Journal of Economic Literature 2014, 52(4), 1075–1118
http://dx.doi.org/10.1257/jel.52.4.1075

Behavioral Contract Theory[†]

Botond Kőszegi[*]

This review provides a critical survey of psychology-and-economics ("behavioral-economics") research in contract theory. First, I introduce the theories of individual decision making most frequently used in behavioral contract theory, and formally illustrate some of their implications in contracting settings. Second, I provide a more comprehensive (but informal) survey of the psychology-and-economics work on classical contract-theoretic topics: moral hazard, screening, mechanism design, and incomplete contracts. I also summarize research on a new topic spawned by psychology and economics, exploitative contracting, that studies contracts designed primarily to take advantage of agent mistakes. (*JEL* A12, D03, D82, D86)

1. *Introduction*

Psychology and economics—also known as behavioral economics—is a mindset for doing economics that espouses the importance of thinking about the psychological accuracy of models. After a history of identifying deviations from classical approaches, modeling these deviations formally, and empirically establishing their importance in economic decisions, the field is in the process of full integration into economic analysis: researchers are using the new psychologically based models to study how individual behavior plays out in organizations and markets, what the welfare consequences are, and how policy should respond to market outcomes—questions in which economists have always been interested.

This review summarizes and organizes one important part of the above development, the rapidly growing literature on behavioral contract theory. In a well-known textbook less than a decade old, Bolton and Dewatripont (2005, p. 10) wrote that "even though there is by now a large literature exploring a wide range of alternative models of individual choice . . . there have been relatively few explorations of the implications for optimal contracting." By the time of Bolton and Dewatripont's book, contract theory had evolved to a curious point: while researchers had explored optimal contracting in many complex economic environments, they largely followed a simplistic approach for

[*]Central European University. I am grateful to Paul Heidhues and Klaus Schmidt for insightful discussions and detailed comments on drafts of this review. I am also hugely indebted to Matthew Rabin for countless conversations over the years that helped clarify my thinking on many of the issues covered in this paper. I thank Mark Armstrong, Petra Bischofberger, Andrea Canidio, Takeshi Murooka, Yuval Salant, Patrick Schmitz, Ferdinand von Siemens, Rani Spiegler, and four anonymous referees for comments, and the European Research Council for support under Starting Grant #313341. Kinga Marczell provided excellent research assistance, including drafting the first version of the auction-theory parts of section 5.

[†]Go to http://dx.doi.org/10.1257/jel.52.4.1075 to visit the article page and view author disclosure statement(s).

modeling agent behavior. In a moral-hazard context, for instance, an agent has an outside option, trades off the monetary benefits and the costs of effort, and optimally chooses whether to participate and what to do.

In contrast, recent research has fruitfully incorporated behavioral-economics ideas into most of the classical contract-theoretic topics, including moral hazard, screening, auction theory, and incomplete contracts. In addition, the recognition that agents often fail to act in their best interests—and that principals might know this better than agents do—has inspired a novel literature on *exploitative contracts*, contracts whose exclusive or primary aim is to take advantage of agent mistakes. For each of these topics, I attempt to organize the main insights researchers have uncovered, as well as point out some of the many additional questions raised by existing work.

Like in any review, it is necessary to draw some, partly arbitrary, lines regarding which papers are included. I chose papers that are "behavioral" in the sense that a formal, relatively general model with psychological foundations or compelling psychological interpretation is involved, and for which some evidence supports the specific decision-making process. This excludes, for instance, the very interesting literature on bounded rationality predicated on optimization with cognitive, effort, or information-acquisition costs (e.g., Tirole 2009 and Bolton and Faure-Grimaud 2010), which does not yet seem to be based solidly on psychology interpretation and evidence.[1] Further, I focus on papers that address design questions related to contracts, mechanisms, or other extended interactions. This excludes purely game-theoretic issues, such as the winner's curse and overbidding in auctions (e.g., Eyster and Rabin 2005), that are studied in behavioral game theory; and pure pricing issues, such as oligopolistic pricing with consumer loss aversion (e.g., Heidhues and Kőszegi 2008) that are studied in behavioral industrial organization.[2]

I begin in section 2 by presenting the theories of individual decision making in psychology and economics that are most used in behavioral contract theory, and by working out some of their implications in simple settings. These theories are (i) loss aversion—whereby an individual evaluates economic outcomes relative to reference points, and weights losses more heavily than similar-sized gains; (ii) present bias—whereby an individual weights an earlier relative to a delayed outcome more heavily when the earlier outcome is in the present; (iii) inequity aversion—whereby people in a social situation dislike advantageous and especially disadvantageous inequality in material outcomes; and (iv) overconfidence—whereby a person displays unrealistically positive beliefs regarding her ability or prospects. In almost all applications, researchers assume that the agent (she) behaves according to one psychologically based model, while the principal (he) is fully rational and has a classical goal (usually profit maximization).

The main sections of the paper cover work from psychology and economics on the main contract-theoretic topics. Deviating from reviews by Rabin (1998) and DellaVigna (2009), I organize the discussion primarily

[1] Many papers in this literature, including those cited above, uncover mechanisms that intuitions suggest would hold and continue to be important in more psychologically based models. In other cases, however, the main results are based on essentially neoclassical optimization principles with novel types of presumed costs. Garicano and Prat (2013) review some of this work. How the insights in this literature connect to those based more closely on psychology evidence is a promising topic for future research.

[2] See Spiegler (2011) for an excellent textbook coverage of many industrial-organization models. A substantial amount of research is either at the intersection of behavioral industrial organization and behavioral contract theory, or is not clearly categorizable, and is hence covered in both this survey and Spiegler's textbook; see especially chapters 2–5 and 11 of the book.

not around behavioral-economics phenomena—whether the underlying model of decision making involves loss aversion, hyperbolic discounting, etc.—but rather according to contract-theoretic principles—whether the interaction qualifies as moral hazard, screening, etc. I do so for two main reasons. First, this organization reflects the fact that the literature has been addressing some of the classical contract-theoretic topics that interest other economists. Second, for some topics, the precise psychological source of the phenomena in question is not important (although for other questions it is).

In section 3, I discuss work on moral hazard, where the principal's main goal in contracting is to provide incentives for the agent to exert costly effort. Much of this literature focuses on sources of nonfinancial motivation that affect the agent's willingness to work hard. As the most important example, researchers recognize—following original discussions by Akerlof (1982) and Akerlof and Yellen (1988, 1990)—that organizations are often social in nature, so that agents' attitudes toward others have an important impact on behavior. Researchers study the nature of these and other nonpecuniary motivations, as well as their interaction with financial incentives.

In section 4, I summarize existing research on asymmetric information and screening, where the principal is attempting to interact with an agent whose private information he does not know. Psychology and economics introduces a number of important novel themes, including a new reason for screening contracts: present-biased agents prefer to write contracts to constrain their own future behavior, while knowing that information pertaining to optimal behavior will arrive in the future. In addition, the psychological phenomena emphasized in behavioral economics also lead to the natural screening issue of how to deal with agents who exhibit these phenomena to different extents.

Section 5 turns to a less extensive literature in behavioral contract theory, mechanism design—where a principal is interacting with multiple agents whose private information he does not know—and its particular application, auction theory. Many of the papers in this area study the implications of reference-dependent utility and loss aversion for auctions. The main results can be categorized around two forces. First, because loss-averse bidders dislike risk in how much they will pay, an auction designer has an incentive to choose auctions that insure them from this risk. Second, a designer may be able to manipulate bidders' reference points to his advantage.

Section 6 considers findings in exploitative contracting, the study of contract designs whose central consideration is exploiting an agent's mistakes. In most of these models, the agent is a consumer who either does not fully understand all features (e.g., all prices and fees) of a product or mispredicts her own behavior with respect to the product, and the principal is a profit-maximizing firm aware of the consumer's tendencies. This literature has reinvigorated the long-recognized but understudied topic of contracting with noncommon priors, but because of its conceptualization of noncommon priors as resulting from consumer mistakes, its focus on specific forms of mistakes, and its resulting ability to make welfare statements, it is usefully categorized as a new topic in contract theory.

In section 7, I discuss two nascent areas of research: incomplete contracts and environment design. In the existing research on incomplete contracts, authors assume that contracts change preferences by affecting feelings of entitlement or the reference point to which parties later compare outcomes, and study the implications for optimal contracting and the optimal allocation of property rights. Finally, section 8 provides a few general thoughts on the state of the literature and the challenges ahead for it.

2. *Key Theories in Psychology and Economics*

This section summarizes the main theories of individual decision making in psychology and economics that are used in the applications below. Rabin (1998) and DellaVigna (2009) provide excellent reviews of the evidence for these models.[3] To simultaneously introduce the model of human behavior, give a sense of how to work with it, as well as to illustrate some of its key economic implications, I present each theory together with an application covered in this review. For brevity, I do not carefully introduce all necessary assumptions in each example, nor present formal proofs.

2.1 *Loss Aversion*

Model. As has been demonstrated in a powerful body of work starting with Kahneman and Tversky (1979) and Tversky and Kahneman (1991), individuals evaluate economic outcomes not just according to an absolute valuation attached to the outcomes in question, but also relative to subjective reference points. The most important property of such reference-dependent preferences is loss aversion: losses relative to a reference point are more painful than equal-sized gains are pleasant.

I present a formal model based on Kőszegi and Rabin (2006, 2007), assuming here that the decision concerns only monetary outcomes such as a wage. If the agent's wage is w and her reference point is r, then her reference-dependent utility is

$$(1)\quad u(w\,|\,r) = \begin{cases} w + \eta(w - r) & \text{if } w \geq r, \text{ and} \\ w + \eta\lambda(w - r) & \text{if } w < r, \end{cases}$$

where $\eta > 0, \lambda > 1$. The first term in the utility function is consumption utility (linear in this example), which corresponds to the conventional notion of outcome-based utility. The second term is gain–loss utility, which captures the agent's sensations of gain or loss relative to the reference point. The parameter λ is the degree of the agent's loss aversion—with $\lambda > 1$ capturing that losses are more painful than gains are pleasant—and η is the weight on gain–loss utility. If there are other outcomes for which the agent evaluates gains and losses separately—for instance, the object in an auction—a similar additively separable utility function can be used to capture reference-dependent utility in those dimensions.

A crucial issue in using loss-averse preferences is the determination of the reference point r. The most common approach in the applications below is to set the reference point equal to the agent's rational expectations as defined by her full probabilistic beliefs (Kőszegi and Rabin 2006, 2007). It is then necessary to extend the above utility function to allow for the reference point to be a distribution $F(r)$:

$$(2)\qquad U(w\,|\,F) = \int u(c\,|\,r)\,dF(r).$$

This formulation captures the notion that the sense of gain or loss from a given consumption outcome derives from comparing it to all outcomes possible under the reference lottery. For example, if the reference lottery is a gamble between \$0 and \$100, an outcome of \$50 feels like a gain relative to \$0 and like a loss relative to \$100, and the overall sensation is a mixture of these two feelings.

Application. I illustrate some implications of the model in the context of wage setting, beginning with a property of risk attitudes that reappears in multiple applications. Suppose $w_L \leq w_M \leq w_H$, and let F denote the lottery that pays w_L with probability

[3] Note, however, that the ability of these models to explain important economic phenomena as explained in this review can—to the extent that other models cannot as convincingly account for the same phenomena—be thought of as additional evidence for the models.

p_L, w_M with probability p_M, and w_H with probability p_H. What is the agent's expected utility from F? An agent with rational expectations correctly anticipates the distribution of outcomes, so her reference point will be F. Equations (1) and (2) then imply:

$$\begin{aligned}
U(w_L|F) &= w_L - p_M\eta\lambda(w_M - w_L) \\
&\quad - p_H\eta\lambda(w_H - w_L) \\
U(w_M|F) &= w_M + p_L\eta(w_M - w_L) \\
&\quad - p_H\eta\lambda(w_H - w_M) \\
U(w_H|F) &= w_H + p_L\eta(w_H - w_L) \\
&\quad + p_M\eta(w_H - w_M)
\end{aligned}$$

Hence, expected utility is

$$\begin{aligned}
&(3)\quad U(F|F) \\
&= p_L U(w_L|F) + p_M U(w_M|F) + p_H U(w_H|F) \\
&= p_L w_L + p_M w_M + p_H w_H \\
&\quad - \eta(\lambda - 1)[p_L p_M(w_M - w_L) \\
&\qquad + p_L p_H(w_H - w_L) \\
&\qquad + p_M p_H(w_H - w_M)].
\end{aligned}$$

<0 unless $w_L = w_M = w_H$

The agent's expected gain–loss utility is always negative for risky outcomes, making her averse to risk. Intuitively, for instance, when $w_H > w_M$, the agent experiences w_M as a loss relative to w_H, and w_H as a gain relative to w_M; but since the sensation of loss outweighs the sensation of gain, her average gain–loss sensation is negative. Furthermore, the agent is first-order averse to risk: her utility decreases linearly with the dispersion in outcomes, for instance linearly in $w_H - w_M$. In particular, the agent greatly values contracts that insure her from risk on any part of the distribution of outcomes. These properties of the agent's aversion to risk are distinct from the properties generated by classical expected utility over wealth, leading to many qualitatively different predictions in economic environments.

As an example of the implications for contracting, I illustrate a loss-aversion-based explanation for "bonus" contracts—binary payment schemes in which the employee has a base salary, and, depending on whether her performance exceeds a threshold, may in addition receive a bonus—by Herweg, Müller, and Weinschenk (2010). Suppose an agent chooses between low effort (e_L) and high effort (e_H), where the cost of effort e_i, $i \in \{L, H\}$, is c_i. The agent's output takes one of three values; low, medium, or high. The following table gives the probabilities for the three levels of output occurring as a function of effort:

	low output	medium output	high output
e_L	2/3	1/3	0
e_H	1/3	1/3	1/3

I assume that parameters are such that the principal would like the agent to exert the high level of effort. Then, if the agent is averse to risk, the informativeness principle in the classical model of moral hazard predicts that the principal pays different wages for different levels of output (Holmstrom 1979). Intuitively, the three levels of output are differently informative regarding whether the agent exerted high effort, and hence should be rewarded differently.

In contrast, the same is not optimal for a loss-averse agent (even though she is averse to risk). Let the wage paid for low, medium, and high output be w_L, w_M, and

w_H, respectively. Clearly, the principal wants $w_H \geq w_M \geq w_L$. Using equation (3), the principal's problem is[4]

$$\begin{aligned}
&\min \tfrac{1}{3}(w_L + w_M + w_H)\\
\text{(PC)}\quad &\text{s.t. } \tfrac{1}{3}(w_L + w_M + w_H)\\
&\quad - \tfrac{1}{9}\eta(\lambda - 1)(w_H - w_M + w_M\\
&\qquad\qquad - w_L + w_H - w_L)\\
&\quad - c_H \geq \underline{u}\\
\text{(IC)}\quad &\tfrac{1}{3}(w_L + w_M + w_H)\\
&\quad - \tfrac{1}{9}\eta(\lambda - 1)(w_H - w_M + w_M\\
&\qquad\qquad - w_L + w_H - w_L) - c_H\\
&\quad \geq \tfrac{2}{3}w_L + \tfrac{1}{3}w_M\\
&\qquad - \tfrac{2}{9}\eta(\lambda - 1)(w_M - w_L) - c_L
\end{aligned}$$

The principal minimizes the expected wage paid to the agent when she exerts high effort, subject to two constraints. The participation constraint (PC) captures that the agent must be willing to take the contract over her best alternative, which has utility $\underline{u}$. The incentive-compatibility constraint (IC) captures that the agent must be willing to exert high effort over low effort. Rewriting the constraints:

$$\begin{aligned}
\text{(PC)}\quad &\tfrac{1}{3}(w_L + w_M + w_H)\\
&\quad - \tfrac{2}{9}\eta(\lambda - 1)(w_H - w_L) \geq c_H + \underline{u}\\
\text{(IC)}\quad &\tfrac{1}{3}(w_H - w_L)\\
&\quad - \tfrac{2}{9}\eta(\lambda - 1)(w_H - w_M) \geq c_H - c_L
\end{aligned}$$

Now I show that $w_H > w_M > w_L$ is not optimal by identifying a better contract if this is the case. First, increase w_M by ϵ and decrease w_H and w_L by $\epsilon/2$. As a result of this, profits are unchanged, PC is still satisfied, and IC becomes slack. Then, decrease w_H by ϵ' and increase w_L by ϵ'. This again leaves profits unchanged, and for a sufficiently small ϵ', IC remains slack, while now PC becomes slack as well. With both constraints slack, the principal can decrease all three wage levels by a sufficiently small ϵ'', increasing profits and still satisfying both constraints.

The above argument implies that the principal chooses at most two different wage levels. Clearly, one wage level (i.e., a constant wage) is not incentive compatible, so in the optimal contract the principal uses *exactly* two wage levels. Hence, either $w_L = w_M < w_H$ or $w_L < w_M = w_H$. Notice that the above improvement applies when $w_L = w_M < w_H$, so this is not optimal either. Hence, in the optimal contract $w_L < w_M = w_H$.

Intuitively, because a loss-averse agent strongly dislikes random variation in the wage due to uncertainty in the environment, the principal has an incentive not to vary wages too finely with output. At the same time, he also needs to provide incentives, so the agent's wage cannot be completely flat. As a result, the principal chooses the minimal amount of wage variation that can still provide incentives: two wage levels, that is, a bonus contract.

2.2 *Present Bias and Time Inconsistency*

Model. In intertemporal decisions, many individuals value the future significantly less than the present, yet place similar weights on different points in time in the future. This

[4] Technically, in this example there are two dimensions of utility, money and effort, so one must also account for gain–loss utility in effort. In this case, however, gain–loss utility in effort is zero, so that total utility in effort is equal to consumption utility: when the agent chooses effort i, with rational expectations both her reference point and her outcome in the effort–cost dimension are c_i.

preference implies that a person would not like to make investments into the future today, but would like to make similar investments in the future. Time inconsistency—a conflict between the person's preferences at different points in time—arises because once the future comes, she would again prefer not to make immediate investments.

Laibson (1997) and O'Donoghue and Rabin (1999a) formalize the above "present bias" using the $\beta - \delta$ model. They assume that self t (the decisionmaker's period-t incarnation) evaluates the stream $u_t, u_{t+1}, u_{t+2}, u_{t+3}, \ldots$ of instantaneous utilities at times $t, t+1, t+2, t+3, \ldots$ as

$$u_t + \beta\delta u_{t+1} + \beta\delta^2 u_{t+2} + \beta\delta^3 u_{t+3} + \cdots,$$

where $\beta, \delta \leq 1$ are the discount factors that parameterize present bias. In this formulation, the discount factor between periods t and $t+1$ is $\beta\delta$, whereas the discount factor between any two consecutive future periods is δ. For $\beta < 1$, therefore, self t discounts between the present and the future more heavily than between different periods in the future, capturing the above pattern of intertemporal preferences. For $\beta = 1$, the model reduces to the exponential discounting model commonly used in economics.

A key consideration in applications of the $\beta - \delta$ model is whether the agent correctly anticipates the present bias that she will (but would prefer not to) have in the future. Following O'Donoghue and Rabin (2001), I suppose that the agent believes that her future β will be $\hat{\beta}$. The assumption $\beta = \hat{\beta}$ corresponds to full sophistication regarding time inconsistency—whereby the agent perfectly understands that she will behave differently than she would now like—whereas $\hat{\beta} = 1$ corresponds to full naïveté—whereby the agent completely ignores her time inconsistency in preferences. Intermediate values of $\hat{\beta}$ capture intermediate values of naïvete.

Application. I present a model of credit contracts based on Heidhues and Kőszegi (2010) and Eliaz and Spiegler (2006). Suppose a consumer with $\beta = 1/2$ and $\delta = 1$ is contracting in period 0 with competitive suppliers of credit, who have funds available to them at an interest rate of zero. The consumer borrows for future consumption and repays her loan in periods 1 and 2. A credit contract consists of a borrowed amount c and a menu of installment plans $r_1 \geq 0$, $r_2 \geq 0$ according to which the consumer can repay her loan in periods 1 and 2. Once the consumer signs a contract with a firm, she cannot borrow from other firms, but she can decide in period 1 which of the installment plans designated in her contract to follow. The consumer's instantaneous utility from consumption is $u(c)$, whereas her instantaneous disutility from making a payment is r. Hence, since borrowing is for future consumption, the consumer's utility in period 0 is $u(c) - r_1 - r_2$; but in period 1, she chooses the installment plan to follow minimizing $r_1 + r_2/2$.

To find the optimal contract, consider first fully sophisticated consumers, for whom $\hat{\beta} = \beta$. Since a sophisticated consumer knows how she will behave, without loss of generality we can suppose that the menu of installment plans specifies only the one plan r_1, r_2 she will choose. Hence, a firm's problem is

$$\max r_1 + r_2 - c$$

$$\text{(PC)} \qquad \text{s.t. } u(c) - r_1 - r_2 \geq \underline{u},$$

where $\underline{u}$ is the consumer's utility from her outside option, and the constraint (the "participation constraint" or PC) captures that the consumer must be willing to accept the principal's contract over her best alternative. Substituting the binding PC into the maximand produces the first-order condition $u'(c) = 1$, which means that the consumer's

consumption is efficient. This simple analysis leads to an important point made by DellaVigna and Malmendier (2004). Even though the consumer suffers from a time-inconsistency problem, the competitive-equilibrium contract maximizes her utility, as it leads her to borrow the optimal amount. Hence, with sophisticated consumers, markets can solve time-inconsistency problems.

The situation is entirely different for a nonsophisticated borrower, for whom $\hat{\beta} > \beta$. To see this, suppose first that c is fixed, so that the contract concerns only the repayment terms. For any contract the consumer is offered, there is a "baseline" installment plan $(\hat{r}_1, \hat{r}_2)$ she believes in period 0 she will follow, and a possibly different, "alternative" installment plan (r_1, r_2) she will actually follow. Without loss of generality, we can assume that the contract contains no other repayment options. The installment plans (r_1, r_2) and $(\hat{r}_1, \hat{r}_2)$ must solve

$$\max r_1 + r_2 - c$$

$$\text{(PC)} \qquad \text{s.t. } u(c) - \hat{r}_1 - \hat{r}_2 \geq \underline{u}$$

$$\text{(PCC)} \qquad \hat{r}_1 + \hat{\beta}\hat{r}_2 \leq r_1 + \hat{\beta} r_2$$

$$\text{(IC)} \qquad r_1 + r_2/2 \leq \hat{r}_1 + \hat{r}_2/2$$

As explained by Heidhues and Kőszegi (2010), the firm faces the following constraints in designing its contract. First, for the borrower to be willing to accept the firm's offer, self 0's utility given how she believes she will behave must be at least $\underline{u}$ (PC). Second, if self 0 is to think that she will choose the baseline option, then given her beliefs $\hat{\beta}$ she must think she will prefer it to the alternative option (perceived-choice constraint or PCC). Third, if self 1 is to actually choose the alternative repayment schedule, she has to prefer it to the baseline repayment schedule (IC).

The solution to this problem is $(\hat{r}_1, \hat{r}_2) = (\underline{u} - u(c), 0)$ and $(r_1, r_2) = (0, 2(\underline{u} - u(c)))$.[5] Given that in a competitive equilibrium firms make zero profits, the equilibrium payments are $(\hat{r}_1, \hat{r}_2) = (c/2, 0)$ and $(r_1, r_2) = (0, c)$. In practice, this type of contract would correspond to a credit arrangement in which, as for instance with many credit cards and nontraditional mortgages, the consumer is asked to repay her loan fast, and delaying repayment carries large penalties.

The optimal contract, therefore, induces false beliefs in the consumer regarding how she will repay, making the contract look more attractive than it really is. In fact, the contract maximizes the consumer's mistake: by asking her to repay only in period 1, the lender maximizes the effect of the consumer's mistaken beliefs regarding her future willingness to pay to delay repayment. This first implication of the example, whereby consumers are led to buy seemingly cheap products, is a general implication of exploitative contracting that appears in many papers discussed in section 6.

The fact that consumers are sold seemingly cheap products can have many types of adverse welfare consequences. In this particular case, a welfare cost arises if the borrowed amount c is endogenous. Since the consumer believes that borrowing c costs only $c/2$, the competitive equilibrium with endogenous c has $u'(c) = 1/2$—the consumer overborrows relative to the optimum. Note that since the borrowing is for future consumption, the overborrowing is not due to the borrower's time inconsistency per se, but due to her false beliefs about her future repayment behavior.

[5] For any $(\hat{r}_1, \hat{r}_2)$, $(r_1, r_2) = (0, 2\hat{r}_1, \hat{r}_2)$ satisfies both IC and PCC (the latter constraint is slack), and no other (r_1, r_2) that satisfies IC achieves higher profits, so it is the optimal (r_1, r_2). Then, among $(\hat{r}_1, \hat{r}_2)$ that satisfy PC, the option that maximizes the principal's resulting profits, $2\hat{r}_1, \hat{r}_2$, has $\hat{r}_2 = 0$.

All of the above holds for *any* $\hat{\beta} > \beta$—that is, for even arbitrarily small degrees of naïveté regarding time inconsistency. This leads to the second important implication of the example: with time inconsistency, even small amounts of naïveté often have large welfare implications. Heidhues and Kőszegi (2010) show that simple restrictions on the contract form can drastically increase welfare. As an illustration, suppose that we impose a sufficiently low constraint p on how much the consumer can be penalized for failing to choose the lowest-cost installment plan in her contract. Then, in one competitive-equilibrium contract for a fixed consumption level c, the consumer chooses between installment plans $(c - p, 0)$ and $(0, c)$.[6] Since the consumer thinks the cost of borrowing c is $c - p$, the competitive-equilibrium level of borrowing becomes $u'(c) = 1$. Hence, with this regulation the consumer borrows the optimal amount, and—although she falsely believes she will get a better deal—ultimately repays exactly that amount.

2.3 *Inequity Aversion*

Model. To explain a number of disparate facts regarding prosocial behavior, Fehr and Schmidt (1999) present a model in which individuals dislike both advantageous and disadvantageous inequality in material outcomes, but dislike being behind more than they dislike being ahead. Suppose that there are N individuals in agent i's "reference group"—the group of individuals (including herself) whose material outcomes she cares about. Then, her utility from material payoffs $x_1, \ldots, x_N$ is

$$x_i - \frac{1}{N-1}\left(\alpha_i \sum_{j:x_j \geq x_i} (x_j - x_i) + \beta_i \sum_{j:x_j < x_i} (x_i - x_j)\right), \tag{4}$$

where β_i satisfying $0 \leq \beta_i < 1$ parameterizes the agent's aversion to being ahead and $\alpha_i \geq 0$ parameterizes her aversion to being behind.

Application. I illustrate the implications of inequity aversion for incentive contracts. First, consider a "gift-exchange" situation, in which the principal pays the agent a fixed output-independent wage $w \geq 0$, and then the agent chooses effort level $e \geq 0$. The agent's material cost of effort is $c(e)$, and this produces output e for the principal. Hence, the agent's material payoff is $x_2 = w - c(e)$, whereas the principal's is $x_1 = e - w$, and the effort level that maximizes social welfare satisfies $c'(e) = 1$. Note that if the principal and the agent were selfish, the agent would exert zero effort for any wage, and hence the principal would pay a wage of zero.

As has long been recognized in the literature, however, inequity aversion can lead to a higher wage and a higher level of effort, potentially increasing material welfare for both the principal and the agent. Suppose that the principal and the agent are inequity averse as defined above, $N = 2$, the principal is player 1 and the agent is player 2, and $\alpha_i \geq \beta_i > 1/2$. We first solve for how the agent reacts to a wage $w > 0$, assuming that the optimal e has $c'(e) \leq 1$ (the analysis below makes clear that this is the case for the principal's optimal contract). Clearly, the agent never chooses an effort level that leads to $x_1 > x_2$; relative to $x_1 = x_2$, this would lead to a lower material payoff for her and generate further disutility from disadvantageous inequality. Hence, the agent maximizes $x_2 - \beta_2(x_2 - x_1) = w(1 - 2\beta_2) + \beta_2 e - (1 - \beta_2)c(e)$ subject to the constraint $x_2 \geq x_1$, or $w - c(e) \geq e - w$. Now it is easy to see that $w - c(e) > e - w$ cannot be optimal: if this were the case, then (using that $\beta_2 > 1/2$ and therefore $\beta_2 > 1 - \beta_2$, and $c'(e) \leq 1$) the agent would want to increase e. As a result, the agent's inequity aversion leads

[6] In this case, there are other competitive-equilibrium contracts: for instance, if $\hat{r}_1 \geq p$ and $\hat{r}_1 + \hat{r}_2 = c - p$, then $(\hat{r}_1, \hat{r}_2)$ and $(r_1, r_2) = (\hat{r}_1 - p, \hat{r}_2 + 2p)$ is a competitive equilibrium.

her to choose her level of effort to equalize material payoffs, setting $w - c(e) = e - w$, or $w = (e + c(e))/2$.

Given that the agent's behavior eliminates inequality, the principal maximizes material payoffs, and her problem can therefore be written as

$$\max \quad e - w \qquad (5)$$
$$\text{s.t.} \quad w = \frac{e + c(e)}{2}.$$

Substituting the constraint into the maximand yields that the principal is maximizing $(e - c(e))/2$, leading to $c'(e) = 1$. This means that inequity aversion generates an efficient outcome: it induces an effort level that maximizes material efficiency, and since this goes along with the lack of inequality, it is also efficient in terms of inequity-averse preferences. To understand the intuition, note that the agent's inequity aversion acts as an incentive device: since a higher wage means that a higher effort is necessary to split the surplus equally, the agent responds to a higher wage by increasing effort. Knowing this and knowing that the agent will give him exactly half of the social surplus, the principal has an incentive to induce the effort level that maximizes social surplus.

An alternative to gift exchange is the "voluntary-bonus" system, in which the agent first chooses her level of effort, and then the principal chooses a bonus $b \geq 0$ to reward her. Similarly to how the agent chooses her level of effort in the case of gift exchange above, the principal chooses the bonus to equalize material payoffs, setting $b - c(e) = e - b$, or $b = (e + c(e))/1$. Again, therefore, inequity aversion creates an incentive device, but it does so in a different way: while the gift-exchange arrangement relies on the inequity aversion of the agent to reward the principal's investment of a high wage, the voluntary-bonus system relies on the principal's inequity aversion to reward the agent's investment of a high effort.

As the principal's bonus-setting behavior results in no inequality, the agent maximizes material payoffs $b - c(e) = (e + c(e))/2 - c(e)$, which again yields $c'(e) = 1$. Hence, the voluntary-bonus arrangement also maximizes social surplus. The intuition is simple: knowing that the principal will split the social surplus $e - c(e)$ equally, the agent has an incentive to maximize social surplus.

The above analysis, however, provides a simplistic view of the power of inequity aversion to generate efficient outcomes. Motivated by Fehr et al. (2007), suppose that a proportion q of individuals in society are inequity averse as above, but a proportion $1 - q$ are selfish. In that case, the gift-exchange and bonus arrangements are neither optimal nor equivalent. To illustrate, I suppose for simplicity that $\alpha_i = \infty$, so that inequity-averse individuals do not tolerate disadvantageous inequality.

Consider first a gift-exchange arrangement. An inequity-averse agent responds to a positive wage as above, but a selfish agent chooses zero effort. Since a positive wage followed by zero effort generates disadvantageous inequality for the principal, an inequity-averse principal—afraid to run into a selfish agent—never chooses a positive wage. Denoting an inequity-averse agent's effort as a function of the wage by $e(w)$, the selfish principal chooses w to maximize $qe(w) - w$, yielding the first-order condition $qe'(w) = 1$. Totally differentiating equation (5) with respect to w, solving for $e'(w)$, and plugging into the first-order condition gives

$$c'(e) = 2q - 1.$$

Hence, effort is lower than optimal for any $q < 1$.

Now I return to the voluntary-bonus arrangement. An inequity-averse agent

would find it intolerable to exert positive effort and then not be rewarded by a selfish principal, so she exerts zero effort. Recalling that an inequity-averse principal chooses $b = (e + c(e))/2$, a selfish agent maximizes $q(e + c(e))/2 - c(e)$, which gives the first-order condition

$$c'(e) = \frac{q}{2 - q}.$$

Again, effort is lower than optimal for any $q < 1$. The intuition for why effort is lower than optimal in the presence of selfish individuals under either arrangement is simple: since a party's attempt to increase the surplus might go unrewarded, she has less of an incentive to increase the social surplus.

But while both arrangements generate lower-than-optimal effort, they do not generate the same effort. Simple arithmetic shows that $q/(2 - q) > 2q - 1$, so effort is higher under the bonus arrangement than under gift exchange. Furthermore, while the bonus system generates effort whenever the agent is selfish, the gift-exchange system generates effort only if the principal is selfish *and* the agent is inequity-averse. Hence, under heterogeneity in inequity aversion, the bonus system is superior. This difference is due to a difference in how much the first mover has to invest in an attempt to increase the social surplus and be rewarded for it by an inequity-averse second mover. For the agent to increase the social surplus under the bonus system, she must pay her marginal cost of effort. But for the principal to increase the social surplus by increasing the wage in the gift-exchange system, he must compensate the agent for her marginal cost of effort *plus* half of the marginal increase in the social surplus (since when splitting the social surplus, the agent will keep half of the increase for herself). When this investment is certain to be rewarded—when there are no selfish individuals around—it does not matter which investment is more costly, as in either case the investing party gets half of the increase in the social surplus. When there is a chance that the investment will not be rewarded, however, investment is lower when its marginal cost is higher.

2.4 *Quasi-Bayesian Models*

Framework. The literature has documented a number of systematic mistakes individuals commit when thinking about statistical properties of random phenomena. A commonly used reduced-form way to capture such mistakes is the quasi-Bayesian approach: one posits that the agent updates her beliefs fundamentally in a Bayesian way, but commits a particular error that is inconsistent with rational inference. These models typically also assume that the agent does not learn about her error from her observations. The quasi-Bayesian approach is not a theory, but merely a framework for thinking about statistical mistakes. A fully-fledged theory specifies the particular mistake the agent is committing. In the models reviewed in this article, there are two general mistakes authors assume: systematically incorrect priors, and mistakes in updating beliefs based on information. The most common version of an incorrect prior is overconfidence: the agent has overly positive views about herself or some of her prospects.

Application. I consider the implications of overconfidence for contracting. First, suppose that an agent produces one of two levels of output, high or low. The true probability of high output is q, but the agent has an incorrect prior in that she believes the probability of high output is $\tilde{q} > q$. The principal offers an output-contingent wage, paying w_H for high output and w_L for low output. The principal is risk neutral, while the agent has mean-variance preferences of the form $\tilde{q}w_H + (1 - \tilde{q})w_L - \tilde{q}(1 - \tilde{q})(w_H - w_L)^2$. For now, suppose that there is no asymmetric information or moral hazard.

In this case, the principal's problem is

$$\text{(PC)}\ \min\ qw_H + (1-q)w_L$$
$$\text{s.t.}\ \tilde{q}w_H + (1-\tilde{q})w_L - \tilde{q}(1-\tilde{q})(w_H - w_L)^2 \geq \underline{u}.$$

Noting that $qw_H + (1-q)w_L = \tilde{q}w_H + (1-\tilde{q})w_L - (\tilde{q}-q)(w_H - w_L)$, we can plug PC into the maximand to get

$$\min \underline{u} + \tilde{q}(1-\tilde{q})(w_H - w_L)^2 - (\tilde{q}-q)(w_H - w_L).$$

This gives the optimal wage spread

$$w_H - w_L = \frac{\tilde{q}-q}{2\tilde{q}(1-\tilde{q})} \equiv \Delta^* \quad (6)$$

Given $w_H - w_L$, the principal sets the levels of the two wages so that PC binds.

Equation (6) identifies a basic exploitation motive with overconfident agents: if the agent is too optimistic that high output will occur, she overvalues the wage paid for high output, so the principal can decrease expected wages by paying her a little more for high output and much less for low output. This is a special case of the well-known observation that individuals with different priors would like to speculate against each other (Harrison and Kreps 1978, for instance). It is also an example of a situation where a type of contract that may appear to have a classical purpose (an output-contingent wage often serves to overcome moral hazard) instead has an exploitative purpose.

The exploitation motive, however, is limited for agents with only slightly overoptimistic beliefs, so that in more realistic settings these agents do not receive exploitative contracts. This finding contrasts with that for time-inconsistent agents above—where even small degrees of naïveté can have large welfare effects—suggesting that the consequences of false beliefs can be less serious under time consistency than under time inconsistency. As a first example of how agents with small overconfidence can fare well, I consider a version of de la Rosa's (2011) moral-hazard model. Suppose that the agent chooses between two levels of effort, high and low. Her output with high effort is as described above, but her output with low effort is low with probability 1. The cost of low effort is zero, and the cost of high effort is e. The agent's incentive-compatibility constraint is then

$$\tilde{q}w_H + (1-\tilde{q})w_L - \tilde{q}(1-\tilde{q})(w_H - w_L)^2 - e \geq w_L.$$

or

$$\tilde{q}(w_H - w_L) - \tilde{q}(1-\tilde{q})(w_H - w_L)^2 \geq e. \quad (7)$$

Let the lowest $w_H - w_L$ satisfying this inequality be Δ^{**}. The comparison of Δ^* and Δ^{**} highlights an important distinction regarding different kinds of contracts in this review. If $\Delta^* > \Delta^{**}$, then $w_H - w_L = \Delta^*$ is optimal, and the agent's IC constraint is not binding. In the sense that the principal's primary consideration is exploiting the agent's mistaken beliefs—and possible constraints derive from this attempt, while other constraints are either nonexistent or not binding—the contract is exploitative by the (somewhat informal) standards of Section 6. In contrast, if $\Delta^* < \Delta^{**}$, then $w_H - w_L = \Delta^{**}$ is optimal. In this case, therefore, the consideration determining $w_H - w_L$ remains to make sure that the agent's incentive-compatibility constraint is satisfied, not to take bets against her. In this sense, the contract is not primarily about exploitation, although the agent's expected wage is still affected by her overconfidence.

It is interesting to note that—as is easy to check using equation (7)—in this latter case the optimal wage spread is *decreasing* in overconfidence. Intuitively, an overconfident agent overestimates how easily she can produce high output, and hence a less output-sensitive incentive contract is sufficient to induce her to choose high effort. This insight has some interesting welfare implications. First, because a lower-powered incentive contract allows the agent to bear less risk, mild overconfidence increases efficiency. Second, when principals compete for the agent, the agent enjoys the full increase in social efficiency, so that having slightly wrong beliefs actually *benefits* her.[7]

As a second limitation on exploiting small degrees of overconfidence, consider screening agents with different degrees of optimism, based on Eliaz and Spiegler (2008). Suppose that there is no moral hazard, but in addition to the agent above, there is an agent with beliefs $\tilde{Q} > \tilde{q}$, and the principal would like to contract with both agents without directly observing their beliefs. Based on equation (6), the principal would like to offer a slightly exploitative contract to the less optimistic agent $\tilde{q}$ and a more exploitative contract to the more optimistic agent $\tilde{Q}$. Given the latter incentive, however, the principal might not exploit agent $\tilde{q}$. I demonstrate this heuristically by showing that the principal prefers to offer agent $\tilde{q}$ a contract with $w_H - w_L = 0$ rather than $w_H - w_L = \epsilon$, where $\epsilon > 0$ is small. With $w_H - w_L = \epsilon$, the less optimistic agent's binding PC becomes

$$w_L + \tilde{q}\epsilon - \tilde{q}(1 - \tilde{q})\epsilon^2 = \underline{u},$$

which for small ϵ implies $w_L \approx \underline{u} - \tilde{q}\epsilon$, yielding $qw_H + (1 - q)w_L = w_L + q\epsilon \approx \underline{u} - (\tilde{q} - q)\epsilon$. Hence, relative to a contract with $w_H - w_L = 0$, the principal makes a profit of approximately $(\tilde{q} - q)\epsilon$ on agent $\tilde{q}$. But how much does she lose on agent $\tilde{Q}$? To see this, note that agent $\tilde{Q}$'s perceived value from taking the above contract is

$$w_L + \tilde{Q}\epsilon - \tilde{Q}(1 - \tilde{Q})\epsilon^2$$

$$\approx w_L + \tilde{Q}\epsilon \approx \underline{u} + (\tilde{Q} - \tilde{q})\epsilon$$

This means that, when contracting with agent $\tilde{Q}$, the principal loses a profit of $(\tilde{Q} - \tilde{q})\epsilon$ because having the less speculative contract around improves agent $\tilde{Q}$'s perceived alternative option. If $\tilde{q} - q$ is sufficiently lower than $\tilde{Q} - q$, therefore, exploiting agent $\tilde{q}$ is not worth it for the principal. Intuitively, while the less optimistic type values a contract with $w_H > w_L$ and therefore allows the principal to decrease expected wages, the more optimistic type values the same contract much more, so that—to avoid having to pay a kind of information rent—the principal would rather not exploit the less optimistic type. This result implies that the principal either does not exploit the agent or exploits her substantially, generating the empirical implication that if we see exploitation, it should be large.

3. *Moral Hazard*

For the rest of this review, I turn to a more comprehensive, but informal, survey of efforts to incorporate psychology-and-economics ideas into contract theory. I organize the topics according to classical contract-theoretic principles, and begin with discussing moral hazard.

Moral hazard arises in contracting situations when a decision taken by the agent after the contract is signed affects the utility

[7] To see this, note that under competition, principals pay the same expected wage to any agent—the agent's expected output conditional on high effort—so a decrease in the wage spread increases the agent's utility.

of the principal, and the action cannot be directly contracted upon. In these settings, there are typically outcomes related to the agent's action that can be contracted upon (e.g., output), and the key issue is how the principal optimally designs a contract that induces the agent to take an action he prefers (captured in the incentive-compatibility constraint), and that the agent will accept (captured in the participation constraint). This section summarizes the existing behavioral-economics literature on moral hazard.[8]

3.1 *Intrinsic Motivation: Reliance on Voluntary Actions*

One important theme that has emerged from the literature is that optimal incentive schemes may rely crucially on "intrinsic motivation"—actions that do not carry a financial reward. This contrasts with classical models of moral hazard, where the only source of motivation is explicit or implicit financial incentives.

Social preferences as a source of intrinsic motivation. Akerlof (1982) was perhaps the first (within economics) to argue that firms might pay high wages to induce high effort as part of a gift exchange. Englmaier and Leider (2012) formalize this idea in a model in which the agent has inequity-averse preferences vis-à-vis the principal. As in previous models, the principal can induce high effort with performance-based pay. But he might also induce high effort with performance-independent pay. Specifically, if the principal pays a low (near-market-clearing) fixed wage, the agent does not work hard, lest the principal enjoy a much higher payoff; but if the principal pays a high fixed wage, the agent responds with high effort so that the principal receives a fair share of the pie. Hence, a high wage is often profit maximizing because it puts the agent ahead and thereby creates intrinsic motivation to help the principal. Englmaier and Leider (2012) also investigate conditions under which it is optimal to use inequity-aversion-based incentives versus performance-based pay, and show that inequity aversion is more likely to be used if output is a poor signal of effort or the agent is highly inequity averse or productive. This is one formalization of Akerlof's distinction between "primary" labor markets (governed by gift exchange and above-market-clearing wages) and "secondary" labor markets (governed by explicit incentives and market-clearing wages).

Voluntary bonuses. Fehr, Klein, and Schmidt (2007) establish that if heterogeneity in inequity aversion is present in the population, a contract based on voluntary bonuses—by which the principal can reward the agent for high effort ex post—often dominates the above high-wage contract based on gift exchange. A bonus contract engages inequity aversion much like a high-wage contract above, except with the order of moves reversed: much like an inequity-averse worker responds to a high wage (and only a high wage) with high effort, an inequity-averse principal is willing to pay the bonus in response to (and only in response to) high effort. Hence, if there are sufficiently many fair-minded principals, both fair-minded and selfish agents are willing to exert effort. The superiority of the bonus scheme arises from the presence of selfish individuals in the population: a party that first takes a costly action (exerting effort or paying a high wage, respectively) faces the risk that a selfish partner will not reward it, and this

[8] Kamenica (2012) reviews evidence of many systematic psychological forces on how individuals respond to incentives, and more specifically, Fehr, Goette, and Zehnder (2009) review evidence on how reference-dependent fairness concerns affect interactions in a labor-market setting.

risk is lower for the bonus contract because the action has lower cost in that case.[9]

3.2 *Reduced Wage Sensitivity*

Another finding in the literature is that—for multiple reasons—an agent's wage may be less responsive to her output than a classical moral-hazard model predicts.

Wage compression. Complementing the research on moral hazard with social preferences reviewed in the previous subsection, it seems natural to assume that employees compare themselves not just to their employer, but also to fellow employees. A very intuitive implication of inequity aversion is then that wages tend to be more compressed than what one would expect based on classical models. Because agents do not like the ex post inequality that can occur by chance as a result of high-powered (i.e., very performance-sensitive) individual incentives, to relax their participation constraint individual incentive pay is reduced. When agents are sufficiently averse to ex post inequality, the principal might institute team-based incentives not because he cannot observe individual performance or wants to encourage cooperation (as in many classical models), but to insure agents against painful inequality (Englmaier and Wambach 2010; Bartling 2011). Furthermore, Bartling and von Siemens (2010) argue that because comparisons are less pronounced across firms than within firms, consistent with the evidence inequity aversion predicts no wage compression across firms.

Despite the general tendency of inequity aversion to generate wage compression, Itoh (2004), Rey-Biel (2008), and Bartling and von Siemens (2010) identify some conditions under which increased pay inequality can result. Specifically, if limited liability prevents the principal from punishing the agent by reducing her pay, he can instead punish her by paying another agent more. In Rey-Biel's (2008) two-agent deterministic setting, for instance, an agent who works hard when the other agent shirks is more than compensated for her cost of effort. In other words, the principal may prefer tournament-type incentives even when classical reasons to prefer such an incentive structure—e.g., common shocks—are absent.

Simple schemes due to loss aversion. In the canonical model of moral hazard in a risky environment, increasing the power of incentives is beneficial to the principal because it aligns the agent's motives with the principal's goals, but it is also costly because it requires exposing the agent to more risk and hence paying her a higher risk premium. Because loss aversion affects a person's attitudes toward risk, it affects this fundamental tradeoff. A finding that reappears in multiple papers is that—because it generates first-order risk aversion—loss aversion leads to the unresponsiveness of transfers to outcomes over some regions of outcomes. As an extreme example in the context of contracting when the consumer is loss averse and her demand for the product is uncertain (e.g., mobile phone contracts), Herweg and Mierendorff (2013) show that—consistent with many real-world examples—the seller might prefer a flat-rate contract. The flat-rate contract leads the consumer to overuse the product, but it is still profit maximizing if this moral-hazard problem is sufficiently weak.

In many moral-hazard settings, especially employment contracts, however, incentives must also be provided, so completely flat

[9] Similarly, Fehr and Schmidt (2004) show that a voluntary-bonus scheme outperforms an explicit incentive contract in a multitasking environment in which performance on both tasks is observable to the principal, but performance on only one is contractible. As is well-known since Holmström and Milgrom (1991), providing incentives on the contractible task leads an agent to ignore the unincentivized task, leading to inefficiency. A scheme in which the principal can reward high effort on both tasks with a voluntary bonus increases efficiency.

contracts are not optimal. In such situations, wages must be responsive to performance somewhere, but nevertheless, loss aversion predicts simpler reward schemes than do typical classical settings. Herweg, Müller, and Weinschenk (2010) establish that if the agent's reference point is her rational expectations about the wage (as in Kőszegi and Rabin 2006, 2007), then the optimal contract often has a "bonus structure," with two possible wage levels. In a dynamic extension of the expectations-based model due to Kőszegi and Rabin (2009), Macera (2012) proves that under some circumstances the current wage is completely unresponsive to performance, with incentives provided by future opportunities.[10] And de Meza and Webb (2007) show that if the reference point is the median wage, then the agent's wage is unresponsive to performance up to the median performance level, but responsive to performance above that level.[11]

Wage stickiness. Eliaz and Spiegler (2014) develop a model of labor-market dynamics that combines the two major features of worker behavior studied separately by the papers above: (i) that workers can be induced to exert discretionary effort by paying them a fair wage; and (ii) that workers' notion of a fair wage is based on their lagged expectations of the wage. In each period, firms face a productivity shock and make a one-period fixed-wage take-it-or-leave offer to workers. The subgame-perfect equilibrium displays a number of interesting and economically important features. First, because firms are reluctant to cut the wage below the reference wage, workers already inside the firm experience downward wage rigidity: in response to moderate decreases in productivity, the wage remains unchanged. Second, if productivity falls too much to pay the reference wage, firms do not induce discretionary effort. Then, depending on the importance of discretionary effort, they either retain the worker with a low wage knowing that this will result in low effort, or lay the worker off, both of which are inefficient. Third, since newly hired workers had expected to be employed with lower probability, they come in with lower expectations regarding how much they will make, so their wage is more flexible than that of existing workers.

3.3 *The Effect of Extrinsic Motivation on Intrinsic Motivation*

As mentioned in section 3.1, a number of theories in behavioral economics imply that an agent might be willing to exert costly effort for reasons other than explicit material incentives. At the same time, all theories assume that agents are also responsive to explicit incentives. The recognition that an agent may be motivated by different types of incentives raises the natural question of how different incentives interact. A number of papers in the literature suggest that extrinsic incentives can crowd out intrinsic motivation, so that it can be optimal to employ weaker explicit incentives than classical models would suggest.

In addition to the stylized fact that explicit performance incentives are often quite weak, different strands of research in this area are motivated by (and consistent with)

[10] In Daido and Murooka (2012), team incentives emerge under some circumstances as an optimal way to manage agents' loss aversion, even when individual incentive pay is feasible. In particular, suppose that a hard-working agent's probability of success in her project is low, and consider first individual incentive pay that is based on the performance of only her own project. Then, working hard creates an expectation of high pay and thereby exposes the agent to a likely sense of loss, dampening the incentive to work. By paying the agent also for other agents' successes, the principal reduces the probability of loss, increasing the incentive to work. From a single agent's point of view, getting paid for another agent's success is equivalent to the principal sometimes "turning a blind eye"—i.e., forgiving failure—but Daido and Murooka argue that team incentives are a credible way to implement such a blind-eye policy.

[11] For related work, see also Jofre et al. (2012).

somewhat different kinds of empirical observations. Some researchers have found circumstances under which explicit incentives have a *contemporaneous* negative effect on effort—that is, effort is lower when greater explicit incentives are in place. In particular, Gneezy and Rustichini (2000) document that imposing a modest fine on parents for picking up their kids late at a day-care center increases the prevalence of late pickups. And Falk and Kosfeld (2006) find in an experiment that a "controlling contract"—which prohibits the lowest contribution levels by the agent—also lowers contributions.[12] This evidence is slightly different from that studied extensively since Deci (1975) and reviewed for instance by Deci, Koestner, and Ryan (1999), whereby providing extrinsic incentives to perform a task leads individuals to engage in the same or similar tasks less often *once the incentives are removed*.

Given the topic of this subsection, I organize existing research according to the mechanism through which explicit incentives act on intrinsic motivation.

Informed principals. A number of theories employ an informed-principal framework in which the principal has superior information regarding a variable that affects both his incentive-design problem and the agent's willingness to exert effort, so that a contract with strong explicit incentives can undermine the agent's motivation through its effect on her beliefs. In Bénabou and Tirole (2003), the principal chooses a bonus for task completion knowing the agent's ability or cost of completing the task, while the agent receives only a signal of this variable. If the principal knows that the cost is high, he is worried that the agent might also have received a signal that the cost is high, so that he chooses a higher bonus to motivate the agent. As a result, a large bonus for completing a task becomes a signal of task difficulty for the agent, which reduces her motivation to work in the future. In a closely related paper with multiple agents, Fang and Moscarini (2005) suppose that the principal receives a private signal about each agent's ability, and can decide whether to make different contract offers for agents with different signals. If agents are overconfident, the principal might prefer not to differentiate contracts, as the information from differentiated contracts would (given their overconfident prior) provide information to agents that is on average bad news and, hence, would lower intrinsic motivation.

Other papers have a similar theoretical structure to that of Bénabou and Tirole, but the information the principal has is of a different nature. In Sliwka (2007), there are three types of agents in the population: selfish, fair, and conformist. Selfish agents care only about their material well-being, whereas fair agents care also about the principal's payoff. Conformists behave like the other type (selfish or fair) that they believe is in the majority in the rest of the population. The principal knows the share of fair types, which affects both his optimal incentive scheme and conformists' willingness to exert effort. And in Herold (2010), the agent may be "trustworthy"—motivated to exert effort even without explicit incentives—or "untrustworthy"—willing to exert effort only if explicitly motivated to do so. The principal has privately held beliefs about the agent's probability of being trustworthy (his "trust" in the agent). This always affects the profitability of different incentive schemes. It can also affect the agent if she takes nontrusting behavior as a hostile act and reciprocates it, or if the principal also takes an action that is

[12] For the above negative effect to occur, it seems crucial that the control is imposed by the principal rather than exogenously. For instance, List (2007) finds that when allocating money between themselves and a recipient, many fewer dictators give a positive amount if they have the option of taking away money than if they do not have this option.

either a complement or a substitute to the agent's action.

The above informed-principal models can naturally explain evidence on the future negative effect of explicit incentives mentioned at the beginning of the subsection: once the agent draws the respective negative conclusion from a high-powered incentive contract in each case, she will be less motivated in future tasks. But these models do not provide full explanations for the *contemporaneous* negative effect of explicit incentives. While the models explain why explicit incentives might be weakened by the crowd-out of intrinsic motivation, they remain positive reinforcers in the sense that, if a given worker receives strong rather than weak incentives (among equilibrium contract offers), she exerts greater effort on average. The intuition derives simply from the negative future effect of explicit incentives: since strong explicit incentives have negative future consequences, the principal uses them only if they increase current effort.[13]

Social signaling. Bénabou and Tirole (2006) develop a model in which agents have heterogeneous degrees of altruism, and all types prefer to be perceived by others as altruistic. If there are no explicit incentives, prosocial behavior signals a high degree of altruism, so that many agents are willing to act prosocially for image reasons. But with explicit incentives for prosocial behavior, such behavior could be due to materialism and is thus a less convincing signal of altruism, reducing the image motivation for prosocial behavior. This crowding out reduces the effectiveness of extrinsic incentives, and under some conditions can induce a negative net effect of extrinsic incentives. Bénabou and Tirole's model explains the evidence on the contemporaneous negative effect of explicit incentives on prosocial behavior mentioned above. Since removing the explicit incentives reinstates the signaling motive, however, the model does not seem to predict a future effect.

Reciprocal behavior. Another potential explanation for the positive contemporaneous impact of weak explicit incentives on effort is reciprocal behavior. The original model of intentions-based reciprocity, Rabin (1993), defines a kind act as one that sacrifices the person's own material well-being to increase the material well-being of the other player, and assumes that a reciprocal player prefers to respond to a kind act with a kind act. While in this model weak incentives cannot generate an increase in effort if the agent is known to be reciprocal—if they did, they would benefit the principal and hence could not be considered kind—von Siemens (2011b) shows that they can do so under some circumstances if both selfish and reciprocal agents are present. Since selfish agents take advantage of weak explicit incentives at the expense of the principal, choosing such incentives is kind vis a vis selfish agents. If there are many selfish agents, therefore, choosing weak explicit incentives is on average a kind act, to which reciprocal agents respond with high effort. Furthermore, this is profitable for the principal if reciprocal agents respond sufficiently strongly to the principal's kindness. Ellingsen and Johannesson (2008) demonstrate a similar result in a closely related model that differs in why exactly the agent responds to a kind act kindly: the principal's choice of incentive contract reveals how nice the principal is, and some agents prefer to be nice to nice principals.

[13] Nevertheless, Bénabou and Tirole (2003) point out that the unconditional correlation between explicit incentives and effort could be negative, which could lead a naïve observer to incorrectly conclude that extrinsic incentives crowd out contemporaneous intrinsic motivation. This can happen because explicit incentives are more likely to be used when the principal's information points to low agent motivation, which is also when the agent is less likely to work.

Anticipatory utility. Immordino et al. (2011) consider the problem of a risk-neutral principal offering an output-contingent wage to a risk-neutral agent with unlimited liability, who receives a private signal about her return to effort before choosing her effort level. The agent derives utility from anticipating high wages, so that she has an incentive to suppress signals that indicate a low return to effort. If anticipatory utility is sufficiently important relative to actual outcomes, then (in contrast to the prediction of a classical model) implementing the first-best level of effort is impossible. Intuitively, the agent exerts high effort only if she is sufficiently rewarded for a good result, but this makes a good result very desirable, increasing the motive to suppress bad news and hence distorting information use.

3.4 *Moral Hazard and Overconfidence*

Individuals often overestimate their ability to do well in a task, either in the sense of *baseline optimism*—overestimating one's performance given an effort level—or in the sense of *control optimism*—overestimating the return to effort—or both.[14] Researchers have explored a few implications of these phenomena for moral hazard.

De la Rosa (2011) and Gervais, Heaton, and Odean (2011) study moral-hazard models in which the risk-averse agent is both baseline optimistic and control optimistic. Since a control-optimistic agent requires lower-powered incentives than a rational agent to implement a given level of effort, the principal responds to moderate overconfidence by lowering the power of incentives. This allows the agent to bear less risk, increasing social welfare. Interestingly, if there is competition between principals, the agent receives all of this increase in social welfare, so that her bias *benefits* her in equilibrium.[15] In contrast, because baseline optimism implies that the agent is overly willing to accept high-powered contracts with a low base wage, the principal responds to large overconfidence with very high-powered contracts to exploit the agent's mistake, and in equilibrium the agent's incentive constraint may not bind. In this range, the contract is more like an exploitative contract as defined in section 6, and the large overconfidence always hurts the agent.[16]

In the context of unemployment insurance and job-search behavior, Spinnewijn (2012) assumes that agents are baseline optimistic and may be control pessimistic or control optimistic, and contrasts the response of a social planner (who maximizes agents' utility subject to a budget constraint) and a competitive profit-maximizing firm. When choosing the level of benefits for a biased agent, the social planner accounts for the effect of a change in search behavior on the agent's welfare, so that he responds to control optimism by increasing benefits to lower search effort. A profit-maximizing firm, in contrast, does not care about the agent's welfare, so it does not face the same consideration. But because the agent's baseline optimism lowers her demand for insurance, the firm responds by lowering benefits.

[14] The above distinction is taken from Spinnewijn (2012).

[15] This result is related to the point made by Mullainathan et al. (2012) that underutilization of medical care due to present bias can lower the moral-hazard cost of providing insurance, potentially benefitting the insured.

[16] See also Santos-Pinto (2008) for a closely related analysis that focuses on identifying conditions under which agent overconfidence benefits the principal in a single-principal setting. If effort is observable—so that only the participation constraint is relevant—the principal benefits from agent overconfidence both because the agent is overly willing to accept a given contract and because he can exploit her as above. If effort is unobservable, the principal benefits if the agent is control optimistic (and other, more technical, conditions hold) as control optimism makes it easier to satisfy the agent's incentive-compatibility conditions as well.

3.5 *Potential Future Directions*

In this section, I discuss some outstanding issues in the literature in the same order in which the related research appears above.

Further progress in the literature exploring optimal moral-hazard contracts with social preferences seems somewhat hindered by the lack of a theory of reference-group formation—the determination of which other individuals' outcomes an agent cares about. This not only makes it more difficult to test or judge the economic importance of the above results, but also prevents researchers from asking natural questions, such as how reorganizations affect the reference group and hence behavior, in a theoretically disciplined manner.

While existing models with inequity-averse decisionmakers can explain gift exchange in the laboratory and in some organizations, they do not predict the phenomenon when—as in many real-life settings—the employer is significantly wealthier than most employees and cares mostly about his own material outcomes. Because in this case a generous wage does not put the worker ahead, gift exchange should not result. And because a selfish principal does not pay a bonus she does not have to, the voluntary-bonus system should not work.

Deepening the above puzzle, Netzer and Schmutzler (2012) show that another major model of social preferences, Rabin's (1993) model of intentions-based reciprocity, cannot convincingly account for gift-exchange behavior, either, so that to date there is no compelling explanation of this simple phenomenon. At first glance, Rabin's model appears to predict gift exchange: offering a high wage seems to be a kind act, to which the agent should naturally respond with her own kind act of a high effort. In contrast, Netzer and Schmutzler show that if the principal is selfish, this mechanism cannot work to increase effort. By definition, a selfish player never sacrifices his own payoff for that of the other player, so the principal's choice of wage will never be considered kind by the agent. As a result, the agent never exceeds the materially cost-minimizing level of performance.[17] This means that gift exchange is either not about reciprocity, or a different or more complex reciprocity mechanism is involved. A model with agent heterogeneity similar to that of von Siemens (2011b), discussed above, might provide a starting point, but whether such a model will prove to be a compelling general explanation requires further research.

The literature on loss aversion does not seem to face similarly central puzzles, but there are also plenty of unanswered questions. For instance, note that the contract terms for loss-averse agents above emerge because the firm offers partial insurance to an agent who is first-order averse to risk. From this perspective, it is interesting to note that (as far as I am aware) no paper has carefully incorporated loss aversion into an insurance model with moral hazard and derived implications for optimal contracts. The logic of Herweg, Müller, and Weinschenk (2010) seems to imply that in many settings, an insurance contract with a simple deductible—whereby the agent pays a deductible if losses exceed a threshold—is optimal.

While the literature on the interaction between extrinsic and intrinsic motivation is one of the most exciting and productive in behavioral contract theory, it also calls for substantial future research on when different sources of motivation are likely to dominate behavior. As explained above, all of the

[17] Nevertheless, if the agent could choose *lower* levels of performance—that is, exert costly effort to punish the principal—her wage can be much higher than that of a selfish agent in the same situation, materially benefiting her and hurting the principal. In this case, the principal chooses a high wage not to induce discretionary high effort, but merely to avoid the punishment that lower wages would trigger.

existing models explain only a part of the evidence and make orthogonal or opposing predictions on the other evidence—with no sound guidelines as to where exactly each model applies. One possible reason for the confusing state of the literature may be that, unlike extrinsic motivation, intrinsic motivation is a complex multifaceted phenomenon that is poorly understood. Individuals may derive intrinsic motivation from many sources, including those in the models above, and perhaps additional motives such as the sense of being in a good organization or the sense of doing good work. Both more evidence and more theory seems necessary to sort out what situations evoke the different kinds of intrinsic motivation identified by researchers.

The observation that a bias—here, slight overconfidence—can benefit the agent raises the natural question of when this is likely to happen. Clearly, fixing the contract the agent signs, a bias always (weakly) reduces her welfare. Hence, for a bias to benefit the agent, it must change contract terms in her favor. This can happen if the agent's bias leads her to either overvalue her outside option or undervalue interacting with the principal—both forcing the principal to offer better terms. The former is the case above, where competing principals offer terms the agent overvalues. The latter seems less likely to happen in many settings, as the principal—in an attempt to induce the agent to accept—will try to write contracts the agent overvalues rather than undervalues. But the question of when biases can benefit an agent has not received sufficient attention in the literature. Research on this question might draw insights from the large classical literatures on commitment and reputation, which also recognizes that appearing nonrational can benefit a player. For instance, the beneficial effect of overconfidence above is similar to that of a commitment to work hard, although the particular type of cost overconfidence implied has, to my knowledge, not been studied in the classical literature.

4. *Asymmetric Information and Screening*

In classical screening models, a principal (e.g., an airline) faces agents whose preferences (e.g., willingness to pay for a ticket) he does not know. Ideally, the principal would like to contract with multiple agent types, differentiating the contract (e.g., the price and conditions of the airline ticket) according to the agent's preferences. The principal's main constraint—captured in the incentive-compatibility constraint that a more profitable type not take a less profitable option—is that the agent may not reveal that she is a profitable type (e.g., that she has a high value for a ticket). The central issue in contract design is how to trade off minimizing this informational advantage with the objective of achieving gains from trade. This section summarizes the bulk of the literature incorporating ideas from psychology and economics into situations of hidden information and screening.[18]

4.1 *Commitment versus Flexibility*

It has been understood that individuals with self-control problems benefit from commitment, and hence might be willing to sign contracts that restrict their choices in some way. At the same time, uncertainty can make it inefficient to remove all choice, thus generating a possible tradeoff between commitment and flexibility. A number of papers explore aspects of this contracting question.

In an early contribution, DellaVigna and Malmendier (2004) study the optimal two-part tariff of a firm facing a consumer who has known present-biased preferences and

[18] Section 6 discusses a different set of screening issues related to exploitative contracts, which arise because different agents might exhibit the mistake the principal is looking to exploit to different extents.

can make an investment such as saving or exercising. Investment carries an immediate cost that is unknown to the firm, and generates a fixed future benefit that is known. If the consumer is sophisticated regarding her present bias, the optimal two-part tariff implements the first-best outcome by imposing a subsidy that equals the amount by which the consumer undervalues the future benefit. Hence, the contract induces more future-oriented behavior without compromising the consumer's flexibility to respond to cost shocks. While this contract achieves first-best despite asymmetric information about the cost, it is clear that the optimality is specific to the problem and form of uncertainty: for instance, if the consumer's benefit was also uncertain, the first-best could not be achieved with a two-part tariff.

Amador, Werning, and Angeletos (2006) study optimal contracting in a setting where the first-best is typically not achievable. They consider a consumption–savings problem in which a present-biased agent faces a privately observed taste or needs shock that affects her marginal instantaneous utility of consumption and that she does not initially know.[19] Thinking of the consumer's late self as the agent and her early self (or a social planner maximizing the expected utility of her early self) as the principal, they characterize the optimal commitment, defined as the optimal subset of the agent's budget set from which she will be allowed to choose her savings level. Amador et al. show that the optimal commitment always features a minimum savings rule akin to those observed in many retirement systems. Intuitively, it is not optimal to allow the highest types—those with the greatest taste for immediate consumption—to choose higher consumption levels than slightly lower types, as this would mean that the highest types are overconsuming from an ex ante point of view. In addition, Amador et al. show that it is often optimal to allow for complete flexibility above the minimum savings requirement. Roughly, whatever flexibility in savings the principal allows above the minimum, the agent tends to overconsume in that range from an ex ante point of view. Since the principal cannot do much about the overconsumption, he might as well allow the agent to adjust consumption to the taste shock.[20]

Galperti (2012) studies a setting where, much like in DellaVigna and Malmendier (2004), the principal can implement first-best if she knows the agent's degree of time inconsistency, but a screening issue arises because there are both time-inconsistent and time-consistent agents in the population. A flexible contract that solves the time-inconsistent agent's commitment problem must offer rewards for saving. But because a time-consistent agent can expect to receive the rewards more often, she derives rent from such a flexible contract. To lower the time-consistent agent's information rent, then, the principal curtails the flexibility of the time-inconsistent agent's contract, restricting both high and low levels of savings. Galperti argues that this feature is consistent with restrictions on retirement savings devices in the US.

Esteban, Miyagawa, and Shum (2007) study a monopolist's nonlinear pricing problem in which the heterogeneity in

[19] Technically, this kind of situation is called a "hidden knowledge" problem, because—as distinct from non-linear pricing or screening problems—exogenous asymmetric information emerges after the contract is signed.

[20] This insight does not always hold due to a subtle consideration in the principal's problem. By eliminating a range of possible consumption levels, the principal forces those types who would otherwise have preferred to consume in that range to consume either less or more. The former response is welfare-increasing, while the latter response is welfare-decreasing. The former force dominates the combined harmful effects of the latter force and lowers flexibility if sufficiently more types respond by consuming less, which is the case if the density of types decreases sufficiently quickly.

agent preferences is due to heterogeneity in temptation disutility in the sense of Gul and Pesendorfer (2001).[21] In Gul and Pesendorfer's model, the agent has preferences consisting of commitment utility—the utility of an option if the agent can perfectly commit to it in advance—and temptation disutility, where she suffers the latter if she does not choose the most tempting option from her choice set. Whereas with hyperbolic discounting the agent would like to commit her future behavior because she will have different preferences in the future than she does now, with temptation disutility the same demand for commitment arises because she wishes to lower temptations. The impact of temptation disutility on the optimal menu turns out to depend on the kind of temptation agents experience. If all agents experience *upward* temptation—whereby high consumption is more tempting than low consumption (e.g., cigarettes)—then the optimal menu is a singleton, so that the seller does not separate consumers with different preferences at all. Intuitively, because trying to take advantage of an agent's temptation to sell her more would make her less willing to participate, the seller just chooses to maximize agents' commitment utility, which is possible with a singleton since (by assumption) agents have homogenous commitment utility. If all agents experience *downward* temptation—whereby lower consumption is more tempting than higher consumption (e.g., exercise)—the optimal menu is identical to that with standard preferences equal to ex post preferences. Intuitively, since not buying is the most tempting alternative and consumers can always choose this alternative, increasing the menu does not increase temptation disutility, so that it does not make consumers less willing to participate. Hence, the seller screens consumers according to ex post preferences. The logic of these results (especially those for upward temptation) relies crucially on the assumption that individuals differ only in temptation disutility, and as the authors point out, it would be interesting to study the same problem if commitment utility also differs between types.

[21] Esteban and Miyagawa (2006) analyze the competitive case.

4.2 *Overconfidence and Screening Issues Related to False Beliefs*

The observation that some individuals have overly positive views about themselves or their prospects raises the question of how to optimally screen agents with different levels of overconfidence. Before discussing research on this question, it is worth noting a general limit to screening beliefs according to their accuracy: because two individuals with the same beliefs about future outcomes and same ex ante preferences choose from a menu in the same way, separating them at the contracting stage by means of self-selection is impossible. In particular, this is the case even if one has correct and the other incorrect beliefs, and hence the two agents will behave differently given the contract. As a result, a contract signed by agents with given beliefs are often priced according to some average of the actual outcomes of these agents. In these situations, a biased agent exerts an externality on rational agents. The externality is negative if the biased agent is less profitable than the rational agent (e.g., in insurance, where an overconfident agent underestimates her risk), and positive if the biased agent is more profitable (e.g., in many exploitative contracts discussed below).

Overconfidence in the insurance market. Sandroni and Squintani (2010) study insurance contracts when some high-risk agents believe that they are low-risk, and show that the presence of overconfident agents has qualitative observable implications in a competitive (but not in a monopolistic) insurance

market. To understand their results, recall that in the seminal competitive insurance model of Rothschild and Stiglitz (1976) with rational agents, high-risk and low-risk agents separate in equilibrium and the price of insurance is actuarially fair for both types, so that the price of insurance in each class does not depend on the proportion of types in the population. In addition, because high-risk agents able to fully insure at actuarially fair prices do not want small amounts of insurance even at low prices, low-risk agents receive at least partial insurance. Overconfidence changes each of these predictions. If some agents who believe themselves to be low-risk are in fact high-risk, the group of low-risk and overconfident—who by the above logic cannot be distinguished because they choose contracts in the same way—receive more expensive insurance that depends on the proportion of overconfident in the population, generating heterogeneity in insurance prices within a risk class. In addition, since the low-risk and overconfident believe that their insurance is overpriced, they may not buy insurance.

Sandroni and Squintani (2007) reconsider the scope for Pareto-improving interventions in the insurance market when overconfident agents are present. Although in a rational competitive insurance market low-risk agents receive cheap insurance, such insurance is partial to discourage high-risk agents from taking it. In this equilibrium, low-risk agents would prefer to buy more insurance at the same price, so that a government policy of mandatory insurance—which allows low-risk agents to get more insurance at reasonable prices—can be Pareto-improving. But with overconfident agents, the group of low-risk and overconfident are offered more expensive insurance, and since this group believes that they are receiving insurance above the actuarially fair price, they may not want much insurance. If this is the case, the group of low-risk and overconfident can choose a perceived-optimal insurance contract that is rejected by high-risk consumers. As a result, low-risk and overconfident consumers prefer *not* to buy more insurance at prevailing prices. Hence, contrary to common intuition that biases increase the scope for government intervention, in this case mandatory insurance is not Pareto-improving.[22]

Contrary to the prediction of classical insurance models that higher-risk consumers buy more insurance, recent empirical research finds a zero or negative correlation between risk and insurance coverage in many circumstances (for instance Chiappori and Salanie 2000; Finkelstein and McGarry 2006). To explain this puzzle, Spinnewijn (2013) studies insurance markets with optimistic agents and both moral hazard and asymmetric information. If a single-crossing property holds, the only way for a principal to separate agents is to offer them different amounts of insurance, and this separates agents (roughly) only according to baseline optimism. But because control optimism affects how much precaution an agent takes when insured, the correlation between risk and insurance coverage depends on the correlation between baseline optimism and control optimism.

Overconfidence and debt financing. Manove and Padilla (1999) consider a model in which overoptimistic entrepreneurs ask for loans from banks to finance overly large projects relative to the most productive use of money, and banks cannot distinguish overoptimistic entrepreneurs from realists. Manove and Padilla's main result

[22] Schumacher (2012) studies a model with a related effect, where—similarly to the overconfident in Sandroni and Squintani (2007)—naïve present-biased agents exert a negative pricing externality on sophisticated agents by not taking enough precautions, making government intervention less desirable. Nevertheless, in Schumacher's (2012) model, government intervention to induce taking more precaution is Pareto-improving if the share of naïve agents is substantial.

is that banks are often willing to fund too many projects, from a social point of view. Intuitively, banks care only about recouping their investment in expectation, and do not internalize the fact that an optimistic entrepreneur could have used the money better elsewhere. Furthermore, because collateral requirements help ensure that banks recoup their loans, they exacerbate the excessive lending problem.

4.3 *Other Topics in Screening*

This section discusses various other research on screening.

Using framing for screening. Salant and Siegel (2013) consider the screening problem of a seller who can temporarily manipulate buyers' preferences. The seller chooses quality and price for two types of buyers, and in addition may employ a "frame"—such as a salient comparison price in a shop or a glitzy environment in a casino—to change the types' utility functions. Crucially, however, the effect of the frame is fleeting, and either (i) the buyer returns the item if she later realizes that she does not value it above her outside option, or (ii) the buyer anticipates the manipulation and stays away from the store if she does not like what she will buy. Salant and Siegel show, roughly, that such framing can help the seller if and only if it relaxes the high type's incentive constraint without increasing the low type's incentive to mimic the high type. In this case, the principal can offer the low type a cheap product while manipulating the high type into buying the expensive product, eliminating the high type's information rent and increasing efficiency of provision to the low type.

Sorting in the labor market. Kosfeld and von Siemens (2011), von Siemens (2011a), and von Siemens (2012) study labor-market outcomes for employees with different types of social preferences. In the monopsony model of von Siemens (2011a), employees might be selfish or inequity-averse with respect to coworkers. Since a worker can affect only her own outcome and both selfish and inequity-averse workers prefer more money to less (holding others' outcomes constant), the firm cannot differentiate selfish and inequity-averse workers who have been hired. In order to employ low-ability inequity-averse individuals, therefore, a firm needs to pay a premium to compensate these workers for falling behind coworkers. To economize on this cost, the firm either reduces the gap between high- and low-ability workers by distorting production, or excludes low-ability inequity-averse workers by paying no premium. In contrast, von Siemens (2012) shows that in a competitive market, the existence of multiple firms allows high- and low-ability workers to sort into different firms, reducing the impact of social comparisons, and hence often allowing low-ability inequity-averse workers to be employed as well. In a related setting, Kosfeld and von Siemens (2011) assume that some workers are selfish—i.e., exert individual effort if compensated for it, but never exert team effort—while other workers are conditional cooperators—i.e., exert team effort if coworkers do too, and prefer to work in a team environment. In equilibrium, firms offer two types of contracts: a high-powered incentive contract designed for the selfish workers, and a low-powered incentive contract designed for the conditional cooperators. To prevent selfish workers from accepting the latter contract, its wage must be relatively low, so that it may be profitable for the firm despite competition. This means that if team effort is sufficiently important, firms with lower pay are more productive.

Loss aversion and screening. Hahn et al. (2010) analyze a monopolist's optimal menu when consumers are loss averse and do not know their willingness to pay in advance. Just

as a loss-averse employee dislikes the risk inherent in a wage schedule that discriminates finely between performance levels (see section 3.2), a consumer who does not initially know her willingness to pay dislikes the risk inherent in a menu that discriminates finely between different willingness-to-pay realizations. Hence, the seller often offers a small number of products relative to the heterogeneity in the population.

Carbajal and Ely (2012) assume instead that consumers know their types in advance and have a type-dependent reference point relative to which they evaluate outcomes. Carbajal and Ely study how the optimal menu depends on the reference-point function, and also derive properties of self-confirming reference consumption plans—where a type's consumption in equilibrium coincides with her reference point. In contrast to an individual-decision-making setting—where an increase in the reference point always hurts the agent—a higher self-confirming reference consumption plan can benefit both the seller and some agents. Intuitively, a higher reference point leads the seller to exclude fewer low types from the market (who, due to their higher reference point, value the product more highly), and as a result of this market expansion, higher types receive higher information rents.

4.4 *Potential Future Directions*

Screening seems to be an understudied topic in behavioral contract theory. Notably, behavioral research has not been incorporated into optimal income taxation, although some psychological phenomena (such as hyperbolic discounting, anticipatory utility, and overconfidence) seem important in labor, leisure, and consumption choices. More generally, there seems to be little research on how insights from psychology and economics affect classical screening problems in which the private information concerns a standard preference parameter; in much of the research above, the contract does not have a classical purpose or the private information concerns a parameter in psychology and economics.

Another important question is the extent to which a social planner can screen agents—some of whom might be behaving optimally, and some suboptimally—in the context of noncoercive interventions aimed at improving individual consumer decisions.[23] For instance, O'Donoghue and Rabin (2003) illustrate that a "sin license," whereby consumers would be required to buy a moderately priced license to smoke, often dominates "sin taxes" because it separates time-consistent consumers from both naïve and sophisticated present-biased consumers. Intuitively, neither type of present-biased consumer obtains the license—naïve consumers because they do not believe they will smoke in the future, and sophisticated consumers because they want to prevent their future selves from smoking—but time-consistent consumers who are making an optimal decision to smoke do.

5. *Auction Theory and Mechanism Design*

Mechanism design is the study of what outcomes can be achieved, and how, when a principal is interacting with multiple agents with private information. The central consideration in the principal's problem is how to set up the game form so that agents reveal their private information. This section reviews applications of behavioral-economics ideas to mechanism design. The bulk of existing research concerns the theory of auctions, a prominent class of mechanisms for selling to buyers with unknown valuations for the product. The section also describes the limited amount of work on other issues in mechanism design.

[23] See Thaler and Sunstein (2008) and Camerer et al. (2003) for arguments in favor of such interventions, as well as some examples.

5.1 *Loss Aversion in Auctions*

When applying loss aversion to auction theory, predictions depend on whether a bidder experiences gain–loss utility separately from money and from the product being sold (separate evaluation), or jointly from the net utility of the transaction (net evaluation). In commodity auctions—where the payment is monetary and the product is nonmonetary—the former assumption seems more appropriate, while in induced-value laboratory experiments—where both the payment and the "product" are monetary—the latter assumption seems to apply.[24] As Lange and Ratan (2010) argue, this implies both that one must use different theories to interpret induced-value laboratory auctions and commodity auctions, and that many experimental findings may not apply to real-world settings.[25]

Lange and Ratan (2010), Eisenhuth (2010), and Eisenhuth and Ewers (2012) analyze the effect of expectations-based loss aversion as modeled in Kőszegi and Rabin (2007) on the revenue ranking of standard sealed-bid private-values auctions. Kőszegi and Rabin's model implies that individuals are first-order averse to local risk, and hence bid more aggressively in auction formats that are better at insuring them against the uncertainty from participation. This behavior generates a revenue ranking between the three most commonly analyzed auction formats—first-price, second-price and all-pay auctions—that depends on whether bidders use separate or net evaluation. Under separate evaluation, an all-pay auction is less risky than a first-price auction because payment is deterministic, and a first-price auction is less risky than a second-price auction because payment conditional on winning is deterministic. In fact, Eisenhuth (2013) establishes that under loss aversion with separate evaluation, any optimal mechanism features a riskless payment, and the optimal auction is an all-pay auction with a minimum bid. In contrast, with net evaluation the ranking between the all-pay and first-price auctions is reversed. Intuitively, the gap between the net utility from winning and not winning equals the product's value with an all-pay auction but only the surplus with a first-price auction, so the latter appears less risky for a net evaluator. In fact, Eisenhuth (2010) proves that in this case, the optimal auction is the first-price auction with a minimum bid.[26]

Another possibility explored by researchers is that the reference point of auction participants is partially determined by the reserve price. In particular, Rosenkranz and Schmitz (2007) specify the reference point in money as a weighted average of the reserve price and an exogenous parameter, so that an increase in the reserve price directly raises bidders' willingness to pay, and hence their bids. When either the number of bidders or the exogenous component of the reference point is high, the bid-raising benefit of increasing the reserve price dwarfs the cost from the increased risk of not being able to sell, and therefore even a small degree of loss

[24] The perspective that individuals use separate evaluation when a real commodity is traded off with money is consistent with a large body of evidence, such as the endowment effect, that is commonly interpreted in terms of loss aversion. Nevertheless, Eisenhuth and Ewers (2012) present some suggestive experimental evidence that a model with net evaluation better describes a commodity auction, as well.

[25] Lange and Ratan (2010) find that with net evaluation, bidders "overbid" in a first-price private-values action, while with separate evaluation they underbid. Intuitively, with net evaluation, loss-averse agents experience a sensation of loss compared to the expected payoff when losing the auction, motivating them to bid higher. In contrast, with separate evaluation, loss aversion in the money dimension kicks in when the agent wins the auction, and thus results in agents underbidding in order to avoid having to pay more than expected.

[26] From a theoretical perspective, a weakness of the above literature is that (as the authors note) it makes predictions qualitatively similar to those of appropriately specified models with classical risk-averse bidders. Of course, the implications of loss aversion could be quantitatively more important.

aversion can lead to a significant increase in the optimal reserve price. Shunda (2009) assumes that bidders' reference points depend on the seller's reserve price as well as his "buy price"—a price at which a bidder may purchase the good from the seller before the auction starts. Increasing the buy price raises the bidders' willingness to pay, and hence leads to a larger pool of participants and higher bids in the auction stage, but lowers the probability that a bidder buys the item in the first stage. The optimal buy price trades off these two effects.

5.2 *Other Topics in Auctions*

Nonequilibrium thinking. Jehiel and Lamy (2012) ask why sellers often employ absolute auctions (auctions with a reserve price of zero) despite positive value for the object, and why sellers often set secret reserve prices. In the spirit of Jehiel (2005) and Eyster and Rabin (2005), Jehiel and Lamy propose that some bidders do not understand how the potential for getting a good deal varies with the auction format. Absolute auctions can be used to attract bidders who underappreciate how product quality and the participation rate depend on the auction format, as these bidders overestimate the quality of goods in such auctions and may also underestimate the participation rate. And auctions with a secret reserve price can be used to attract bidders who do not understand how secret reserve prices differ in distribution from public reserve prices, failing to anticipate that a secret reserve price is likely to be higher than public reserve prices.

Augenblick (2011) studies the penny auction, an auction format in which agents bid for items in predefined increments (often 1 cent) and have to pay a nonrefundable fixed price for each bid. Augenblick documents that bidders severely overbid in these auctions, generating for instance an average profit margin of 104 percent in auctions for direct cash payments. He argues that the penny auction engages a naïve sunk-cost fallacy, whereby bidders are reluctant to stop bidding in an auction when they have spent money on it, and do not predict the full extent of this effect. He shows that the profit-maximizing supply of auctions is constrained by the consideration that an auction attracting few bidders tends to end early and generate large losses for the auctioneer. The same consideration also creates an effective barrier to entry, as competitors who cannot initially attract many bidders must absorb large losses.

Crawford et al. (2009) begin studying optimal auction design with level-k bidders. Following Camerer, Ho, and Chong (2004) and Crawford and Iriberri (2007), their model distinguishes between bidders of different levels of strategic reasoning. L0 types follow some simple bidding strategy, such as bidding a random number or always bidding their true valuation. L1 types formulate their strategy as the best response to the L0 strategy. For example, if L0 types bid randomly, then L1 types believe that winning does not reveal any information about the value of the object, so that they—similarly to fully cursed players in Eyster and Rabin (2005)—overbid compared to the predictions of the classical model. L2 types best respond to L1 types. Crawford et al. (2009) find that the optimal reserve price may either be higher or lower than in an equilibrium model, while seller revenue—due to overbidding by many players—is usually higher. They also give an example of an exotic auction specifically designed to exploit bidders' nonequilibrium thinking that can yield arbitrarily large revenues.

Anticipated regret. Engelbrecht-Wiggans (1989) and Filiz-Ozbay and Ozbay (2007) study a model in which bidders anticipate feelings of regret in case their bid turns out to be suboptimal ex post. Losers of a first-price sealed-bid auction or a Dutch auction may feel regret when confronted with the information that the winning bid was lower than

their valuation, which means that they could have walked away with a positive surplus had they followed a more aggressive strategy. On the other hand, winners of a first-price auction may also feel regret when discovering that the second-highest bid was lower than their own, so that they could have won at a lower price. Anticipating "loser regret" induces participants to make higher bids than otherwise, while anticipating "winner regret" induces them to make lower bids. These emotions are only triggered when the regret-inducing information is revealed to participants, so a seller can influence bidder behavior through information provision. In a setting that triggers only loser regret, participants overbid relative to their consumption utility.[27]

5.3 *Mechanism Design*

Beyond auction theory, there is little work on the psychology and economics of mechanism design. One exception concerns the implications of social preferences in public-interest situations, those in which everyone benefits if everyone gives up personal material gain for the sake of others. In a number of different models, authors find that social preferences tend to increase efficiency. Kucuksenel (2012) considers mechanism design with altruistic agents and shows that more altruistic agents—those who attach a larger weight to others' material payoffs—are more likely to produce a public good that is efficient to produce, and are more likely to trade when it is efficient to do so. The intuition is simple: more altruistic agents care more about social efficiency, so it is easier to implement socially efficient outcomes with them. Bierbrauer and Netzer (2012) establish a similar result for agents with intentions-based social preferences in the sense of Rabin (1993), where a player is willing to sacrifice her own material payoff for the sake of the other player if and only if she believes the other player is similarly kind. The codependence of intentions makes kindness an equilibrium phenomenon. If in equilibrium agents have positive intentions toward each other, then the observation that they can achieve better outcomes is similar to that in Kucuksenel (2012). But since intentions are endogenous, a key part of the designer's problem is to set up a mechanism in which agents can develop positive intentions toward each other. Bierbrauer and Netzer show that the designer can achieve this by adding unused options to enrich oneself to the mechanism, so that truth-telling will be considered a kind action. Even so, there is always an equilibrium in which players are unkind to each other, so the designer must make sure that the kind equilibrium is played—for which the theory provides no guidance. Finally, Carmichael and MacLeod (2003) make a related point in the context of holdup: they show that caring about the other party's investment—a notion akin to a preference for equity—can soften the holdup problem in contracting.[28,29]

[27] See also Cramton et al. (2012) for an application of this concept to clock auctions such as those commonly used for selling radio spectrums, electricity, or gas. And Roider and Schmitz (2012) study optimal reserve prices in a related model that assumes bidders get utility directly from winning or losing, independently of whether they could have done better with another strategy.

[28] Desiraju and Sappington (2007) consider a nonlinear pricing problem with two agents who are averse to inequity in the total ex post payoff they receive, a setting that does not neatly fit the public-interest environment above. While with standard preferences a two-agent screening problem reduces to a single-agent screening problem, with inequity aversion it is fundamentally multi-agent mechanism design, as each agent cares about the other's surplus. Desiraju and Sappington show that if the agents are ex ante identical, the principal can implement the standard outcomes while eliminating all ex post inequality. In one such mechanism, a low-type agent gets paid more if the other agent is a high type than if the other agent is also a low type. If the agents are ex ante different, however, there is no way to implement the standard outcomes while eliminating inequality for all type realizations.

[29] There is a also a mini-literature on implementation when agents have a preference for reporting their types honestly, showing that even a weak such preference can greatly help the principal (Alger and Renault 2006;

5.4 *Potential Future Directions*

Several topics in mechanism design are natural candidates to link to behavioral ideas. A notable aspect of auctions is that they seem to trigger some unique emotions related to competition that are not triggered in many other settings. In particular, as noted by Malmendier and Lee (2011), it has been understood since ancient times that auctions invoke "bidding fever," while Morgan, Steiglitz, and Reis (2003) and Cox, Smith, and Walker (1988) argue that auctions engage a spite motive and the joy of winning, respectively. An interesting issue is how these emotions work and how they affect optimal auction design. For instance, it is not immediately clear why these emotions are triggered by auctions, while other situations (including many in this review) are more likely to generate positive social preferences.

While classical mechanism design recognizes that public revelation of information can help the principal—e.g., in an auction if bidders' valuations are positively correlated (Milgrom and Weber 1982)—behavioral economics introduces at least two novel reasons for why information could matter. First, as in the case of regret above and in the case of anticipatory utility discussed in other sections of the review, information can directly affect agents' preferences. Second, information can lead strategically naïve agents to change their views about the strategic situation, either directly through providing information on how others play or indirectly through focusing their attention on specific aspects of the game. For instance, in a second-price auction a bidder must rely exclusively on her predictions regarding the distribution of bids, but in an English auction she receives some information on this distribution. Such informational issues seem understudied in auction theory and mechanism design.

6. *Exploitative Contracting*

This section considers contracts designed exclusively or primarily to exploit false beliefs by the agent. While there does not seem to be a precise formal way to distinguish such contracts from nonexploitative contracts that are affected by agent mistakes, an informal and subjective distinction seems useful to make. Namely, a contract is exploitative if the economically central considerations driving it derive from trying to profit from the agent's mistake, and other considerations or constraints are nonexistent, not binding, or not central.

The literature has explored two broad forms of false beliefs. First, an agent's false beliefs may be about the contract itself; in Gabaix and Laibson (2006), for instance, naïve consumers underappreciate that the product they are purchasing will require some add-on purchases. Alternatively, an agent might believe that regulation prevents certain charges, when in fact it does not (Armstrong and Vickers, 2012). Second, an agent's false beliefs may be about her own behavior given the contract; in DellaVigna and Malmendier (2004), for instance, naïve hyperbolic discounters are aware of all contract features, but they mispredict how they will behave given a contract.

Methodological issues. The models discussed in this section assume that the principal knows the agent's tendency to commit mistakes, so that in this domain, the principal is more informed about the agent than she is herself. While unusual from a classical point of view—where it is commonly assumed that the agent's tendencies are

Matsushima 2008). These papers, however, make specific assumptions about the structure of honesty preferences that do not seem to be based on psychology evidence, so I do not review them here.

her private information—this assumption often makes sense when the principal is a firm with the capacity and willingness to collect and analyze tremendous amounts of data about consumers, and the agent is an individual consumer. The theories are also unlike informed-principal models with novel types of principal information because the agent does not make inferences about her tendencies from the contract she is offered. Hence, as discussed by Eliaz and Spiegler (2006), these models can be treated formally as contracting with noncommon priors. While the classical literature also recognizes the potential relevance of noncommon priors for contracting and other settings, the recent literature has both reinvigorated research on such models and changed its style and emphasis: the theories posit a particular form of agent beliefs based on psychology evidence, take a stance on which beliefs are correct, and tend to consider welfare issues.

6.1 *Nature and Implications of Exploitative Contracts*

Seemingly cheap products. Exploitative contracts often make products appear cheaper than they really are. The reason is simple: firms want to encourage consumers to buy, so if they choose multiple prices (e.g., a basic price and an add-on price) that consumers will pay, they aim to obtain revenues more from the prices consumers underappreciate. In a competitive market, this does not necessarily affect firm profits or consumer or social welfare: similar to the logic of loss-leader pricing (Lal and Matutes 1994) as well as that of many switching-cost models (e.g., Farrell and Klemperer 2007), firms compete aggressively for valuable consumers ex ante and bid down the more noticeable component of the price until they eliminate net profits. In this sense, naïve consumers can be protected from their mistakes by market forces.

Nevertheless, the fact that firms distort prices to take advantage of consumer mistakes can have a number of efficiency implications, so that the above "safety in markets" is very partial. Consumers might buy products whose value is below the true price but above the misperceived low price, generating inefficiency in participation decisions (Heidhues and Kőszegi 2013). Other inefficiencies obtain due to the high prices naïve consumers do not appreciate. In Gabaix and Laibson (2006) and Armstrong and Vickers (2012), naïve consumers ignore add-on prices, firms respond by choosing high add-on prices, and this leads sophisticated consumers to take socially inefficient steps (such as arranging for a substitute with higher production costs) to avoid the add-on. In Grubb (2009), consumers believe they can predict their consumption more precisely than they actually can, firms respond by setting high marginal prices for high usage, and this can lead consumers to underutilize the service ex post. In Gottlieb and Smetters (2012), life-insurance buyers do not properly account for the probability that they will lapse their policies, firms respond by front-loading premiums to make the policy look better, and this distorts consumption smoothing. At the same time, distorted prices could also have beneficial effects if they correct another problem. For instance, firms take advantage of naïve present-biased agents by offering high per-usage fees for pleasurable goods whose consumption the agent underestimates, and these high prices have the beneficial effect of lowering the present-biased agent's consumption (DellaVigna and Malmendier, 2004). Further, as argued by Bar-Gill (2006), if a consumer's mistake leads her to consume too little, the higher consumption induced by deceptively low prices can benefit her.

Cross-subsidies. Another recurring feature of exploitative contracts, first discussed in detail by Gabaix and Laibson (2006), is

that sophisticated consumers benefit from the presence of naïve consumers. Because naïve consumers do not anticipate some fees that they will pay, they tend to generate higher profits than sophisticated consumers buying the same contract. In a competitive market—where firms make zero profits on average—it must therefore be that firms make money on naïve consumers and lose money on sophisticated consumers, in effect acting as a tool for cross-subsidizing the latter type.

Exacerbation of mistakes using high fees. DellaVigna and Malmendier (2004) were the first to point out that firms might fine-tune contracts to exacerbate consumers' mistakes. A number of papers explore the specific features of such exploitative contracts.

Eliaz and Spiegler (2008) study optimal dynamic contracting when an agent's preferences are uncertain at the time of signing the contract, and the agent may be more optimistic than the principal about the better state occurring. As an example, suppose that all law-school students will end up in the nonprofit sector or in the corporate sector with the same fixed probabilities, but students have different, to the school unobservable, overoptimistic beliefs that they will end up in the nonprofit sector. The school's optimal screening menu includes a standard nonexploitative loan contract that is not contingent on subsequent outcomes, as well as exploitative loan contracts that amount to speculation on where a student will end up. With the latter type of loan, the student pays somewhat less if she ends up in the nonprofit sector and much more if she ends up in the corporate sector. An interesting feature of the optimal menu is that students with beliefs close to those of the school receive a nonexploitative contract. Intuitively, although the school could make a little money on slightly overoptimistic students by taking a small bet on what the student will do, much more overoptimistic students would value such a contract much more highly, forcing the school to pay a kind of information rent. This implies that contracts should either involve no speculation or substantial levels of speculation.

Eliaz and Spiegler (2006) analyze a closely related problem in which the principal knows that the agent's preferences will change between the time of signing the contract and the time of taking an action, whereas the agent may assign positive probability to her preferences remaining the same. The time inconsistency in the agent's preferences introduces an additional value from contracting similar to that in section 4.1: the agent prefers the contract to restrain her future behavior. Again, if the agent's beliefs are sufficiently close to that of the principal, she receives the ex ante optimal contract that fully commits her behavior, but if she is sufficiently naïve regarding her time inconsistency, she receives an exploitative contract in which she is effectively rewarded if she does not change her mind and punished if she does change her mind.

Heidhues and Kőszegi (2010) study exploitative credit contracts, and analyze the kinds of mistakes that are exacerbated by firms and lead to large welfare losses. In their model, present-biased consumers who might be partially naïve about their time inconsistency can sign exclusive contracts with competitive suppliers of credit, agreeing to a menu of installment plans according to which credit can be repaid in the future. In the competitive equilibrium firms offer seemingly cheap credit to be repaid quickly, but introduce large penalties for falling behind this front-loaded repayment schedule. The contracts are designed so that borrowers with even an arbitrarily small degree of naïveté both pay the penalties and repay in an ex ante suboptimal back-loaded manner more often than they predict. Intuitively, a lender chooses the repayment options so that, when deciding how to repay in the future, the consumer will

be indifferent between the front-loaded and back-loaded schedules. As a result, if she is naïve about her time inconsistency—no matter by how little—she falsely believes that she will repay quickly. To make matters worse, the same misprediction leads the consumer to underestimate the cost of credit and borrow too much—despite borrowing being for future consumption.

6.2 *Helping Consumers*

Many researchers have asked whether there are ways of improving consumer welfare in the types of situations described in the previous subsection. The theme that emerges from this literature so far is that both market-based and regulatory solutions face severe problems, but some forms of intervention may nevertheless increase welfare. As has been pointed out by Armstrong and Vickers (2012), for instance, one common feature of welfare-increasing interventions is redistribution of wealth from sophisticated to naïve consumers through the reduction of cross-subsidies. While some researchers and observers use this as an argument against intervention on the basis that we should not hurt rational agents who are paying attention and making individually optimal decisions, from an economic perspective this seems no better a reason to refrain from intervention than the fact that pollution taxes redistribute wealth from potential polluters to the rest of society. In fact, to the extent that sophisticated consumers are wealthier than naïve consumers, redistribution might be an additional argument for intervention.

Lack of market incentives to educate consumers. The fact that consumers underestimate some costs associated with product use raises the question of whether market participants have incentives to educate consumers or offer more transparent products or contracts. First-pass logic suggests that in a sufficiently competitive industry, competitors should unshroud price components underappreciated by consumers, and then compete on them. While it is unclear whether and how one can educate consumers, and this question requires further research, a number of papers have investigated the incentive to unshroud assuming that firms can costlessly do so. Gabaix and Laibson (2006), Spiegler (2006), and Heidhues, Kőszegi, and Murooka (2012) show that unshrouding is often unprofitable because it turns profitable naïve consumers into unprofitable sophisticated consumers. Hence, deceptive products or contracts can often survive in markets. Furthermore, Heidhues, Kőszegi, and Murooka (2012) identify a perverse aspect of when this can happen: products that generate lower social surplus than the best alternative facilitate deception precisely because they would not survive in the market if consumers understood hidden fees, and therefore firms often make profits on exactly such products.

Problems with financial advice. Armstrong and Zhou (2011), Stoughton, Wu, and Zechner (2011) and Inderst and Ottaviani (2012) show that if consumers are naïve, firms pay commissions to financial advisors to steer consumers to buy their high-priced products, and financial advisors do so even if this is suboptimal for a consumer. Murooka (2013) establishes that even perfect competition among financial advisors does not push commissions down to a competitive level, as deceptive firms must pay significantly higher commissions to stop advisors from explaining to consumers that other products are superior. Since the high commissions are ultimately paid by consumers, in this situation the presence of financial advisors who can explain deceptive practices *lowers* consumer welfare.

Interventions to improve consumer decision making. A number of researchers recognize that when consumers might mispredict their own behavior, classical disclosure, which

typically limits information to those related to the contract itself, is insufficient, and informing consumers about themselves is necessary (for example Bar-Gill and Ferrari 2010). An approach by Thaler and Sunstein (2008) and Kamenica, Mullainathan, and Thaler (2011) would require firms to provide contract details and individualized usage information to all consumers in a standard form, so that market-based "choice engines" can emerge that help consumers choose, making consumer *choices* but not necessarily making consumers more sophisticated.

Welfare-improving interventions that limit contract forms. A number of authors have shown that limiting what firms can do—thereby targeting the tools firms use to exploit consumers—can raise social welfare. For instance, Heidhues and Kőszegi (2010) show that because the large welfare losses partially naïve consumers suffer in credit markets are driven by large fees, prohibiting large fees for small deviations from contract terms often raises welfare for *any* combination of naïve and sophisticated consumers in the population. Similarly, Inderst and Ottaviani (2012) and Murooka (2013) show that because the distorted advice financial advisors provide is driven by the large commissions firms pay, capping or banning commissions, or requiring commissions to be uniform across products, can raise consumer welfare.[30]

[30] Korobkin (2003) provides an interesting legal perspective on exploitative consumer contracts. Similarly to the economic research in this section, he argues that when consumers do not understand all aspects of the contract, sellers have an incentive to include inefficient terms that favor themselves. As a remedy, he proposes a modification of the unconscionability doctrine, a legal doctrine that renders contract features a party had no effective choice over invalid. Although this is not the prevailing legal interpretation, Korobkin argues that a consumer who did not understand part of the contract could not have had effective choice over it. And in a completely different approach, Bubb and Kaufman (2011) propose that different ownership structures, such as customer-owned firms, can lower firms' incentive to exploit consumer mistakes.

Problems with government intervention. A number of authors point out potential problems with market intervention. Grubb (2012) considers services, such as mobile phones and bank overdraft protection, for which consumers may not know the marginal price of the next unit of service, and asks whether requiring firms to disclose this information at the point of sale increases welfare. If consumers correctly anticipate their probability of running into penalties, such regulation can actually hurt because it interferes with efficient screening by firms. Intuitively, penalty fees for high usage prevent high-value consumers from taking the contracts offered to low-value consumers; yet because consumers do not know when they apply, these fees do not distort the consumption of low-value consumers.

Kosfeld and Schüwer (2011) establish that, since the inefficiency in Gabaix and Laibson's 2006 model of shrouded attributes is due to sophisticated consumers' efforts to avoid the add-on, in such a setting, education—modeled as turning a portion of naïve consumers into sophisticated consumers—can lower social welfare. And Warren and Wood (2010) show that when naïve consumers misperceive add-on prices, there is no budget-balanced regulation that consumers will perceive as improving their welfare, so that there is no welfare-improving regulation that citizens will vote for. Intuitively, a competitive market redistributes income from naive to sophisticated consumers, and since no consumer believes that she is naïve, she believes she benefits from the redistribution.

6.3 *Potential Future Directions*

While researchers have identified a number of features of exploitative contracts, overall this literature seems to be in its infancy, with a variety of open questions. I highlight some major issues. First, while many researchers employ a static framework or assume that firms and consumers sign

exclusive contracts, it seems important to understand the effect of ex post competition and the possibility of switching for contract structure. Gottlieb (2008) takes a first step in this research agenda, exploring the dynamics of contracting with present-biased agents in the setting of DellaVigna and Malmendier (2004). He shows that ex post competition leaves contracts for products with immediate costs and future benefits unaffected, but renders contracts for products with immediate benefits and future costs equivalent to spot markets. As an example in the context of smoking, while a retailer and a present-biased consumer might prefer to sign an exclusive contract in which (to restrain her future smoking) the consumer is paid a lump sum ex ante and buys overpriced cigarettes from the retailer ex post, this contract is ineffective if the consumer can go to other retailers. Nevertheless, in some markets for products with immediate benefits, such as credit cards, consumers do not switch easily to competitors, and firms seem to take advantage of this (Ausubel 1991). Procrastination on switching offers a plausible explanation, but this explanation and especially its implications for equilibrium contracts have not been formally analyzed.

A noticeable aspect of existing exploitative contracts seems to be that exploitative features are often masked as having alternative uses. For instance, while the primary purpose of a high credit-card late fee is likely to exploit a consumer's inattention to this fee or her optimism that she will not pay late, one can make the argument that the purpose is to overcome the moral-hazard problem that she might delay repayment. An interesting question for behavioral contract theory is why exploitative features that can in no plausible environment be useful are so rare. Some form of metasophistication on the part of consumers—whereby consumers might realize the possibility of exploitation, and pure exploitative features might seem suspicious to them—could be involved. Alternatively, this pattern may be due to regulation or the threat of regulation, making it more difficult to conclusively establish that the feature in question is deceptive.

Another poorly understood issue is how to improve consumer decision making, including how to educate or inform consumers about their own behavior. The likely impact of proposed choice engines to improve consumer decision making, and their optimal regulation, also deserves closer theoretical scrutiny. The concern arises that the platform for exploitation shifts from firms to choice engines—which become like other intermediaries who deceive consumers—and it is unclear what business model for choice engines will prevent this. In particular, if choice engines are paid for finding a match between buyers and sellers, then to attract consumers, they appear to have similar incentives to mislead consumers as firms do. Furthermore, because choice engines can analyze offers much more quickly, their presence can lead to a proliferation of products, including products that take advantage of very specific or rare mistakes by consumers.

Finally, researchers have studied contracting to exploit agents who are naïve about their suboptimal behavior much more than they have studied contracting to help agents who are sophisticated about their suboptimal behavior (e.g., commitment contracts)—although such sophisticated agents should in principle demand help. Indeed, this asymmetry in the literature seems to reflect an asymmetry in the real-life contracts we observe. Whether there are more exploitative than "helping" contracts, and, if so, why, is an important area for future research. There are at least two main possibilities. First, the market mechanism might, for some reason, favor the exploitation of biases rather than their mitigation. Second, it might be the case that most individuals are not sufficiently sophisticated to demand help, or are

averse to helping contracts for another reason, such as being averse to committing their own behavior.

7. *Other Topics*

This section covers two further small literatures in behavioral contract theory.

7.1 *Incomplete Contracts*

A contract is incomplete if it leaves the details of some transaction that the parties care about to be determined at a later time. In the classical literature on incomplete contracts, such as Grossman and Hart (1986) and Hart and Moore (1990), parties make interim investments into the relationship, and because the incomplete contract cannot provide direct incentives for these investments, they are typically inefficient. The main contracting problem is to maximize the efficiency of the interim decisions, primarily through the allocation of ownership and control rights.

The psychology-and-economics literature on incomplete contracts is relatively small. Papers in this literature study ways in which individuals' reactions to a contract generate inefficiencies ex post (in contrast to the classical literature, where inefficiencies arise at the investment stage) and how this affects optimal contracts and the optimal allocation of ownership rights.

In the most influential contribution of this literature, Hart and Moore (2008) analyze a buyer–seller relationship when the contract parties write initially cannot condition on the state of the world that determines the optimal terms of trade, and at the ex post stage an aggrieved party can inefficiently "shade" on her performance in a way that is impossible to prevent contractually. As a simple example, while a court may be able to verify that a professor delivered an agreed-to lecture, it cannot verify that the lecture was good, and hence a professor can shade by delivering a bad lecture. Hart and Moore assume that a party is aggrieved and shades if she gets a worse outcome than the best possible under the initial contract.[31] The way to avoid aggrieving a party and shading, therefore, is to fix or restrict the terms in the ex ante contract; for instance, the parties may agree to trading at a fixed price. This will result in a failure to trade in circumstances when the price does not fall between the buyer's value and the seller's cost, but economizes on shading costs, and the optimal contract trades off these two considerations. Furthermore, among multiple terms the parties eventually decide on, it is most important to fix the ones that create the most conflict of interest—and therefore the most potential for shading. Indeed, since the parties have a direct conflict regarding the price, an optimal contract might fix the price but not the full description of the good to be delivered. Finally, to minimize shading, the party who cares more about the good should have the right to specify it ex post. In the extreme, if the seller is almost indifferent between different products but the buyer cares a lot—such as in a firm where the employee might perform a number of similarly difficult tasks, only a few of which are useful—it is optimal for the buyer to have control rights. If the opposite is the case—such as when the seller is subject to large cost shocks—the seller should be in control.[32]

Hart (2009) applies a similar model to a holdup problem, where holding up the other party results in inefficient shading in the

[31] Psychologically, this means that a party is not aggrieved if the terms of the initial contract are bad, but is aggrieved if the terms of the renegotiated transaction are bad (in her view). Hart and Moore (2008) explain this distinction by arguing that there is a change in circumstances: the final agreement often occurs in situations of bilateral monopoly, whereas the initial contract is agreed to in a more competitive, "objective" setting.

[32] See also Hart and Holmstrom (2010) for an application of this approach to firm scope—whether and when units should integrate to a single firm or merely cooperate.

renegotiated contract. A buyer and a seller agree to a fixed price ex ante and, when the state of the world is realized, decide whether to hold up the other party by refusing to trade. Because holdup results in shading, a party only chooses it if the resulting price is sufficiently better than the price agreed to in the contract. Hart shows that an appropriate allocation of asset ownership can lower the probability of inefficient holdup. In particular, it is optimal to assign asset ownership to the party who faces more uncertainty regarding her value from trade; since this increases the party's outside option exactly when her value from trade is high, it makes holding her up less profitable. This implies that—unlike in the classical model, where asset ownership is determined by the importance of interim investments—asset ownership is determined by the importance of uncertainty in a party's value from trade.

Herweg and Schmidt (2012) take a somewhat different approach to how contracts affect preferences over trades. They suppose that a loss-averse buyer and a loss-averse seller write a contract specifying the nature and price of the good to be delivered at a later date, and the contract serves as a reference point for outcomes ex post. Since a party evaluates a better-than-contracted term (e.g., an increase in price for the seller) as a gain and a worse-than-contracted term as a loss, and loss aversion implies that she is more sensitive to the loss, parties are reluctant to compromise and trade away from the contract. In particular, if the ex post optimal terms are close to the specified terms, the parties stick to the contract, and even if the ex post optimal terms are far, the parties adjust the terms only partially, in either case generating (material) inefficiency. Here, inefficiency arises when there is a specific contract, whereas in Hart and Moore (2008), inefficiency arises when there is only a vague contract. An immediate implication is that the parties might prefer not to write a contract so as not to set a reference point, or—to optimize the insufficient adjustment ex post—might prefer to write a kind of "compromise" contract that they know they will trade away from. The paper also derives two less immediate implications. First, it is optimal to write a specific contract rather than rely on ownership rights to incentivize relationship-specific investments if there is little uncertainty or parties are not very loss averse. Second, an employment contract dominates a fixed performance contract if the scope for inefficient abuse is small relative to the renegotiation costs generated by loss aversion due to uncertainty.

7.2 *Environment Design*

A few papers investigate, under various psychologically motivated assumptions, how to design individuals' decision-making environment to achieve specific goals.

Optimal nudges. Psychology and economics has had a major practical impact in motivating "soft paternalism"—designing policies such that individuals with a tendency to behave suboptimally make better decisions, but others are either free to choose as they would otherwise (Thaler and Sunstein 2003; Thaler and Sunstein 2008), or are not hurt much (Camerer et al., 2003). Some work can be thought of as providing theoretical guidance as to how such "nudges" should be designed.

O'Donoghue and Rabin (1999b) analyze the problem of a principal employing a naïve present-biased agent to complete a single task with an uncertain cost of effort, where he can pay the agent depending on when she completes the task. The principal faces a delay cost, yet because in any given period the effort cost might be high, efficiency requires that the agent sometimes wait. Consistent with the soft paternalist agenda but quite distinct from the literature on exploitative contracts, O'Donoghue

and Rabin assume that the principal's goal is to choose the most efficient (rather than profit-maximizing) incentive scheme. Note that with time-consistent preferences, such an incentive scheme is obvious: the principal imposes a punishment for delay that is exactly equal to his delay cost, generating a first-best outcome. When the agent is present-biased and the principal does not know the task-cost distribution, in contrast, the first-best outcome is not achievable. Intuitively, since for high average costs the agent is more prone to procrastinate, the principal needs severe punishment for delay to induce her to complete the task efficiently; but if he imposes such a punishment, an agent with low average costs will tend to complete the task too soon. The optimal contract, then, resembles a deadline: punishment for delay is relatively mild for a while to allow an agent with low average cost to delay if necessary, but severe after a deadline to discourage an agent with high average costs from procrastinating.

Carroll et al. (2009) study optimal defaults for retirement savings decisions when employees have time-inconsistent tastes for immediate gratification, and have heterogeneous optimal savings rates. Employees are initially assigned the default savings rate, and in each period draw a stochastic cost at which they can change their savings rate to the optimal one, suffering a flow utility loss if they do not. Employees' problem is that (due to their time inconsistency) they may procrastinate paying the cost. Carroll et al. establish that if employees have a large time-inconsistency problem and are not too heterogeneous, it is optimal to set the default at the optimal rate for the population distribution. If employees are very heterogenous or the time inconsistency is not so serious, in contrast, the optimal policy involves "active decisions": the default is set at such an unattractive level that all employees are forced to pay the cost immediately. And in in-between cases, a compromise policy is optimal: the default is set such that it is good enough for part of the population, while the rest of the population is forced to pay the cost and switch to the optimal savings rate.

Information optimization with emotions. Caplin and Eliaz (2003) and Schweizer and Szech (2013) characterize optimal medical tests when individuals derive anticipatory utility from their beliefs about their health status, and find bad news more aversive than good news pleasant. In both papers, the principal can commit to an information-revelation policy that is conditional on an individual's status. Caplin and Eliaz establish that an optimal policy for stopping the spread of HIV gives only noisy bad news: it certifies some, but not all, individuals who are uninfected, and gives the same signal to everyone else. This policy protects individuals from receiving very bad news, yet provides sufficient information to help match uninfected individuals. Using methods similar to those in Kamenica and Gentzkow (2011), Schweizer and Szech (2013) show that such a policy also emerges as the optimal way to reveal information about infection status to an individual who uses the information to make better decisions.

7.3 *Potential Future Directions*

Most of the literature on incomplete contracts is based on theories of how a contract agreed to by the parties affects future preferences over outcomes. While this approach has produced important insights, it seems both too broad and too narrow for incomplete contracts. It is too broad because the question of how contracts affect preferences is an important topic for research quite independently of whether the contract is incomplete. Identifying the psychological underpinnings of how contracts affect preferences—including the extent to which these can be derived from psychological phenomena, such as the role of expectations or social preferences, that

go beyond contracts, and the extent to which they are specific to contracts—seems to be an important agenda for future research. At the same time, the question of how contracts affect preferences is also too narrow for incomplete contracts because psychological phenomena that are not directly about this question are also likely to have implications for incomplete contracts.

A long-recognized issue at the foundations of incomplete contracts is whether and when it is reasonable to assume that contracts will be incomplete. Since rational agents can often write optimal contracts that render standard arguments for incomplete contracts irrelevant (for example Moore 1992; Maskin and Tirole 1999), some researchers have conjectured that ideas from psychology and economics might help provide more compelling foundations for the incompleteness of contracts. This is still an important problem for future research.

Relative to the attention that has been paid to the question of soft paternalism at least since the success of Thaler and Benartzi's (2004) Save More Tomorrow plan, the dearth of theoretical research on such policies is striking. One potential reason is that most real-life nudges take advantage of the default effect—that individuals tend to stick with an option they have chosen or to which they have been automatically assigned—and we do not have good theories for why defaults influence behavior in such a powerful way.[33] Yet providing much better theoretical guidance on soft paternalism seems essential for the continued practical success of the agenda, especially in figuring out what kinds of policies are likely to work and which policies push people in the right direction. Theoretical work is also necessary for delineating the limits of the nudge agenda—that is, when more heavy-handed policies are desirable.

In research on emotions and information in contract theory, authors have largely confined themselves to Bayesian models, although anticipatory utility might lead to non-Bayesian information processing, and beliefs formed in a non-Bayesian way can change emotions. Oster, Shoulson, and Dorsey (2013), for instance, argue that facts regarding testing and economic behavior among patients at risk for Huntington's disease are only consistent with non-Bayesian models of anticipatory utility, such as Brunnermeier and Parker (2005). Thinking about how to optimize information revelation with non-Bayesian beliefs seems like an interesting agenda.

8. *Conclusion*

This section discusses a few general issues regarding the state and direction of the literature on behavioral contract theory, and applied behavioral-economics theory more generally. When applying a behavioral-economics theory to an economic setting, the all-too common tendency of much research is to study how the parameters of the behavioral theory affect predictions. In studying the implications of present bias for credit, for instance, a researcher might analyze how less and more present-biased individuals behave. While such analyses are useful for understanding the model and comparing it to a classical one, ultimately economists are more interested in comparative statics with respect to variables typically studied in economic analysis. To continue with the previous example, a researcher could study how the same present-biased individual responds to different kinds of credit or to changes in the interest rate. Such predictions are both economically more relevant and easier to test than those deriving from manipulating the behavioral model itself.

[33] Most likely, there are multiple psychological reasons for people not to switch away from defaults, and the effect is robust and powerful because in most situations multiple forces operate at the same time.

A central constraint on progress in behavioral economics is the scarcity of compelling portable models of individual decision making that researchers can apply to economic settings. Much of behavioral contract theory uses only the four models introduced in section 2, highlighting both that useful models can be taken up by many researchers, and that there are too few portable models. As one particular example, developing portable psychologically based models of limited or distorted attention is a first-order question.

References

Akerlof, George A. 1982. "Labor Contracts as Partial Gift Exchange." *Quarterly Journal of Economics* 97 (4): 543–69.

Akerlof, George A., and Janet L. Yellen. 1988. "Fairness and Unemployment." *American Economic Review* 78 (2): 44–49.

Akerlof, George A., and Janet L. Yellen. 1990. "The Fair Wage-Effort Hypothesis and Unemployment." *Quarterly Journal of Economics* 105 (2): 255–83.

Alger, Ingela, and Régis Renault. 2006. "Screening Ethics When Honest Agents Care about Fairness." *International Economic Review* 47 (1): 59–85.

Amador, Manuel, Iván Werning, and George-Marios Angeletos. 2006. "Commitment vs. Flexibility." *Econometrica* 74 (2): 365–96.

Armstrong, Mark, and John Vickers. 2012. "Consumer Protection and Contingent Charges." *Journal of Economic Literature* 50 (2): 477–93.

Armstrong, Mark, and Jidong Zhou. 2011. "Paying for Prominence." *Economic Journal* 121 (556): F368–95.

Augenblick, Ned. 2011. "Consumer and Producer Behavior in the Market for Penny Auctions: A Theoretical and Empirical Analysis." Unpublished.

Ausubel, Lawrence M. 1991. "The Failure of Competition in the Credit Card Market." *American Economic Review* 81 (1): 50–81.

Bar-Gill, Oren. 2006. "Bundling and Consumer Misperception." *University of Chicago Law Review* 73 (1): 33–61.

Bar-Gill, Oren, and Franco Ferrari. 2010. "Informing Consumers about Themselves." *Erasmus Law Review* 3 (2): 93–119.

Bartling, Björn. 2011. "Relative Performance or Team Evaluation? Optimal Contracts for Other-Regarding Agents." *Journal of Economic Behavior and Organization* 79 (3): 183–93.

Bartling, Björn, and Ferdinand A. von Siemens. 2010. "The Intensity of Incentives in Firms and Markets: Moral Hazard with Envious Agents." *Labour Economics* 17 (3): 598–607.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Bierbrauer, Felix, and Nick Netzer. 2012. "Mechanism Design and Intentions." University of Cologne Department of Economics Working Paper 53.

Bolton, Patrick, and Mathias Dewatripont. 2005. *Contract Theory*. Cambridge, Mass. and London: MIT Press.

Bolton, Patrick, and Antoine Faure-Grimaud. 2010. "Satisficing Contracts." *Review of Economic Studies* 77 (3): 937–71.

Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95 (4): 1092–1118.

Bubb, Ryan, and Alex Kaufman. 2011. "Consumer Biases and Mutual Ownership." New York University School of Law Law and Economics Research Paper 11-35.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *Quarterly Journal of Economics* 119 (3): 861–98.

Camerer, Colin F., Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin. 2003. "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism'." *University of Pennsylvania Law Review* 151 (3): 1211–54.

Caplin, Andrew, and Kfir Eliaz. 2003. "AIDS Policy and Psychology: A Mechanism-Design Approach." *RAND Journal of Economics* 34 (4): 631–46.

Carbajal, Juan Carlos, and Jeffrey C. Ely. 2012. "Optimal Contracts for Loss Averse Consumers." Unpublished.

Carmichael, Lorne, and W. Bentley MacLeod. 2003. "Caring About Sunk Costs: A Behavioral Solution to Holdup Problems with Small Stakes." *Journal of Law, Economics and Organization* 19 (1): 106–18.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124 (4): 1639–74.

Chiappori, Pierre-Andre, and Bernard Salanie. 2000. "Testing for Asymmetric Information in Insurance Markets." *Journal of Political Economy* 108 (1): 56–78.

Cox, James C., Vernon L. Smith, and James M. Walker. 1988. "Theory and Individual Behavior of First-Price Auctions." *Journal of Risk and Uncertainty* 1 (1): 61–99.

Cramton, Peter, Emel Filiz-Ozbay, Erkut Y. Ozbay, and Pacharasut Sujarittanonta. 2012. "Fear of Losing in a Clock Auction." *Review of Economic Design* 16 (2–3): 119–34.

Crawford, Vincent P., and Nagore Iriberri. 2007. "Level-*k* Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica* 75 (6): 1721–70.

Crawford, Vincent P., Tamar Kugler, Zvika Neeman, and Ady Pauzner. 2009. "Behaviorally Optimal Auction Design: Examples and Observations." *Journal of the European Economic Association* 7 (2–3): 377–87.

Daido, Kohei, and Takeshi Murooka. 2012. "Team Incentives and Reference-Dependent Preferences." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1922366.

Deci, Edward L. 1975. *Intrinsic Motivation*. New York: Plenum Publishing Company.

Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. "A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125 (6): 627–68.

de la Rosa, Leonidas Enrique. 2011. "Overconfidence and Moral Hazard." *Games and Economic Behavior* 73 (2): 429–51.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47 (2): 315–72.

DellaVigna, Stefano, and Ulrike Malmendier. 2004. "Contract Design and Self-Control: Theory and Evidence." *Quarterly Journal of Economics* 119 (2): 353–402.

de Meza, David, and David C. Webb. 2007. "Incentive Design under Loss Aversion." *Journal of the European Economic Association* 5 (1): 66–92.

Desiraju, Ramarao, and David E. M. Sappington. 2007. "Equity and Adverse Selection." *Journal of Economics and Management Strategy* 16 (2): 285–318.

Eisenhuth, Roland. 2010. "Auction Design with Loss Averse Bidders: The Optimality of All Pay Mechanisms." Munich Personal RePEc Archive Paper 23357.

Eisenhuth, Roland. 2013. "Reference Dependent Mechanism Design." Unpublished.

Eisenhuth, Roland, and Mara Ewers. 2012. "Auctions with Loss Averse Bidders." Unpublished.

Eliaz, Kfir, and Ran Spiegler. 2006. "Contracting with Diversely Naïve Agents." *Review of Economic Studies* 73 (3): 689–714.

Eliaz, Kfir, and Ran Spiegler. 2008. "Consumer Optimism and Price Discrimination." *Theoretical Economics* 3 (4): 459–97.

Eliaz, Kfir, and Ran Spiegler. 2014. "Reference-Dependence and Labor-Market Fluctuations." In *NBER Macroeconomics Annual 2013*, edited by Jonathan Parker and Michael Woodford, 159–200. Chicago and London: University of Chicago Press.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98 (3): 990–1008.

Engelbrecht-Wiggans, Richard. 1989. "The Effect of Regret on Optimal Bidding in Auctions." *Management Science* 35 (6): 685–92.

Englmaier, Florian, and Stephen Leider. 2012. "Contractual and Organizational Structure with Reciprocal Agents." *American Economic Journal: Microeconomics* 4 (2): 146–83.

Englmaier, Florian, and Achim Wambach. 2010. "Optimal Incentive Contracts under Inequity Aversion." *Games and Economic Behavior* 69 (2): 312–28.

Esteban, Susanna, and Eiichi Miyagawa. 2006. "Temptation, Self-Control, and Competitive Nonlinear Pricing." *Economics Letters* 90 (3): 348–55.

Esteban, Susanna, Eiichi Miyagawa, and Matthew Shum. 2007. "Nonlinear Pricing with Self-Control Preferences." *Journal of Economic Theory* 135 (1): 306–38.

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica* 73 (5): 1623–72.

Falk, Armin, and Michael Kosfeld. 2006. "The Hidden Costs of Control." *American Economic Review* 96 (5): 1611–30.

Fang, Hanming, and Giuseppe Moscarini. 2005. "Morale Hazard." *Journal of Monetary Economics* 52 (4): 749–77.

Farrell, Joseph, and Paul Klemperer. 2007. "Coordination and Lock-In: Competition with Switching Costs and Network Effects." In *Handbook of Industrial Organization, Volume 3*, edited by Mark Armstrong and Robert H. Porter, 1967–2072. Amsterdam: Elsevier, North-Holland.

Fehr, Ernst, Lorenz Goette, and Christian Zehnder. 2009. "A Behavioral Account of the Labor Market: The Role of Fairness Concerns." *Annual Review of Economics* 1: 355–84.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract Design." *Econometrica* 75 (1): 121–54.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114 (3): 817–68.

Fehr, Ernst, and Klaus M. Schmidt. 2004. "Fairness and Incentives in a Multi-task Principal–Agent Model." *Scandinavian Journal of Economics* 106 (3): 453–74.

Filiz-Ozbay, Emel, and Erkut Y. Ozbay. 2007. "Auctions with Anticipated Regret: Theory and Experiment." *American Economic Review* 97 (4): 1407–18.

Finkelstein, Amy, and Kathleen McGarry. 2006. "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market." *American Economic Review* 96 (4): 938–58.

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121 (2): 505–40.

Galperti, Simone. 2012. "Commitment, Flexibility, and Optimal Screening of Time Inconsistency." Unpublished.Garicano, Luis, and Andrea Prat. 2013. "Organizational Economics with Cognitive Costs." In *Advances in Economics and Econometrics: Tenth World Congress, Volume I, Economic Theory*, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 342–88. Cambridge and New York: Cambridge University Press.

Gervais, Simon, J. B. Heaton, and Terrance Odean. 2011. "Overconfidence, Compensation Contracts, and Capital Budgeting." *Journal of Finance* 66 (5): 1735–77.

Gneezy, Uri, and Aldo Rustichini. 2000. "A Fine Is a

Price." *Journal of Legal Studies* 29 (1): 1–17.
Gottlieb, Daniel. 2008. "Competition over Time-Inconsistent Consumers." *Journal of Public Economic Theory* 10 (4): 673–84.
Gottlieb, Daniel, and Kent Smetters. 2012. "Narrow Framing and Life Insurance." National Bureau of Economic Research Working Paper 18601.
Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy* 94 (4): 691–719.
Grubb, Michael D. 2009. "Selling to Overconfident Consumers." *American Economic Review* 99 (5): 1770–1807.
Grubb, Michael D. 2012. "Consumer Inattention and Bill-Shock Regulation." Massachusetts Institute of Technology Sloan School Working Paper 4987-12.
Gul, Faruk, and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica* 69 (6): 1403–35.
Hahn, Jong-Hee, Jinwoo Kim, Sang-Hyun Kim, and Jihong Lee. 2010. "Screening Loss Averse Consumers." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1678825.
Harrison, J. Michael, and David M. Kreps. 1978. "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations." *Quarterly Journal of Economics* 92 (2): 323–36.
Hart, Oliver. 2009. "Hold-Up, Asset Ownership, and Reference Points." *Quarterly Journal of Economics* 124 (1): 267–300.
Hart, Oliver, and Bengt Holmstrom. 2010. "A Theory of Firm Scope." *Quarterly Journal of Economics* 125 (2): 483–513.
Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy* 98 (6): 1119–58.
Hart, Oliver, and John Moore. 2008. "Contracts as Reference Points." *Quarterly Journal of Economics* 123 (1): 1–48.
Heidhues, Paul, and Botond Kőszegi. 2008. "Competition and Price Variation When Consumers Are Loss Averse." *American Economic Review* 98 (4): 1245–68.
Heidhues, Paul, and Botond Kőszegi. 2010. "Exploiting Naivete about Self-Control in the Credit Market." *American Economic Review* 100 (5): 2279–303.
Heidhues, Paul, and Botond Kőszegi. 2013. "Seller Information about Consumer Naivete Lowers Welfare." Unpublished.
Heidhues, Paul, Botond Kőszegi, and Takeshi Murooka. 2012. "Inferior Products and Profitable Deception." Unpublished.
Herold, Florian. 2010. "Contractual Incompleteness as a Signal of Trust." *Games and Economic Behavior* 68 (1): 180–91.
Herweg, Fabian, and Klaus M. Schmidt. 2012. "A Theory of Ex Post Inefficient Renegotiation." Unpublished.
Herweg, Fabian, and Konrad Mierendorff. 2013. "Uncertain Demand, Consumer Loss Aversion, and Flat-Rate Tariffs." *Journal of the European Economic Association* 11 (2): 399–432.
Herweg, Fabian, Daniel Müller, and Philipp Weinschenk. 2010. "Binary Payment Schemes: Moral Hazard and Loss Aversion." *American Economic Review* 100 (5): 2451–77.
Holmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10 (1): 74–91.
Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization* 7 (Special Issue): 24–52.
Immordino, Giovanni, Anna Maria C. Menichini, and Maria Grazia Romano. 2011. "A Simple Impossibility Result in Behavioral Contract Theory." *Economics Letters* 113 (3): 307–09.
Inderst, Roman, and Marco Ottaviani. 2012. "Financial Advice." *Journal of Economic Literature* 50 (2): 494–512.
Itoh, Hideshi. 2004. "Moral Hazard and Other-Regarding Preferences." *Japanese Economic Review* 55 (1): 18–45.
Jehiel, Philippe. 2005. "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory* 123 (2): 81–104.
Jehiel, Philippe, and Laurent Lamy. 2012. "Absolute Auctions and Secret Reserve Prices: Why Are They Used?" Unpublished.
Jofre, Alejandro, Sofia Moroni, and Andrea Repetto. 2012. "Dynamic Contracts under Loss Aversion." Unpublished.
Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–92.
Kamenica, Emir. 2012. "Behavioral Economics and Psychology of Incentives." *Annual Review of Economics* 4: 427–52.
Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101 (6): 2590–615.
Kamenica, Emir, Sendhil Mullainathan, and Richard Thaler. 2011. "Helping Consumers Know Themselves." *American Economic Review* 101 (3): 417–22.
Korobkin, Russell. 2003. "Bounded Rationality, Standard Form Contracts, and Unconscionability." *University of Chicago Law Review* 70 (4): 1203–95.
Kosfeld, Michael, and Ulrich Schüwer. 2011. "Add-on Pricing, Naïve Consumers, and the Hidden Welfare Costs of Education." Centre for Economic Policy Research Discussion Paper 8636.
Kosfeld, Michael, and Ferdinand A. von Siemens. 2011. "Competition, Cooperation, and Corporate Culture." *RAND Journal of Economics* 42 (1): 23–43.
Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics* 121 (4): 1133–65.
Kőszegi, Botond, and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes." *American Economic Review* 97 (4): 1047–73.
Kőszegi, Botond, and Matthew Rabin. 2009. "Refer-

ence-Dependent Consumption Plans." *American Economic Review* 99 (3): 909–36.
Kucuksenel, Serkan. 2012. "Behavioral Mechanism Design." *Journal of Public Economic Theory* 14 (5): 767–89.
Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2): 443–78.
Lal, Rajiv, and Carmen Matutes. 1994. "Retail Pricing and Advertising Strategies." *Journal of Business* 67 (3): 345–70.
Lange, Andreas, and Anmol Ratan. 2010. "Multi-dimensional Reference-Dependent Preferences in Sealed-Bid Auctions—How (Most) Laboratory Experiments Differ From the Field." *Games and Economic Behavior* 68 (2): 634–45.
List, John A. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115 (3): 482–93.
Macera, Rosario. 2012. "Intertemporal Incentives under Loss Aversion." Unpublished.
Malmendier, Ulrike, and Young Han Lee. 2011. "The Bidder's Curse." *American Economic Review* 101 (2): 749–87.
Manove, Michael, and A. Jorge Padilla. 1999. "Banking (Conservatively) with Optimists." *RAND Journal of Economics* 30 (2): 324–50.
Maskin, Eric, and Jean Tirole. 1999. "Unforeseen Contingencies and Incomplete Contracts." *Review of Economic Studies* 66 (1): 83–114.
Matsushima, Hitoshi. 2008. "Role of Honesty in Full Implementation." *Journal of Economic Theory* 139 (1): 353–59.
Milgrom, Paul R., and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50 (5): 1089–1122.
Moore, John. 1992. "Implementation in Environments with Complete Information." In *Advances in Economic Theory*, edited by Jean-Jacques Laffont, 181–282. Cambridge and New York: Cambridge University Press.
Morgan, John, Ken Steiglitz, and George Reis. 2003. "The Spite Motive and Equilibrium Behavior in Auctions." *B. E. Journal of Economic Analysis and Policy* 2 (1): 1–27.
Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. "A Reduced-Form Approach to Behavioral Public Finance." *Annual Review of Economics* 4: 511–40.
Murooka, Takeshi. 2013. "Deception under Competitive Intermediation." Unpublished.
Netzer, Nick, and Armin Schmutzler. 2012. "Explaining Gift-Exchange—The Limits of Good Intentions." Unpublished.O'Donoghue, Ted, and Matthew Rabin. 1999a. "Doing It Now or Later." *American Economic Review* 89 (1): 103–24.
O'Donoghue, Ted, and Matthew Rabin. 1999b. "Incentives for Procrastinators." *Quarterly Journal of Economics* 114 (3): 769–816.
O'Donoghue, Ted, and Matthew Rabin. 2001. "Choice and Procrastination." *Quarterly Journal of Economics* 116 (1): 121–60.
O'Donoghue, Ted, and Matthew Rabin. 2003. "Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes." *American Economic Review* 93 (2): 186–91.
Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease." *American Economic Review* 103 (2): 804–30.
Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83 (5): 1281–1302.
Rabin, Matthew. 1998. "Psychology and Economics." *Journal of Economic Literature* 36 (1): 11–46.
Rey-Biel, Pedro. 2008. "Inequity Aversion and Team Incentives." *Scandinavian Journal of Economics* 110 (2): 297–320.
Roider, Andreas, and Patrick W. Schmitz. 2012. "Auctions with Anticipated Emotions: Overbidding, Underbidding, and Optimal Reserve Prices." *Scandinavian Journal of Economics* 114 (3): 808–30.
Rosenkranz, Stephanie, and Patrick W. Schmitz. 2007. "Reserve Prices in Auctions as Reference Points." *Economic Journal* 117 (520): 637–53.
Rothschild, Michael, and Joseph Stiglitz. 1976. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information." *Quarterly Journal of Economics* 90 (4): 629–49.
Salant, Yuval, and Ron Siegel. 2013. "Framing in Mechanism Design." Unpublished.
Sandroni, Alvaro, and Francesco Squintani. 2007. "Overconfidence, Insurance, and Paternalism." *American Economic Review* 97 (5): 1994–2004.
Sandroni, Alvaro, and Francesco Squintani. 2010. "Overconfidence and Asymmetric Information: The Case of Insurance." Unpublished.
Santos-Pinto, Luís. 2008. "Positive Self-Image and Incentives in Organisations." *Economic Journal* 118 (531): 1315–32.
Schumacher, Heiner. 2012. "Insurance, Self-Control and Commitment." Unpublished.
Schweizer, Nikolaus, and Nora Szech. 2013. "Optimal Revelation of Life-Changing Information." Unpublished.
Shunda, Nicholas. 2009. "Auctions with a Buy Price: The Case of Reference-Dependent Preferences." *Games and Economic Behavior* 67 (2): 645–64.
Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review* 97 (3): 999–1012.
Spiegler, Ran. 2006. "The Market for Quacks." *Review of Economic Studies* 73 (4): 1113–31.
Spiegler, Ran. 2011. *Bounded Rationality and Industrial Organization*. Oxford and New York: Oxford University Press.
Spinnewijn, Johannes. 2012. "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs." Unpublished.
Spinnewijn, Johannes. 2013. "Insurance and Perceptions: How to Screen Optimists and Pessimists." *Economic Journal* 123 (569): 606–33.

Stoughton, Neal M., Youchang Wu, and Josef Zechner. 2011. "Intermediated Investment Management." *Journal of Finance* 66 (3): 947–80.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112 (Special Issue 1): S164–87.

Thaler, Richard H., and Cass R. Sunstein. 2003. "Libertarian Paternalism." *American Economic Review* 93 (2): 175–79.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York: Penguin Books.

Tirole, Jean. 2009. "Cognition and Incomplete Contracts." *American Economic Review* 99 (1): 265–94.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics* 106 (4): 1039–61.

von Siemens, Ferdinand A. 2011a. "Heterogeneous Social Preferences, Screening, and Employment Contracts." *Oxford Economic Papers* 63 (3): 499–522.

von Siemens, Ferdinand A. 2011b. "Intention-Based Reciprocity and the Hidden Costs of Control." Unpublished.

von Siemens, Ferdinand A. 2012. "Social Preferences, Sorting, and Competition." *Scandinavian Journal of Economics* 114 (3): 780–807.

Warren, Patrick L., and Daniel H. Wood. 2010. "Will Governments Fix What Markets Cannot? The Positive Political Economy of Regulation in Markets with Overconfident Consumers." Unpublished.