perturbed case, assumed to move to $x = 1 - \epsilon$. In the perturbed case, those clustering at $x = 1 - \epsilon$ include the re-distributed population (moved from $\delta_0 < x < \delta_1$ to $1 - \epsilon < x < 1$ under perturbation, and then to the point $x = 1 - \epsilon$ by the policy). Thus, the perturbed social-welfare function evaluated at ϵ is:

$$\begin{aligned} \tilde{\mathrm{SW}}_\theta(\epsilon) &= \int_0^{\delta_0} w(x) f_\theta(x)\mathrm{d}x + \int_{\delta_1}^{1-\epsilon} w(x) f_\theta(x)\mathrm{d}x \\ &\quad + w(1-\epsilon)\left[\int_{1-\epsilon}^{1} f_\theta(x)\mathrm{d}x + \int_{\delta_0}^{\delta_1} f_\theta(x)\mathrm{d}x\right] - k(\epsilon). \end{aligned} \tag{7}$$

The difference of $\tilde{\mathrm{SW}}_\theta(\epsilon)$ and $\tilde{\mathrm{SW}}_\theta(0)$ yields the perturbed rationale for paternalism:

$$\tilde{\mathrm{SW}}_\theta(\epsilon) - \tilde{\mathrm{SW}}_\theta(0) = \int_{1-\epsilon}^{1} [w(1-\epsilon) - w(x)] f_\theta(x)\mathrm{d}x - k(\epsilon)$$
$$+\left[w(1-\epsilon) - \int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon\right] \int_{\delta_0}^{\delta_1} f_\theta(x)\mathrm{d}x \qquad (8)$$

The expression above is the rationale for paternalism computed by the policy maker after adopting an alternative behavioral theory with a slightly heavier upper tail. The first integral is the sum of social savings attributable to the policy based on the un-perturbed distribution just as before (c.f., Eq. (4)). The second bracketed expression is the average social savings attributable to the policy among individuals who were re-distributed to the upper tail by the perturbation. Under regulation, re-distributed individuals cluster at $x = 1 - \epsilon$ and are assigned weight $w(1 - \epsilon)$. The average pre-regulation weight assigned to them was $\int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon$.

The difference-in-differences calculation measuring the response of the rationale for paternalism with respect to a small change in ρ_θ while holding ϵ constant is:

$$\mathrm{d}\Delta\mathrm{SW}_\theta(\epsilon)|_\epsilon \approx \left[\tilde{\mathrm{SW}}_\theta(\epsilon) - \tilde{\mathrm{SW}}_\theta(0)\right] - [\mathrm{SW}_\theta(\epsilon) - \mathrm{SW}_\theta(0)]$$
$$= \left[w(1-\epsilon) - \int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon\right] \int_{\delta_0}^{\delta_1} f_\theta(x)\mathrm{d}x. \qquad (9)$$

The expression provides the numerator for the approximation of the derivative. The derivative can now be computed as:

$$\frac{\mathrm{d}}{\mathrm{d}\rho_\theta}\Delta\mathrm{SW}_\theta(\epsilon)|_\epsilon = w(1-\epsilon) - \int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon. \qquad (10)$$

Result 1: Provided $w(1 - \epsilon) - \int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon > 0$, the social-welfare rationale for paternalism $\Delta\mathrm{SW}_\theta(\epsilon)$ is monotonically increasing in the rate of excessive risk-taking ρ_θ. Thus, as measured by $\mathrm{SW}_\theta(\epsilon)$, paternalistic policies ϵ, $0 < \epsilon < 1$, become more compelling the greater the proportion of the population that violates the maximum allowable risk threshold $x = 1 - \epsilon$.

One obvious case in which the sign of the derivative is positive and monotonicity therefore holds is when $w(x)$ is decreasing to the right of $1 - \epsilon$, as would be the case when social or psychological costs accumulate faster than the individual benefits of risk-taking at levels above the maximum allowable risk threshold. Result 1 provides a social-welfare rationalization for relying on ρ_θ as a parsimonious proxy for the policy maker's inclination toward paternalism – parsimonious in the sense that ρ_θ is independent of context-specific social-welfare weights and policy costs. When considering two behavioral hypotheses, the policy maker's willingness to incur costs to implement the proposed restriction ϵ will be greater under the hypothesis that predicts the heavier upper tail. This does not imply that ϵ will necessarily be worth its price, $k(\epsilon)$, according to either theory. To answer the more difficult question of whether implementation should actually occur, knowledge of $w(x)$ and $k(\epsilon)$ is required. The advantage of the difference-in-differences metric in Eq. (9) is that it provides a monotonic transformation of the change in willingness-to-pay for a proposed restriction on risk-taking under distinct behavioral hypotheses, without requiring all parameters needed to specify the social-welfare function. Thus, one may investigate whether hypothesis θ_0 or θ_1 leads to a greater willingness-to-pay for ϵ simply by examining the upper-tail probabilities of the unregulated risk distributions, ρ_{θ_0} and ρ_{θ_1}. The rate of excessive risk-taking ρ_θ provides a means of ranking behavioral hypotheses according to the degree to which they favor risk-limiting paternalism without fully specifying the social-welfare function.

A limitation of the social-welfare perturbation as modeled above is that the cost of policy $k(\epsilon)$ is assumed constant with respect to changes in ρ_θ. This may be reasonable for small changes in ρ_θ and for some policies whose costs are relatively insensitive to individual risk-taking within fairly wide bounds (e.g., the cost of lifeguards and markers for safe swimming areas at public beaches). One could also generalize the social-welfare function by allowing $k(\epsilon)$ to respond positively to perturbations of ρ_θ. The relevant condition for monotonicity would then require that $[w(1-\epsilon) - \int_{1-\epsilon}^{1} w(x)\mathrm{d}x/\epsilon]$ is large relative to $\frac{\mathrm{d}}{\mathrm{d}\rho_\theta}k(\epsilon)$. This would once again require knowledge of difficult-to-measure social-welfare parameters, however, and we do not pursue the generalization here.

2.5 Endogenizing ϵ

This section presents an example featuring highly stylized functional forms with which it is straightforward to endogenize the policy maker's choice of ϵ in terms of exogenous social-cost and policy-cost parameters. Suppose the weighting function and policy cost functions are quadratic:

$$w(x) = w_0 - cx^2/2, \quad w_0 > 0, \ c > 0, \ \text{and} \quad k(x) = \kappa\epsilon^2/2, \quad \kappa > 0. \tag{11}$$

Assume that the policy maker adopts a uniform prior on x_θ, so that $f_\theta(x) = 1$ for $0 \le x \le 1$, and 0 elsewhere. In this case, the social-welfare function (2) takes on the form:

$$\mathrm{SW}_\theta(\epsilon) = w_0 - c(1-\epsilon)^2(1+2\epsilon)/6 - \kappa\epsilon^2/2. \tag{12}$$

The first and second derivatives of (12) with respect to ϵ are:

$$\mathrm{SW}'_\theta(\epsilon) = c(1-\epsilon)\epsilon - \kappa\epsilon, \quad \text{and} \quad \mathrm{SW}''_\theta(\epsilon) = c(1-2\epsilon) - \kappa. \tag{13}$$

Solving the first-order condition for a social-welfare-maximizing level of paternalism on the interior of the unit interval yields the formula:

$$\epsilon^* = 1 - \frac{\kappa}{c}. \tag{14}$$

The endpoints $\epsilon = 0$ and $\epsilon = 1$ must be checked to find out when ϵ^* is the unique global maximizer. The maximum possible level of paternalism $\epsilon = 1$ can be ruled out because $\mathrm{SW}'_\theta(1) = -\kappa$, implying that $\mathrm{SW}_\theta(\epsilon)$ is decreasing at $\epsilon = 1$ and therefore that social welfare can always be increased by reducing paternalism. The expressions in (13) show that the first derivative of the social-welfare function is zero at both $\epsilon = 0$ and $\epsilon = \epsilon^*$, and that the second derivative evaluated at those points has opposite signs: $SW''_\theta(0) = c - \kappa$ and $SW''_\theta(\epsilon^*) = \kappa - c$. Thus, the global social-welfare maximizer is ϵ^* if $c > \kappa$, and zero otherwise. This condition is intuitive because it requires that the social costs (therefore, the social savings attributable to the regulation) are large relative to the policy's costs as a necessary condition for positive levels of paternalism. The formula for ϵ^* shows that, as one would expect, the optimal level of paternalism is decreasing in the policy-cost parameter κ and increasing in the social-cost-of-risk-taking parameter c that scales the social savings from paternalism. The optimal level of paternalism approaches absolute abolishment of risk-taking as social costs grow infinitely larger than policy costs.

3 A descriptive model of prescriptive behavioral modeling

Previous sections provided definitions of the paternalism parameter, the maximum allowable risk threshold, the policy maker's notion of excessive risk, and the upper-tail probability ρ_θ, referred to as the rate of excessive risk-taking. Previous sections also provided a social-welfare rationalization for quantifying the rationale for paternalism in terms of ρ_θ. Specifically, we modeled the policy maker's comparison of two behavioral hypotheses θ_0 and θ_1 in relative terms using the difference $(\rho_{\theta_0} - \rho_{\theta_1})$ as a proxy for the fully specified social-welfare equivalent $(\Delta\mathrm{SW}_{\theta_0}(\epsilon) - \Delta\mathrm{SW}_{\theta_1}(\epsilon))$.

The remainder of the paper makes extensive use of this result to demonstrate that departures from expected-utility maximization produce changes in the rationale for paternalism that have an indeterminate sign. Given a policy proposal ϵ and returns-generating process γ, we compute rates of excessive risk-taking for the expected-utility-maximization hypothesis (ρ_α) and the satisficing hypothesis (ρ_β) to reveal that behavioral-hypothesis-induced changes in ρ_θ may be either negative or positive. The sign indeterminacy of $(\rho_\alpha - \rho_\beta)$ turns out to be generic rather than special, with departures from maximization decreasing the rationale for paternalism in roughly 40% of cases (i.e., among possible combinations of ϵ and γ). Thus, failure to maximize does not imply an increased rationale for paternalism.

3.1 Flows of risk-taking opportunities

Technological innovation and institutional evolution generate an ongoing sequence of opportunities for individuals to take new forms of risk in pursuit of anticipated benefits. Innovation in medical science is one source of such opportunities (e.g., the advent of heart transplantation technology, the increasing availability of diagnostic tests for disease, and the expanding menu of cosmetic medical services). Cultural innovation produces a virtually continuous, although not steady, flow of fashion and lifestyle choices that entail risk-return trade-offs (e.g., extreme sports, new consumer products, new types of illicit drugs, and novel forms of coupling activity and sexual behavior). Financial institutions, both private and regulatory, represent another important channel through which new opportunities for risk-taking flow (e.g., bank deregulation, the introduction of new financial products, and technological innovations of the kind that enabled widespread online equity trading).

The unfamiliarity of newly arrived risks imposes bounds on what individuals can know about the shape of the relevant risk distributions. Novel risks may even have unknown event spaces. Without assuming that individuals see the space of possible outcomes or understand the shape of the returns-generating process, we model the newly arrived risk by the simple gamble $R(x)$ whose distribution depends on the individual's choice of risk-level $x \in [0, 1]$:

$$R(x) = \begin{cases} x^{\gamma} + x & \text{with probability } 1/2 \\ x^{\gamma} - x & \text{with probability } 1/2, \end{cases} \tag{15}$$

where $0 < \gamma < 2$. The parameter γ indicates whether the marginal expected return with respect to x is increasing ($\gamma > 1$), linear ($\gamma = 1$), or decreasing ($\gamma < 1$). Thus, γ summarizes the shape of the risk-return environment. The distribution $R(x)$ has mean x^{γ} and standard deviation x, although these should not automatically be interpreted as appropriate measures of subjective reward and risk without specifying a behavioral theory, such as expected-utility maximization, for which they are relevant statistics.

3.2 Paternalism and novel risk

Novel forms of risk are necessary for economic growth. At the same time, they are among the most difficult to analyze for the purpose of choosing a reasonable regulatory approach. It is usually unclear whether forward-looking behavior and self-correcting feedback brought about by market competition provide sufficient safeguards against the downside risks for which policy makers are held responsible. Even the most sophisticated statistical tools provide only limited insight when new risks with unknown frequencies and unknowable event spaces are concerned. Novel risks therefore imply nontrivial questions about the desirability of paternalism.

Concerning issues such as genetically modified food, nuclear power, and global warming, some argue that private incentives lead small groups to dishonestly generate exaggerated fears. At the same time, others argue that private incentives lead to the suppression of important evidence about risks. It may also be the case that the available data are fundamentally ambiguous and cannot possibly provide a factual basis for regulatory consensus. When the empirical evidence is weak, few constraints bound even the most rigorous attempts to apply scientific reasoning to policy design. Consequently, there are many degrees of freedom in formulating logically coherent approaches to policy. The policy maker's choice of behavioral hypotheses in making predictions about the effects of different policies and their benefits and costs is an important variable that can lead to markedly different decisions about regulatory approach. The enlarged set of behavioral hypotheses recently entering into economics from psychology, which includes numerous alternatives to expected-utility maximization, therefore represents an interesting new source of variation in economic arguments for different levels of paternalism.

3.3 Two behavioral hypotheses

In this section, we introduce two specific behavioral hypotheses, labeled α and β, which give rise to distributions of risk-taking behavior, denoted x_a and x_b, both with support [0, 1]. The hypotheses correspond to two idealized models of behavior, one psychological and boundedly rational, and the other economic and unboundedly rational. This set-up would apply, for example, to a policy maker who consults with two sets of economic advisors, the first of which offers advice based on psychological models drawn from behavioral economics, and the second of which bases its advice on standard expected-utility theory.

Hypothesis α: Individuals satisfice aspirations without possessing or requiring global knowledge of the relationship between risk-taking and expected returns. In particular, individuals have no beliefs about $\mathrm{E}[R(x)]$, no knowledge of the support of $R(x)$, and no subjective probabilities concerning possible values that $R(x)$ might take on. Satisficers search along the x dimension starting from the point $x = 0$ since the risky opportunity did not previously exist and, prior to its existence, risk-taking (in this particular domain) was necessarily zero. They are assumed to search upward along the risk spectrum, sampling returns by observation until aspiration levels are satisficed in expectation. Heterogeneity in the population derives from heterogeneous aspirations. An aspiration distribution describes the population's heterogeneity and a random draw from this distribution is denoted a. Hypothesis α represents one version of bounded rationality, or psychological man, as hypothesized by Simon (1982). Aspiration-seekers are referred to as satisficers, or α-types.

Hypothesis β: Individuals possess global knowledge of the risk-reward relationship and the frequency distribution associated with every possible choice of x. Individuals share a common mean–variance objective function that differs only by the risk-aversion parameter. Population heterogeneity derives solely from heterogeneous degrees of risk aversion, whose distribution is represented by b, which denotes a randomly drawn individual's risk-aversion parameter. Individuals know their own risk-aversion parameter. The unbounded rationality of β-types is reflected by their objectively correct beliefs, unlimited capacity to optimize, and the fact that their behavior results from a systematic weighting of anticipated benefits and costs across all feasible alternatives.

It is commonly assumed that societies of β-types are better off in an objective social-welfare sense and that normative decision theory should unequivocally aim to encourage people to become more β-like. We demonstrate, however, that societies of β-types are often more difficult to manage, requiring a greater degree of paternalistic intervention than societies of α-types.

3.4 Heterogeneity in aspirations and risk-aversion

Individual risk-taking in the real world is richly heterogeneous. Some try to avoid risks that others eagerly pursue. Most choose intermediate degrees of risk, insuring against some contingencies while betting on others. Both α- and β-hypotheses allow for such heterogeneity of risk-taking behavior. We assume that aspiration-seekers' aspirations and EU-maximizers' inverse risk aversion parameters are uniformly distributed on the unit interval:

$$a \sim U[0,1] \quad \text{and} \quad \frac{1}{b} \sim U[0,1]. \tag{16}$$

The assumption implies that both aspirations and willingness to bear risk have finite upper bounds.

3.5 Decision rules

Recall from hypothesis α that an α-type with aspiration a begins at $x = 0$ and increases x until the aspiration condition $\mathrm{E}[R(x)] = a$ is met. No beliefs about the expected return or prior knowledge are required. The α-type simply observes other individuals' returns at each position along the search path, obtaining arbitrarily good estimates of average returns at each position, and stops search as soon as a good-enough average return is discovered. Solving $\mathrm{E}[R(x)] = a$ for x yields the α-type's decision rule:

$$x_\alpha = a^{\frac{1}{\gamma}}. \tag{17}$$

In contrast, a β-type with risk-aversion parameter b is assumed to see the entire risk-return schedule $R(x)$ instantaneously, with precise knowledge of its mean and variance, and choose x on the closed unit interval to maximize the expected-utility function:

$$u(x) = x^\gamma - \frac{bx^2}{2}. \tag{18}$$

The first- and second-order conditions for an interior local maximum are:

$$u'(x) = \gamma x^{\gamma-1} - bx = 0, \quad \text{and} \quad u''(x) = \gamma(\gamma-1)x^{\gamma-2} - b < 0. \tag{19}$$

What is needed is an expression for the global maximizer that describes the behavior of a randomly drawn β-type in terms of the risk-aversion parameter b and returns-generating parameter γ. The first-order condition has two solutions, 0 and x^*, defined as:

$$x^* \equiv \left(\frac{\gamma}{b}\right)^{\frac{1}{2-\gamma}}. \tag{20}$$

It turns out that $x = 0$ is never optimal because $u(0) = 0$ is always dominated by $u(x^*) = (\frac{\gamma}{b})^{\frac{\gamma}{2-\gamma}}(1 - \gamma/2)$ or $u(1) = 1 - b/2$. For $\gamma < b$, x^* is in the admissible range (i.e., the interior of the unit interval) and the objective function is decreasing at $x = 1$: $u'(1) = \gamma - b < 0$. Therefore, for $\gamma < b$, the global maximizer must be x^*. For $\gamma > b$, x^* is out of the admissible range (i.e., $x^* > 1$) and $u(1) > u(0)$, implying that $x = 1$ is the global maximizer. Putting these results together, the β-type's decision rule can be expressed as:

$$x_\beta = \min\{x^*, 1\}. \tag{21}$$

We note that there is one special case, $\gamma = 1$, in which psychological and economic hypotheses cannot be distinguished using aggregate-level statistics because x_α and x_β have the same distribution. The distributions are always distinct, however, for environments with $\gamma \neq 1$.

Result 2: For $\gamma = 1$:

$$x_\alpha = a, \quad x_\beta = \frac{1}{b}, \quad \text{and} \quad x_\alpha \sim x_\beta. \tag{22}$$

3.6 Rationales for paternalism

Based on the earlier argument that rates of excessive risk-taking serve as a parsimonious proxy for the rationale for paternalism, we compute ρ_α and ρ_β to reveal whether satisficing or expected-utility maximization leads to a greater rationale for paternalism at level ϵ. Subsequent sections consider how the sign of $\rho_\alpha - \rho_\beta$ depends on policy and environment parameters ϵ and γ.

Using the decision rule (17) and the assumption that a is uniformly distributed on the unit interval, the α-type population's rate of excessive risk-taking is computed as:

$$\rho_\alpha \equiv \Pr\left[x_\alpha > 1-\epsilon\right] = \Pr\left[a^{\frac{1}{\gamma}} > 1-\epsilon\right] = \Pr\left[a > (1-\epsilon)^\gamma\right] = 1-(1-\epsilon)^\gamma. \tag{23}$$

The computation of the rate of excessive risk-taking among β-types must take into account clustering at the maximum-risk point $x = 1$, which occurs whenever $\gamma > b$. Making use of the assumption that $\frac{1}{b}$ is uniformly distributed on the unit interval, the β-type population's rate of excessive risk-taking is:

$$\begin{aligned}\rho_\beta &\equiv \Pr\left[x_\beta > 1-\epsilon\right] = \Pr\left[\min\left\{\left(\frac{\gamma}{b}\right)^{\frac{1}{2-\gamma}}, 1\right\} > 1-\epsilon\right]\\ &= \Pr\left[\left(\frac{\gamma}{b}\right)^{\frac{1}{2-\gamma}} > 1-\epsilon\right]\\ &= 1 - \min\left\{(1-\epsilon)^{2-\gamma}\frac{1}{\gamma}, 1\right\}.\end{aligned} \tag{24}$$

As one would expect, after integrating over sources of individual heterogeneity, a and b, the rationales for paternalism as proxied by ρ_α and ρ_β depend on only two factors: the policy maker's maximum allowable risk threshold and the shape of the returns-generating environment.

3.7 Determinants of $\rho_\alpha > \rho_\beta$ versus $\rho_\alpha < \rho_\beta$

Table 1 presents rates of excessive risk-taking for different combinations of excessive risk thresholds $(1-\epsilon)$ and environments (γ). Table 1 clearly shows that, for every policy objective, either inequality may occur: $\rho_\alpha > \rho_\beta$ or $\rho_\alpha < \rho_\beta$. According to the first row of Table 1, a combination of a high preference for paternalism (only 10 percent of the risk spectrum is acceptable at $1-\epsilon = 0.10$) and rapidly decreasing marginal returns $(\gamma = 0.25)$ leads to rates of excessive risk-taking that are more than twice as severe under the expected-utility-maximization hypothesis as under the satisficing hypothesis (93 as opposed to 44%). Given the same preference for paternalism $(1-\epsilon = 0.10)$ and an increasing returns environment such

Table 1 Rates of excessive risk-taking among satisficers and expected utility maximizers (ρ_α and ρ_β) as a function of the policy objective ($1 - \epsilon$) and environment (γ)

Excessive risk threshold[a]	Environment type (shape of return risk schedule)	Rate of excessive risk taking	
		alpha types	beta types
$1 - \epsilon$	γ	ρ_α	ρ_β
0.10	0.25	0.44	0.93
0.10	0.50	0.68	0.94
0.10	0.75	0.82	0.93
0.10	1.00	0.90	0.90
0.10	1.25	0.94	0.86
0.10	1.50	0.97	0.79
0.10	1.75	0.98	0.68
0.25	0.25	0.29	0.65
0.25	0.50	0.50	0.75
0.25	0.75	0.65	0.76
0.25	1.00	0.75	0.75
0.25	1.25	0.82	0.72
0.25	1.50	0.88	0.67
0.25	1.75	0.91	0.60
0.50	0.25	0.16	0
0.50	0.50	0.29	0.29
0.50	0.75	0.41	0.44
0.50	1.00	0.50	0.50
0.50	1.25	0.58	0.52
0.50	1.50	0.65	0.53
0.50	1.75	0.70	0.52
0.75	0.25	0.07	0
0.75	0.50	0.13	0
0.75	0.75	0.19	0.07
0.75	1.00	0.25	0.25
0.75	1.25	0.30	0.36
0.75	1.50	0.35	0.42
0.75	1.75	0.40	0.47
0.90	0.25	0.03	0
0.90	0.50	0.05	0
0.90	0.75	0.08	0
0.90	1.00	0.10	0.10
0.90	1.25	0.12	0.26
0.90	1.50	0.15	0.37
0.90	1.75	0.17	0.44

[a] The excessive risk-taking threshold is exogenously given and enjoys no special normative status in computations or interpretations

as $\gamma = 1.75$, however, the expected-utility-maximization hypothesis predicts a smaller rate of excessive risk-taking (68 compared to 98% among satisficers). The bottom block of Table 1 shows values of ρ_α and ρ_β corresponding to a far weaker degree of paternalism, a maximum allowable risk threshold of $1 - \epsilon = 0.90$. Again, we observe reversals in the sign of $\rho_\alpha - \rho_\beta$ depending on whether the marginal returns on risk-taking are increasing or decreasing.

Another feature visible in Table 1 is the fluctuating ranges of ρ_α and ρ_β. For example, when the policy is $1 - \epsilon = 0.25$, the dispersion of ρ_α is larger than that of ρ_β, with ranges of 29–91 versus 60–76%. The relative dispersions are reversed for $1-\epsilon = 0.90$, with ranges of 3–17 for ρ_α versus 0–44% for ρ_β. In general, the magnitudes of the percentage-point differences in rates of excessive risk-taking are large, clearly enough to sway a marginal policy maker toward or away from paternalism depending on which behavioral hypothesis is adopted. Also worth observing is the non-monotonicity of ρ_β with respect to γ in the second and third blocks of Table 1 (corresponding to $1-\epsilon = 0.25$ and 0.50, respectively). Non-monotonicity complicates attempts to generalize about the determinants of the magnitude of $\rho_\alpha - \rho_\beta$.

We can generalize about the factors that determine the sign of $\rho_\alpha - \rho_\beta$. Figure 1 plots all combinations of γ and $1-\epsilon$ for which $\rho_\alpha > \rho_\beta$. The shaded regions therefore indicate environment-and-policy combinations for which hypothesized departures from maximization lead to a greater rationale for paternalism. The unshaded region, in contrast, indicates environments and policies for which psychological hypotheses decrease the rationale for risk-limiting intervention, implying that individual maximization reduces aggregate economic performance according to the social-welfare objective. The four regions of Fig. 1 are defined by the curves $1-\epsilon = \gamma^{\frac{1}{2(1-\gamma)}}$ and $\gamma = 1$. Rates of excessive risk-taking are equal along these boundaries. The Appendix provides additional details on the derivation of Fig. 1 and its boundaries.

The policy parameter is constant along every horizontal line through Fig. 1. A key observation about Fig. 1 is that every iso-policy horizontal intersects both shaded and unshaded regions, covering one region where $\rho_\alpha > \rho_\beta$ and another where $\rho_\alpha < \rho_\beta$. The single exception is the iso-policy horizontal $1-\epsilon = e^{-1/2}$ along which the inequality $\rho_\alpha \geq \rho_\beta$ uniformly holds. (One can show using

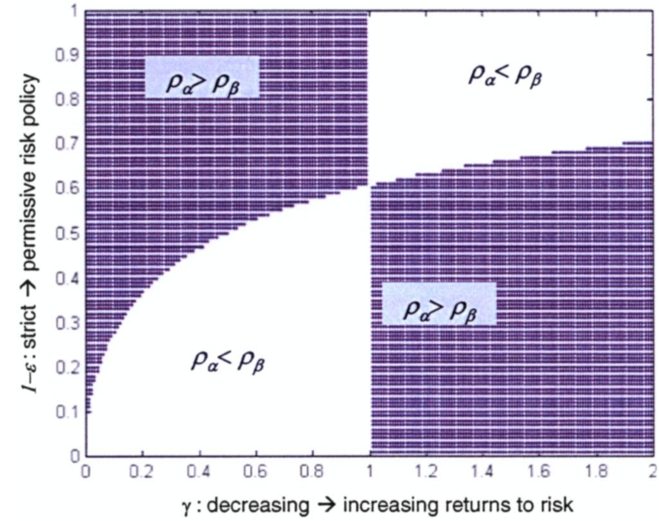

Fig. 1 Shaded region depicts environment-policy pairs $(\gamma, 1-\epsilon)$ for which the hypothesis of satisficing (α) provides a greater rationale for paternalism $(\rho_\alpha > \rho_\beta)$. The *unshaded region* depicts $(\gamma, 1-\epsilon)$ pairs for which expected utility maximizers provide a larger rationale for paternalism

L'Hospital's Rule that $\lim_{\gamma \to 1} \gamma^{\frac{1}{2(1-\gamma)}} = \mathrm{e}^{-1/2}$ and that the intersection of the interior boundaries of Fig. 1 occurs at the point $(\gamma, 1-\epsilon) = (1, \mathrm{e}^{-1/2})$.) Thus, given virtually any policy objective, there is a dense set of environments for which psychological man provides a smaller rationale for paternalism than economic man. The rectangle depicted in Fig. 1 (with an area of 2) represents all possible pairs of $(\gamma, 1-\epsilon)$, and the fraction of those pairs for which psychology leads to a smaller rationale for paternalism can be computed as $\left[\int_0^1 \gamma^{\frac{1}{2(1-\gamma)}} d\gamma + \int_1^2 (1-\gamma^{\frac{1}{2(1-\gamma)}}) \mathrm{d}\gamma\right]/2 \approx 0.40$.

Result 3: For every policy objective $1-\epsilon \neq \mathrm{e}^{-1/2}$, there exist two dense sets of environments, one with $\rho_\alpha > \rho_\beta$ and the other with $\rho_\alpha < \rho_\beta$.

3.8 Taxing risk

Consider a unit tax on risk-taking at the rate $\tau < 1$. The after-tax return for an individual who chooses risk-taking x is $R(x) - \tau x$, and the after-tax decision rules are:

$$x_\alpha = [\frac{a}{1-\tau}]^{\frac{1}{\gamma}}, \quad \text{and} \quad x_\beta = \min\left\{\left[\frac{\gamma(1-\tau)}{b}\right]^{\frac{1}{2-\gamma}}, 1\right\}. \tag{25}$$

Because x_α conditional on a is increasing in τ for all a, the upper tail ρ_α must also be increasing in τ. In contrast, x_β is non-increasing in τ and so too is ρ_β.

Result 4: For positive tax rates on risk-taking, satisficers choose more risk than they would under a zero-tax regime, attempting to reach internally fixed aspirations at reduced after-tax expected returns. Expected utility maximizers and satisficers have qualitatively opposite responses to the tax.

Similar results follow for other kinds of taxes such as user fees and lump-sum taxes. For example, imposing a user fee $\phi > 0$ on any individual who chooses a strictly positive level of risk-taking results in an after-tax return of $R(x) - \phi$ if $x > 0$, and 0 if $x = 0$. The resulting decision rules under the user-fee regime are:

$$x_\alpha = (a+\phi)^{\frac{1}{\gamma}} \text{ and } x_\beta = \begin{cases} x^{**} \equiv \min\left\{\left(\frac{\gamma}{b}\right)^{\frac{1}{2-\gamma}}, 1\right\} & \text{if } (x^{**})^\gamma - \frac{b}{2}(x^{**})^2 - \phi > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

The first formula indicates that α-types take on more risk in response to a positive user fee and ρ_α increases as a result. The second formula implies that ρ_β is non-increasing in ϕ.

4 Discussion

This paper attempts to model a policy maker's reasoning process about whether to impose paternalistic regulation that limits individual risk-taking. Given a proposed maximum allowable risk-taking threshold, the rationale for the paternalistic proposal is defined as the difference in social welfare post- and pre-regulation. We provide a condition under which this social-welfare difference is monotonic in the upper tail of the risk-taking distribution cut off by the policy maker's maximum allowable risk threshold. Monotonicity allows one to examine how various behavioral hypotheses affect the social-welfare difference by studying their effects on the much simpler rate of excessive risk-taking. The simplification provides a means of ranking the rationales for paternalism associated with different behavioral hypotheses without estimating difficult-to-measure parameters in the social-welfare function, such as those that scale the social costs of risk-taking and administrative costs of the policy. According to the model, a change in behavioral hypotheses changes the rationale for paternalism if and only if it makes distinct predictions about the upper tail of the risk-taking distribution and its response to interventions such as bans or taxes. The policy maker is more inclined to accept a paternalistic proposal (i.e., more willing to incur policy costs) under behavioral hypotheses that imply a heaver upper tail. The intuition for this is straightforward: the more severe excessive risk-taking is without regulation, the greater the social savings from enforcing a maximum allowable risk threshold is.

The paper's main result is that psychological departures from expected-utility maximization may either increase or decrease the policy maker's rationale for paternalism. For every policy objective, there exist two dense sets of environments: on the first set, satisficing implies a greater rationale for paternalism than expected-utility maximization does, and on the second set the inequality is reversed. The key point is that departures from maximization can reduce the costs of externalities and make laissez faire policy more attractive.

Despite the monotonicity result's motivation, which was to avoid having to fully specify the social-welfare function, the indeterminacy result described in the preceding paragraph underscores the need for context-specific policy analysis rather than generalizations based on ideology. The generalization we argue most vehemently against is the claim, now common in the behavioral economics literature, that when individuals fail to conform to the usual normative decision-making axioms, government should try to change individual behavior so that it more closely conforms. The model shows that for a large subset of environments, departures from maximization improve social welfare. Efforts to bring individuals into conformity with content-blind normative axioms (i.e., those that are applied invariantly across all decision-making domains) such as those based on formal logic and probability theory (e.g., the Savage axioms underlying the expected-utility hypothesis) may hurt the economy. Thus, conformity with content-blind norms is not an appropriate policy-making objective in itself. Rather, full specification of the social-welfare function is needed in each specific instance to provide a context-sensitive tally of the benefits and costs that determine whether departures from decision-making norms are harmful, helpful or benign.

A counterintuitive implication of the satisficing hypothesis is that a tax on risk-taking induces greater risk-taking. The result depends on the fact that aspiration levels are fixed rather than fast-adjusting. When risk-taking is taxed, satisficers take on extra risk so that their after-tax expected returns reach the old aspiration level. We describe three examples in which the counterintuitive satisficing response to taxation seems plausible.

Example 1 (Drugs and sex): The addiction literature frequently finds that aggregate consumption of alcohol and illicit drugs is price insensitive and that consumption takes on more risky forms when availability is restricted (McGeorge and Aitken 1997; Österberg 1990; Popham et al. 1972). The satisficing hypothesis would therefore seem a worthwhile consideration in trying to anticipate the effects of proposals aimed at dealing with alcohol and drug consumption. The satisficing hypothesis again seems relevant in debates about sexual risk-taking, where proposed interventions seek to reduce sexually transmitted disease, discourage teenage pregnancy, or expand opportunities for women in labor markets. Insofar as individuals possess relatively fixed aspiration levels for sexual satisfaction, satisficing theory suggests that sanctions against sexual risk-taking would lead to additional risk-taking, which is the opposite of the intended effect. Aspiration seekers would respond to easy availability of low-risk forms of sexual gratification by reducing risky behavior. The policy in Germany and a handful of European countries of permitting soft pornography to be broadcast over publicly owned frequencies could, for example, be understood as a risk-deterring subsidy of sexual gratification. The interpretation depends on the assumption that a fraction of gratification seekers are satisficed by pornographic broadcasts and consequently forego riskier forms of consumption.

Example 2 (Social safety nets and insufficient risk-taking): The argument is sometimes put forward that government provision of social safety nets in the form of medical, unemployment, and pension insurance encourages entrepreneurship. If the next best alternative to self-employment is working for a large firm that provides those benefits, the opportunity cost of starting a new business would be smaller the more extensive the safety net is. Much has been written on entrepreneurship and self-employment and we only wish to point out that certain reduced-form correlations, such as the highly entrepreneurial cultures of China and the U.S. compared to those in countries with more extensive social safety nets, seem to match the qualitative predictions of a naive satisficing model. Thus, it could be that reductions in risk provided by safety nets do not promote risk-taking as much as high aspirations for wealth do.

Example 3 (Retirement saving): Given that financial risks are difficult to quantify or even define, it may be reasonable for retirement savers to aim for expected-return targets rather than maximize a mean-variance objective. There is evidence that labor-supply decisions sometimes follow a target-satisficing process (Camerer et al. 1997).[6] If so, policies designed to promote financial risk-taking, for example in transition economies whose publics have relatively little investment experience, would need to consider the contrasting implications of satisficing versus expected-utility maximization hypotheses. Subsidies such as tax-free savings accounts and defined-benefit pension guarantees in a population of satisficers could lead to reductions in aggregate financial risk-taking, because those policies satisfice savers at lower levels of risk-taking.

A limitation of the model presented in this paper is its consideration of only two behavioral hypotheses. As researchers seek psychological explanations for

[6] Doubts have been put forward concerning the negative wage elasticity of cab drivers' labor supply (e.g., Farber (2003) and those cited in Goette et al. 2004). The finding has been replicated, however, in Singapore (Chou 2002). Satisficing wage laborers have been observed independently in other settings, such as Weber's (1958, pp. 59–60) account of "pre-capitalist" agricultural workers who reportedly worked fewer hours at high-wage harvest times.

social and individual-level decision-making phenomena they observe, satisficing is only one among many directions to consider, although Simon (1982) argued that it should be a prominent one. Another limitation of the model is that risk-taking has no formal price. This feature of the model may apply literally to thrill-seeking opportunities in public places, such as swimming on public beaches and rock climbing in public parks or hunting and foraging opportunities in our evolutionary past. For goods with positive prices, the risky return in the model should be interpreted as net surplus, the difference between reservation price and market price. Instead of forgone consumption or time costs, the reasons why individuals do not typically maximize returns are risk aversion among expected utility maximizers and lack of global knowledge about the structure of the environment among satisficers. The model abstracts from general equilibrium effects by focusing solely on a single-dimensional risky decision.

The model describes one mechanism through which behavioral hypotheses can influence a benevolent policy maker's willingness to adopt paternalistic regulation. Psychological models of man do not necessarily increase the rationale for paternalism because psychological or behavioral hypotheses do not necessarily imply losses in economic efficiency. Depending on the returns-generating environment and the policy maker's desired level of paternalism, the satisficing hypothesis may increase or decrease the apparent benefits of risk-limiting paternalistic intervention.

5 Appendix

Figure 1 partitions $(\gamma,\ 1-\epsilon)$-space into four regions characterized by the sign of $\rho_\alpha - \rho_\beta$. The interior boundaries occur at points where $\rho_\alpha = \rho_\beta$, or equivalently, where $\min\{(1-\epsilon)^{2-\gamma}/\gamma, 1\} = (1-\epsilon)^\gamma$. We list three conditions that exhaustively characterize points where $\rho_\alpha > \rho_\beta$ and simplify them with a single condition below.

To begin with, note that $\min\{(1-\epsilon)^{2-\gamma}/\gamma, 1\} = 1$ whenever $\gamma^{\frac{1}{2-\gamma}} < 1-\epsilon$, in which case $\rho_\beta = 0$ and $\rho_\alpha - \rho_\beta > 0$. Therefore:

$$\gamma^{\frac{1}{2-\gamma}} < 1-\epsilon \Rightarrow \rho_\alpha > \rho_\beta. \tag{27}$$

When $\min\{(1-\epsilon)^{2-\gamma}/\gamma, 1\} = (1-\epsilon)^{2-\gamma}/\gamma$, then $1-\epsilon < \gamma^{\frac{1}{2-\gamma}}$, and the inequality $\rho_\alpha > \rho_\beta$ requires $(1-\epsilon)^\gamma < (1-\epsilon)^{2-\gamma}/\gamma$. This inequality can be simplified to $\gamma < (1-\epsilon)^{2(1-\gamma)}$, but further simplification requires separate consideration of two cases. For $0 < \gamma < 1$, $\gamma < (1-\epsilon)^{2(1-\gamma)}$ is equivalent to $\gamma^{\frac{1}{2(1-\gamma)}} < (1-\epsilon)$, and therefore:

$$0 < \gamma < 1 \quad \text{and} \quad \gamma^{\frac{1}{2(1-\gamma)}} < 1-\epsilon < \gamma^{\frac{1}{2-\gamma}} \Rightarrow \rho_\alpha > \rho_\beta. \tag{28}$$

For $1 < \gamma < 2$, the exponent $2(1-\gamma)$ is negative and $\gamma < (1-\epsilon)^{2(1-\gamma)}$ is equivalent to $\gamma^{\frac{1}{2(1-\gamma)}} > 1-\epsilon$, so that:

$$1 < \gamma < 2 \quad \text{and} \quad 1-\epsilon < \gamma^{\frac{1}{2(1-\gamma)}} \Rightarrow \rho_\alpha > \rho_\beta. \tag{29}$$

It is straightforward to show that points covered by conditions (27), (28) and (29) are equivalent to the single condition:

$$\rho_\alpha > \rho_\beta \text{ iff } \left[0<\gamma<1 \text{ and } \gamma^{\frac{1}{2(1-\gamma)}} < 1-\epsilon\right] \text{ or } \left[1<\gamma<2 \text{ and } 1-\epsilon<\gamma^{\frac{1}{2(1-\gamma)}}\right]. \quad (30)$$

Acknowledgements We thank Alexis Benlianin and participants at the behavioral economics session of the VI Annual International Scientific Conference, State University-Higher School of Economics, Moscow, for lively discussions and valuable criticism.

References

Aaron HJ (1999) Behavioral dimensions of retirement economics. Brookings Institution Press

Arthur WB (1994) Inductive reasoning and bounded rationality: The El Farol problem. Am Econ Rev 84:406–411

Berg N, Lien D (2003) Tracking error decision rules and accumulated wealth. Appl Math Finance 10(2):91–119

Berg N, Lien D (2005) Does society benefit from investor overconfidence in the ability of financial market experts? J Econ Behav Organization 58:95–116

Bernheim BD, Rangel A (2006) Behavioral public economics: How to do welfare analysis when individuals can make mistakes. In: Diamond P, Vartiainen H (eds) Economic institutions and behavioral economics. Princeton University Press, Princeton (in press)

Bookstaber R, Langsam J (1985) On the optimality of coarse behavior rules. J Theor Biol 116:161–193

Buchanan JM, Wagner RE (1977) Democracy in deficit: the political legacy of Lord Keynes. Academic Press, New York

Caldwell B (2004) Introduction adapted from Hayeks challenge. Ama-Gi 6(2):16–18

Camerer C, Babcock L, Loewenstein G, Thaler R (1997) Labor supply of New York City cabdrivers: One day at a time. Q J Econ 112(2):407–441

Camerer C, Issacharoff S, Loewenstein G, O'Donoghue, Rabin M (2003) Regulation for conservatives: behavioral economics and the case for asymmetric paternalism. Univ Penn Law Rev 151:1211–1254

Caplan B (1999) Rational ignorance versus rational irrationality. Working Paper, George Mason University

Caplin A, Leahy J (2003) Behavioral policy. In: Brocas I, Carrillo J (eds) Essays in psychology and economics. Oxford University Press, Oxford

Choi JJ, Laibson D, Madrian BC, Metrick A (2005) Optimal defaults and active decisions, NBER Working Paper W11074

Chou YK (2002) Testing alternative models of labor supply: evidence from cabdrivers in Singapore. Singapore Econ Rev 47:1747

Cosmides L, Tooby J (1994) Better than rational: evolutionary psychology and the invisible hand. Am Econ Rev 84(2):327–332

Elster J (1992) Local justice: how institutions allocate scarce goods and necessary burdens. Sage, New York

Epstein R (1995) Simple rules for a complex world. Harvard University Press, Cambridge

Epstein R (2004) The optimal complexity of legal rules. Olin Working Paper 210, University of Chicago

Farber HS (2003) Is tomorrow another day? The labor supply of New York cabdrivers. NBER Working Paper 9706

Frey B, Stutzer A (2004) Economic consequences of mispredicting utility. IEW Working Paper 218, University of Zurich

Gigerenzer G, Todd PM, the ABC Research Group (1999) Simple heuristics that make us smart. Oxford University Press, New York

Goette L, Fehr E, Huffman D (2004) Loss aversion and labor supply. J Eur Econ Assoc 2:216–228

Gruber J (2002) Smoking's 'internalities'. Regulation 25(4):52–57

Hayek F (1945) The use of knowledge in society. Am Econ Rev 35(4):519–530

Hayek F (1952) The sensory order: an inquiry into the foundations of theoretical psychology. University of Chicago Press, Chicago

Herrmann-Pillath C (1994) Evolutionary rationality, "homo economicus," and the foundations of social order. J Soc Evol Syst 17(1):41–69

Iyengar SS, Lepper MR (2000) When choice is demotivating: can one desire too much of a good thing? J Personality Soc Psychol 79:995–1006

Kirman AP (1983) Mistaken beliefs and resultant equilibria. In: Frydman R, Phelps E (eds) Individual forecasting and collective outcomes. Cambridge University Press, Cambridge

Kirman AP (1993) Ants, rationality and recruitment. Q J Econ 108:137–156

Kopcke RW, Sneddon Little J, Tootell GMB (2004) How humans behave: Implications for economics and economic policy. New England Econ Rev 2004 (Final Issue):3–35

Kysar DA, Ayton P, Frank RH, Frey BS, Gigerenzer G, Glimcher PW, Korobkin R, Langevoort DC, Magen S (2006) Are heuristics a problem or a solution? In: Gigerenzer G, Engel C (eds) Heuristics and the Law. MIT Press, Cambridge (in press)

Lesourne J (1992) The economics of order and disorder. Clarendon Press, Oxford

McGeorge J, Aitken CK (1997) Effects of cannabis decriminalization in the Australian capital territory on university students' patterns of use. J Drug Issues 27(4):785–793

Ng Y-K (1999) Utility, informed preference, or happiness? Soc Choice Welfare 16(2):197–216

Ng Y-K (2003) From preference to happiness: towards a more complete welfare economics. Soc Choice Welfare 20(2):307–350

O'Donoghue T, Rabin M (2003) Studying optimal paternalism, illustrated by a model of sin taxes. Am Econ Rev 93(2):186–191

Österberg E (1990) Would a more liberal control policy increase alcohol consumption? Contemporary Drug Problems 17:545–573

Popham RE, Schmidt W, De Lint J (1972) The effects of legal restraint on drinking. In: Kissin B, Begleiter H (eds) The biology of alcoholism vol 4. Plenum, New York

Raiffa H (1982) The art and science of negotiation. Harvard University Press, Cambridge

Rossi P (2004) The risk of paternalism. Ama-Gi 6(1):10–13

Schneider CE (1998) The practice of autonomy: patients, doctors, and medical decisions. Oxford University Press, Cambridge

Schwartz B (2004) The paradox of choice: Why more is less. Harper Collins

Sheshinski E (2002) Bounded rationality and socially optimal limits on choice in a self-selection model. Working Paper, Hebrew University of Jerusalem

Sheshinski E (2003) Optimal policy to influence individual choice probabilities. Working Paper, Hebrew University of Jerusalem

Simon HA (1978) Rationality as process and as product of thought. Am Econ Rev 68(2):1–16

Simon HA (1982) Models of bounded rationality. MIT Press, Cambridge

Slovic P (2000) The Perception of risk. Earthscan, London

Smith VL (2003) Constructivist and ecological rationality in economics. Am Econ Rev 93(3):465–508

Suber P (1999) Paternalism. In: Gray CB (ed) Philosophy of law: an encyclopedia. Garland, Newyork

Sunstein CR (1997) Behavioral analysis of law. Univ Chicago Law Rev 64:1175–1195

Sunstein CR, Schkade D, Kahneman D (2000) Do people want optimal deterrence? J Legal Stud 29:237–249

Sunstein CR, Thaler R (2003) Libertarian paternalism is not an oxymoron. Univ Chicago Law Rev 70:1159–1202

Surowiecki J (2004) The wisdom of crowds. Doubleday, New York

Thaler RH, Benartzi S (2004) Save more tomorrow: using behavioral economics to increase employee saving. J Polit Econ 112:164–187

Thaler RH, Sunstein CR (2003) Libertarian paternalism. Am Econ Rev 93:175–179

Tisdell C (1996) Bounded rationality and economic evolution: a contribution to decision making, economics and management. Elgar, Northampton

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185:1124–1131

VanDeVeer D (1986) Paternalistic intervention: the moral bounds of benevolence. Princeton University Press, Princeton

Vriend NJ (1995) Self-organization of markets: an example of a computational approach. Comput Econ 8(3):205–231

Weber M (1958) The protestant ethic and the spirit of capitalism. Scribners, New York