Private Paternalism, the Commitment Puzzle, and Model-Free Equilibrium*

David Laibson
Harvard University

February 1, 2018

Abstract

Paternalism is a policy that advances an individual's interests by restricting his or her freedom. In a setting with present-biased agents, I characterize the scope of *private paternalism* – paternalism implemented by private institutions. Private paternalism arises from two channels: (i) agents who seek commitment because they hold sophisticated beliefs about their present bias, and (ii) agents (naive or sophisticated) who use *model-free forecasts* to choose organizations that have a history of generating high experienced utility flows for their members (O'Donoghue and Rabin 1999b). When naive consumers are common, private paternalism will be *shrouded*, explaining why commitment mechanisms are typically shrouded in the labor market (*the commitment puzzle*). Private paternalism has greater traction when production occurs in the formal sector instead of the informal (household) sector, where monitors are not always present, able, or willing to implement socially efficient forcing mechanisms.

JEL Codes: D15, D21, D6, D9, G4, H1, J3, L2, M5

*This paper describes the content of the Richard T. Ely Lecture, presented at the ASSA meeting on January 5, 2018. Omeed Maghzian provided excellent research assistance. I also received helpful advice from Dan Benjamin, John Beshears, Olivier Blanchard, James Choi, Xavier Gabaix, Matthew Gentzkow, Brigitte Madrian, Matthew Rabin, and colleagues at the Hebrew University, the University of Pennsylvania (CHIBE), and the ASSA annual meeting. I gratefully acknowledge financial support from the National Institute on Aging (P01AG005842, P30AG034532, R01AG021650) and the Pershing Square Fund for Research on the Foundations of Human Behavior.

1 Introduction

College instructors adopt course policies that force students to focus on their coursework: pop quizzes, classroom attendance requirements, cold calling, graded problem sets,[1] deadlines,[2] classroom wifi blocking, and classroom laptop bans.[3,4] In my experience, most students don't welcome these paternalistic restrictions. The unpopularity of such teaching policies is implicitly revealed by the way that colleges market themselves: public relations campaigns don't mention their paternalistic policies. No marketing materials would boast: "Choose our University because we have classrooms enabled with wifi blocking." Colleges are more likely to discuss national rankings, famous graduates, student endorsements, recreation centers, climbing walls, saunas, juice bars, golf simulators, driving ranges, ropes courses, water parks, and lazy rivers.[5] Even colleges that market their academic strength talk about research prowess and Nobel Laureates instead of explaining how effort and focus is coaxed from students.

Despite the lack of student enthusiasm, most colleges have embraced programs of *private paternalism*. I define paternalism as a policy that advances an individual's interests by restricting his or her freedom. Private paternalism is paternalism implemented by private institutions.[6] Universities aren't unique in their use of parternalistic policies. Many other private organizations deploy paternalistic policies, including firms in the for-profit sector. Such private paternalism is often shrouded, which suggests that these organizations' workers/customers/students/members do not explicitly demand paternalism. A firm doesn't recruit workers by telling them how it's going to limit their freedom. But firms do limit their workers' freedom with intermediate deadlines, progress reports, production tar-

[1] In my undergraduate classes, I now give two small problem sets each week.

[2] For example, a term paper might have more than one binding deadline: e.g., a deadline for a topic choice, a deadline for an initial draft, and a deadline for a final draft. Likewise, doctoral students are required to meet many intermediate milestones (each with a deadline) before receiving their degree: general exams, field exams, dissertation prospectus, semester-frequency seminar presentations, etc.

[3] Rockmore, Dan. 2014. "The Case for Banning Laptop in the Classroom." The New Yorker, June 6. https://www.newyorker.com/tech/elements/the-case-for-banning-laptops-in-the-classroom.

[4] If externalities were the sole problem with laptops, then those who wish to use laptops could be allowed to self-segregate into some area of the classroom, where their screens wouldn't be visible to those who prefer not to be near laptop users.

[5] Rubin, Courtney. 2014. "Making a Splash on Campus: College Recreation Now Includes Pool Parties and River Rides." September 14. https://www.nytimes.com/2014/09/21/fashion/college-recreation-now-includes-pool-parties-and-river-rides.html

[6] I contrast private paternalism with libertarian paternalism (Thaler and Sunstein 2003) in Subsection 7.1. In a nutshell, private paternalism is choice-restricting, unlike the pure form of libertarian paternalism.

gets, 'attendance' requirements during working hours, frequent evaluations, and mandatory health/retirement benefits. Firms *could* move towards more of a piece rate system (i.e., the worker is able to produce as much or as little as they want on a linear payment schedule), but, by and large, formal sector firms don't use this compensation system. Firms look more like classrooms in which bosses/teachers successfully coax productivity out of workers/students and then reward them for their high output.[7]

Private paternalism is the first half of a phenomena that I call the commitment puzzle: (i) private institutions help people solve self-control problems by using lots of forcing mechanisms (i.e., private paternalism) and (ii) these institutions don't market/advertise these forcing mechanisms, so the mechanisms remain at least partially shrouded. These forcing mechanisms are shrouded commitment technologies that are indirectly 'chosen' by agents who voluntarily join organizations that limit their freedom. In this lecture, I explain why such private paternalism is an equilibrium, even without explicit demand from its beneficiaries (e.g., workers, customers, students). As a corollary, I explain why private paternalism is typically shrouded. I describe the benefits of private paternalism relative to public paternalism and identify the domains where private paternalism will fail to be socially efficient.

The ideas in this lecture are related to the arguments made by Gregory Clark (1994). Clark views the work arrangements that arose during the Industrial Revolution as an implicit solution to self-control problems. "Whatever the workers themselves thought, they effectively hired the capitalists to discipline and coerce them. Even in the factories of the Industrial Revolution they were the ultimate masters of their fate, but weakness of will meant they delegated that mastery to the capitalists."[8,9] Clark also notes that modern work arrangements have continued in this vein: "When we look at the organization of work from the perspective of the twentieth century, the prevailing system, factory discipline, seems the natural and timeless way of organizing work. Under factory discipline workers face a very constrained choice. In return for their wage, they surrender to the employer complete com-

[7] Cadena et al. (2011) report the results of a field experiment – run collaboratively with a financial firm – which pursued these goals.

[8] Clark (1994), p. 131.

[9] This lecture was named by the AEA after Richard T. Ely, the organization's founder and first Secretary, who was also an early leader of the Progressive movement. Relative to Clark, Ely took a far more critical view of 19th century factory work conditions. However, Ely will not be remembered only as a champion of workers' rights. His writings are also fraught with eugenicism and racism. For analysis of his views, see: Weisberger, Bernard A., and Marshall Steinbaum. 2016. "The Exclusionist Legacy of Progressive Economics." Democracy. March 15. https://democracyjournal.org/arguments/the-exclusionist-legacy-of-progressive-economics/.

mand of their labor for a fixed period each day. The employer sets the pace of work and also dictates how workers will conduct themselves on the job."[10]

The specific approach that I take follows an argument proposed by O'Donoghue and Rabin (1999b). They characterize the socially efficient incentives that organizations should use if those organizations were motivated to help naive present-biased agents avoid procrastination. They assume that organizations are interested in maximizing their reputations as employers, arguing that "reputational pressures may induce firms to offer incentive contracts that are ex post acceptable to agents, which would imply that firms wish to induce efficient behavior."[11] O'Donoghue and Rabin's mechanism will serve as the fulcrum of this lecture: when prospective employees use backward-looking ('model-free') data to compare employers, firms have socially efficient incentives to create shrouded forcing/commitment technologies that coax high effort out of naive present-biased workers, which in turn engenders high levels of compensation in competitive equilibrium. I call this a *model-free equilibrium* because agents make forecasts about their future payoffs by using historical payoff data and *not* by using a structural model of how they will respond to future incentives.

The *model-free equilibrium* that I analyze is also related to the literature on fictitious play (Brown 1951; Fudenberg and Levine 1998). The fictitious play equilibrium concept assumes that agents playing a dynamic game use the past play of their opponents as a forecast of those opponents' future play. This experiential forecasting rule bears some resemblance to the model-free equilibrium that I will analyze. In both fictitious play and model-free equilibria, agents use historical data to predict future outcomes, without having a complete structural model of strategic play. In the model-free equilibria that I study, historical and experiential data are used to predict one's *own* future payoffs, rather than the actions of other players.

Section 2 introduces an illustrative model that I will use throughout this lecture. I assume that agents have present bias, which induces a self-control problem. Present bias (without commitment) leads to low, socially inefficient levels of effort. This section also describes the normative assumptions that are made throughout this paper. Section 3 characterizes equilibrium in an economy with forward-looking, *sophisticated* present-biased agents. Section 4 characterizes equilibrium – including employer-based exploitation of workers – in an economy with forward-looking, *naive* present-biased agents. Section 5 describes the commitment puz-

[10] Clark (1994), p. 131.

[11] O'Donoghue and Rabin (1999b), p. 783. There is a large economics literature on the role of worker and firm reputations: e.g., Kanemoto and MacLeod (1992), Holmstrom (1999), Levin (2003), Ely and Valimaki (2003), and Halac and Prat (2016).

zle: the existence of *shrouded private paternalism*. Section 6 resolves the commitment puzzle by assuming that people choose employers by using backward-looking experienced utility instead of using forward-looking predicted utility. I use the terms *model-based forecasting* and *model-free forecasting* to respectively distinguish the forward-looking and backward-looking equilibrium concepts that I study (e.g., Dayan and Niv 2008). I call the equilibrium based on backward-looking forecasts a *model-free equilibrium* because it does not rely on a structural theory of one's own future behavior. The model contains both a formal sector, where forcing mechanisms endogenously arise in equilibrium, and an informal (household) sector, where forcing mechanisms have no traction. Section 7 relates these results to the concept of private paternalism – paternalism implemented by private institutions. Section 8 discusses the limits and scope of private paternalism, including a discussion of the relative strengths of public vs. private paternalism, and the difficulty of generating efficient private paternalism in the informal sector. Section 9 concludes and discusses directions for future work.

2 Present-biased discounting and an effort task

In this lecture I focus on present bias, which is a specific bias that private paternalism can offset. In Subsection 8.4 I discuss the generalization of private paternalism to a wider class of psychological biases.

Time is indexed by t, with $t \in \{0, 1, 2, ..., T\}$, and that agents have intertemporal preferences characterized by present bias:[12]

$$U_t = u_t + \beta\delta u_{t+1} + \beta\delta^2 u_{t+2} + \beta\delta^3 u_{t+3} + \ldots + \beta\delta^{T-t} u_T \tag{1}$$

$$= u_t + \beta\left[\delta u_{t+1} + \delta^2 u_{t+2} + \delta^3 u_{t+3} + \ldots + \delta^{T-t} u_T\right]. \tag{2}$$

Accordingly, from self t's perspective, one util at date $t+1$ is worth $\beta\delta$ utils at date t. However, one util at date $t+2$ is worth δ utils at date $t+1$. Present bias is a wedge between the *present* and the *future*, but, from today's perspective, it does not affect tradeoffs between two future dates. The present bias parameter is β, and present bias is $1-\beta$ (the 'extra' discounting between the present and the future). Accordingly, when $\beta = 1$, discounting is

[12] See Strotz (1955), Phelps and Pollak (1967), Akerlof (1992), Laibson (1997), and O'Donoghue and Rabin (1999a). Laibson (1997) uses this functional form to capture the properties of hyperbolic discounting (see Chung and Herrnstein 1967; Ainslie 1975, 1992; and Loewenstein and Prelec 1992). See Cohen et al. (2017) for a review of the intertemporal choice literature.

purely exponential, and there is no present bias.

Present bias induces dynamically inconsistent preferences. For example, the preferences from the perspective of self t don't agree with the preferences of self $t+1$. Self $t+1$ has preferences

$$U_{t+1} = u_{t+1} + \beta\delta u_{t+2} + \beta\delta^2 u_{t+3} + \beta\delta^3 u_{t+4} + \ldots + \beta\delta^{T-(t+1)} u_T \tag{3}$$

$$= u_{t+1} + \beta\left[\delta u_{t+2} + \delta^2 u_{t+3} + \delta^3 u_{t+4} + \ldots + \delta^{T-(t+1)} u_T\right] \tag{4}$$

From self $(t+1)$'s perspective, one util at date $t+2$ is worth $\beta\delta$ utils at date $t+1$, contradicting the preferences of self t.

In general, I assume that $\beta < 1$ and δ is close to one. For a 'daily model' (in which each time period is one day), δ will be very close to unity. With this case in mind, I assume $\beta < 1$ and $\delta = 1$. Setting $\delta = 1$ also simplifies the notation. I'll note where this assumption affects my argument.

From a normative perspective, it is common to treat present bias as a bias (and not a normative preference) and strip it out of the planner's objective.[13] Indeed, even the agent agrees with this perspective when it comes to tradeoffs that involve all future selves. Accordingly, when we discuss normative benchmarks we will assume that the planner's objective is

$$U = u_0 + \delta u_1 + \delta^2 u_2 + \delta^3 u_3 + \ldots + \delta^T u_T.$$

2.1 An illustrative example

Consider the following simple model of effort. When an agent exerts effort e, she will experience immediate effort cost (measured in utils) of $-\frac{1}{2}e^2$ and she will experience a one-period delayed reward (measured in utils) of e. Assume that effort is exerted one period before the reward is experienced. I also assume that e has an upper bound, $\overline{e}$, and this upper bound is larger than β. In other words, $\overline{e} > \beta$.[14]

At date t, the worker's current objective is

$$\max_{e_t}\left(-\frac{1}{2}e_t^2 + \beta e_t\right).$$

[13] For example, see O'Donoghue and Rabin (1999b), Choi et al. (2003), or Carroll et al. (2009).

[14] This bound will play a role in Subsection 4.2.

This implies that her effort level will be given by the FOC

$$-e_t^* + \beta = 0,$$

where e_t^* is her equilibrium level of effort. Accordingly,

$$e_t^* = \beta.$$

However, when she plans at date t for her effort level at date $t+1$, she would prefer to maximize this objective

$$\max_{e_{t+1}} \beta \left(-\frac{1}{2} e_{t+1}^2 + e_{t+1} \right).$$

The associated FOC is

$$\beta \left(-e_{t+1}^\dagger + 1 \right) = 0,$$

where $e_{t+1}^\dagger$ is her desired future effort level. It follows that

$$e_{t+1}^\dagger = 1 > \beta = e_{t+1}^*.$$

Note that $e_{t+1}^* = \beta$ for the same reason that $e_t^* = \beta$.

When $\beta = 1$, self t and self $t+1$ fully agree on the best course of action for period $t+1$. When $\beta < 1$, self t wants self $t+1$ to work hard ($e_{t+1}^\dagger = 1$) and self $t+1$ prefers to work less hard ($e_{t+1}^* = \beta < 1$). This is the essence of the dynamic inconsistency problem that afflicts present-biased agents: self t and self $t+1$ don't agree on how to behave.

2.2 Beliefs about the choices of future selves

Since the beginning of the contemporary literature[15] on dynamically inconsistent preferences (Strotz 1955), there has been active discussion of the extent to which agents correctly anticipate their own future behavior. At one extreme, *sophisticated agents* understand that their future selves will hold preferences that contradict the preferences of the current self. In the illustrative example that I have presented, early sophisticated selves will choose to commit later selves to exert effort $e = 1$ in every future period if commitment is free. If commitment

[15] Many earlier scholars, including Adam Smith and David Hume, also understood these psychological forces and wrote about them. See Ashraf, Camerer, and Loewenstein (2005) and Cohen et al. (2017) for selected references.

is not free (or if non-contractible uncertainty makes commitment counterproductive), such commitments may not be chosen (see Laibson 2015).

At the other extreme, *naive agents* believe that their future selves will hold preferences that match the preferences of the current self (with respect to utility tradeoffs in the future). Accordingly, naive agents have no motive to commit their future selves, unless commitment is the ancillary (undesired) consequence of an action that is actually motivated by some other goal: e.g., pre-purchasing theatre tickets to a musical that is likely to sell out.

O'Donoghue and Rabin (1999a) introduced the concept of *partially naive agents* who have a partial understanding of their dynamic inconsistency. Specifically, partially naive agents believe that future selves will have present bias characterized by a present-bias parameter $\widehat{\beta}$, where $\widehat{\beta}$ is strictly greater than the true value of β. As $\widehat{\beta}$ falls to β, partially naive agents become fully sophisticated. As $\widehat{\beta}$ rises to 1, partially naive agents become fully naive. Because partially naive agents recognize some of their dynamic inconsistency ($\beta < \widehat{\beta} < 1$), they will also choose commitment devices when commitment is free.

3 Sophistication and commitment

If agents are sufficiently sophisticated, non-contractible uncertainty is sufficiently low, and commitment technologies are sufficiently inexpensive, then agents will choose pure commitment technologies (Laibson 1997, 2015). In a rapidly growing literature, economists have demonstrated that there is demand for a wide range of explicit commitment technologies. This literature is reviewed in Bryan, Karlan, and Nelson (2010) and Cohen et al. (2017).[16]

This new empirical literature has two interesting limitations. First, almost all of the papers demonstrating a demand for pure commitment are lab or field experiments that study commitment mechanisms designed by behavioral economists, as opposed to analysis of pure commitments that already exist in markets.[17] Second, the experimental participants who

[16] For example, see Ariely and Wertenbroch (2002), Ashraf, Karlan, and Yin (2006), Houser et al. (2010), Karlan, Giné, and Zinman (2010), Kaur, Kremer, and Mullainathan (2010, 2015), Royer, Stehr, and Sydnor (2012), Bisin and Hyndman (2014), Augenblick, Niederle, and Sprenger (2015), Schilbach (2015), and Beshears et al. (2017), Alsan et al. (2017), and Cho and Rust (2017).

[17] One exception is Cho and Rust (2017), who show that many Korean consumers reject free credit. Wertenbroch (1998) is another possible exception, depending on how one interprets this paper. Wertenbroch studies the fact that consumers tend to buy candy and other 'sin' goods in small packages (e.g., 'single serving'), implicitly limiting their future consumption. However, this may not be commitment, but instead a forecast by the consumer that she only wants to eat a single serving of these foods and hence shouldn't buy in volume (at a per-unit discount). Naive agents (who do not seek commitment) will make such low-consumption forecasts.

are willing to make commitments are rarely willing to *pay* to make those commitments. In other words, some experimental participants are willing to restrict their choice set, but these subjects are usually not willing to pay for this privilege.[18]

If one counts only pure commitment mechanisms that have not been created by behavioral economists,[19] the list is small. Market-based examples include, web-blocking software (e.g., the Freedom app) and alarm clocks that jump off nightstands.[20] It is not easy to expand this list if one keeps the conceptual bar high (i.e., pure commitments, as opposed to commitments that have ancillary benefits that may explain the choice-reducing behavior). In Korea, consumers turn down free credit (Cho and Rust 2017). Many people delete apps from their smartphone or iPad to force themselves to stop using the app (rather than freeing up memory).[21] The list of pure commitments is not long.

When commitments exist in real markets, they are usually embedded in a broader set of mechanisms/policies, so they aren't pure, and they aren't even perceived as self-generated. In other words, there is very little commitment for commitment's sake, but people often find themselves working/studying in institutions that use forcing mechanisms: e.g., a firm with frequent progress reports, supervisor check-ins, and binding deadlines. In this sense, people don't appear to be choosing to commit themselves, but rather accepting restrictions on their freedom as part of a grand bargain with certain counterparties.

4 Naiveté, freedom, and exploitation

Naive agents believe that the preferences they hold today will be identical to the preferences that they will hold in the future. In other words, naive agents believe that their preferences are dynamically consistent (and anchored by their current preferences). So the plans they make today will be faithfully executed by their future selves. W.C. Fields expressed this view succinctly when he said, "Now don't say you can't swear off drinking; it's easy. I've

[18] Schilbach (2015) is a notable exception. Beshears et al. (2017) is also an exception but the amounts that are being paid for commitment in this study are small (in most cases, less than $1 per year).

[19] The self-control website stickK was created by behavioral economists (Dean Karlan and Ian Ayres). Behavioral economists also report using a host of commitment behaviors of their own design: for example, a coauthor who commits to pay a penalty if he doesn't finish a draft on time, doctoral students who create a communal fund that they pay into whenever they oversleep, a doctoral student that signs up for a research presentation as a commitment device, and a faculty member that bets that he will lose a pre-specified amount of weight by a deadline.

[20] For the jumping alarm clocks, see https://nandahome.com/.

[21] I have saved thousands of hours by deleting a chess app from my iPad.

done it a thousand times."[22]

For naive agents, there is no reason to commit themselves because they believe their future selves will always agree with their current selves. Naive agents will never choose commitment for its own sake. All else equal, naive agents prefer freedom. They would never choose to bind themselves unless there was some compensating tradeoff.

Most of the evidence from the field supports a substantial degree of naiveté, and some studies even find evidence for full naiveté: e.g., Dellavigna and Malmendier (gyms; 2006); Acland and Levy (gyms; 2015); Goda et al. (retirement savings; 2015); Augenblick and Rabin (experimental effort task; 2017); Levy et al. (smoking cessation; 2017); and Kuchler and Pagel (credit cards; 2017).

In a pair of experiments – one in a large class and one online – Annastasia Fedyk (2017) reports evidence for *asymmetric* naiveté. Specifically, people are relatively sophisticated about *others'* present bias, but fail to anticipate their *own* future present bias. In the classroom experiment, students systematically underestimate how late they will turn in an assignment but hold much more accurate beliefs about their classmates' propensity to procrastinate. In Fedyk's online experiment, participants engage in a real-effort task (cf. Augenblick, Niederle, and Sprenger 2015, and Augenblick and Rabin 2017). Fedyk estimates $\beta = 0.82$ for her sample of experimental participants. These participants perceive other participants' β to be 0.87 (i.e., $\widehat{\beta}_{Other} = 0.87$), implying a substantial degree of interpersonal sophistication. Note that perfect interpersonal sophistication would correspond to $\beta = \widehat{\beta}_{Other}$. By contrast, experimental participants perceive $\widehat{\beta}_{Self} = 1.03$,[23] which implies perfect *intra*personal naiveté. Fedyk's work provides formal evidence for a hypothesis that has long been folk wisdom. As Mark Twain put it: "Nothing so needs reforming as other people's habits."[24]

Indirect evidence for naiveté comes from the absence of a thriving commitment industry. Very few pure commitment devices are currently being marketed, with the exception of the small set of products reviewed in Section 3. When commitment is present in markets, it is almost always bundled with other features. So the worker/student/consumer is not buying commitment purely for commitment's sake, but is instead buying commitment along with some other features. For example, a mortgage that provides credit for purchasing a home *and* an ancillary system of forced savings (principal repayments). Later in this lecture,

[22] W.C. Fields, (1938). *The Temperance Lecture*, radio broadcast. See Shapiro (2006), p. 256.

[23] This estimate is statistically indistinguishable from unity.

[24] Mark Twain, (1894), *Pudd'nhead Wilson*, Chapter XV prefatory quote. Charles L. Webster & Company.

I will explain why commitments are almost always bundled with other products and work arrangements, and thereby shrouded.

In light of the evidence for naiveté, I will hereafter study the extreme case of perfect naiveté (i.e., $\beta << 1$ but $\widehat{\beta}_{Self} = 1$). Though it is likely that people have some (domain-specific, partial) awareness of their own present bias, I study the limiting case, $\widehat{\beta}_{Self} = 1$, for simplicity.

4.1 Naive agents in home production

If agents are naive, their behavior will depend upon the setting in which they are operating and the counterparties with whom they negotiate. If the counterparty is a firm, the agent may be exploited, a possibility that I consider in the next subection. In home production, by contrast, the agent will simply fail to live up to her good intentions for high effort.

I now revisit the earlier example, where I showed that an agent will expect (one period ahead) to work with effort $\widehat{e} = 1$, when in fact, she will work with effort $e = \beta$. Accordingly, she anticipates payoff

$$-\frac{1}{2}\widehat{e}^2 + \widehat{e} = \frac{1}{2}.$$

However, she will consistently underperform relative to this benchmark. Her actual (undiscounted) payoff will be

$$-\frac{1}{2}\beta^2 + \beta.$$

This payoff will be strictly less than the worker's anticipated payoff, $\frac{1}{2}$ (as long as $\beta \neq 1$).[25]

4.2 Naive agents in a monopsony labor market

If a naive, present-biased agent/worker interacts with a sophisticated firm (i.e., a firm that understands the worker's present bias), and the worker evaluates the firm's offers using a forward-looking (structural) theory of her own future actions, then the firm will typically exploit the agent's naiveté and may cause the agent to earn less than her outside option (e.g., O'Donoghue and Rabin 1999b). I illustrate this exploitative equilibrium in the current subsection (monopsony) and the subsection that follows (competitive equilibrium). In the next section of the paper, I critique the underlying assumptions that support these exploitative equilibria, and explain why these equilibria will sometimes *not* arise in practice.

[25] Note that $-\frac{1}{2}\beta^2 + \beta$ is a concave parabola with a maximum of $\frac{1}{2}$ at $\beta = 1$.

Readers who want to see the details of the exploitative equilibrium should read the current subsection. Readers wishing to skip ahead can take the exploitative equilibria for granted (Subsections 4.2 and 4.3), and jump to Section 5, where I critique such equilibria and argue that they describe only a subset of the relationships between firms and workers/customers.

To understand exploitative equilibria, consider the following game played between a monopsonist firm and a worker. The game has four periods, which are summarized below.

Period 1: Firm commits to an enforceable wage schedule mapping effort to wage: $w(e)$.

Period 2: Worker chooses between the firm's wage schedule and the worker's outside option, which has undiscounted payoff z and is received in period 4). In other words, the worker's choice is

$$\max\left\{\beta\max_{e}\left[-\frac{1}{2}e^2+w(e)\right],\ \beta z\right\}=\max\left\{\max_{e}\left[-\frac{1}{2}e^2+w(e)\right],\ z\right\}.$$

This choice is binding for the next period.[26]

Period 3: If the worker chose the firm's wage schedule, she chooses effort e subject to present bias that she had not previously anticipated. The cost of effort is experienced in this period. In other words, the worker chooses e in the following maximization problem:

$$\max_{e}\left\{-\frac{1}{2}e^2+\beta w(e)\right\}.$$

Period 4: If the worker chose the firm's wage schedule, she experiences a utility flow derived from her wage payment, $w(e)$, or, if she chose her outside option, she experiences utility flow z.

In equilibrium, the firm may exploit the worker, even if the firm has competition in the labor market (see the next subsection for the competitive labor market case). In this subsection, I study a monopsonist firm.

The firm will offer a wage schedule that generates a perceived payoff that is at least as good as z. The firm's problem can be reduced to the choice of four scalar values: $\widehat{e}, w(\widehat{e}), e^*, w(e^*)$. Effort level $\widehat{e}$ is the quantity of effort that the worker expects to exert once

[26] The agent can choose again once the game "ends" after period 4. There are many (unmodeled) reasons to believe that such temporary stickiness exists, including frictional transition costs.

she starts working (the expectation that is formed in period 2). Wage $w(\widehat{e})$ is the wage that the firm will pay the worker for effort $\widehat{e}$. Effort level e^* is the quantity of effort that the worker will revert to once the moment to work arrives (period 3); she is unexpectedly subject to present bias, which she hadn't foreseen when joining the firm. Finally, wage $w(e^*)$ is the wage the firm promises to pay the worker for effort e^*. Firms receive output e^* and pay wage $w(e^*)$, so firms maximize the following expression subject to a participation constraint (P) and an incentive compatability constraint (IC):[27]

$$\max_{e^*,\, w(e^*),\widehat{e},\, w(\widehat{e})} e^* - w(e^*), \tag{5}$$

subject to the constraints[28]

$$-\frac{1}{2}\left(\widehat{e}\right)^2 + w(\widehat{e}) \geq z \tag{P}$$

$$-\frac{1}{2}\left(e^*\right)^2 + \beta w(e^*) \geq -\frac{1}{2}\widehat{e}^2 + \beta w(\widehat{e}). \tag{IC}$$

In equilibrium, the firm recruits the worker with a misleading offer: work with effort $\widehat{e}$ and get paid wage $w(\widehat{e})$. The firm sets up the contract so that the naive worker has an unanticipated incentive to switch (after taking the job) to work effort e^*, with payment $w(e^*)$. The firm sets $\widehat{e}$ and $w(\widehat{e})$ so the worker is indifferent between the firm's offer and the worker's outside option, z (this is the participation constraint, P). The firm sets e^* with payment $w(e^*)$ so the worker is indifferent between staying with her original effort plan and switching to this new level of effort (this is the incentive compatability constraint, IC). P and IC are both satisfied at the firm's optimum.

The appendix derives the solution to the firm's problem. The appendix shows that the firm exploits the worker by offering a high level of compensation for the maximal level of effort. Specifically, the firm offers the worker the highest feasible level of effort, $\widehat{e} = \overline{e}$, and a sufficiently high wage, $w(\widehat{e})$, to make this level of effort appear to be desirable ex-ante to the naive worker:

$$w(\widehat{e}) = z + \frac{1}{2}\overline{e}^2. \tag{6}$$

But the firm allows the worker to opt out of this arrangment (once she is on the job) and

[27] To study the case of economic interest, we assume that $z \leq \frac{1}{2}\beta + \frac{1}{2}\left(\frac{1}{\beta} - 1\right)\overline{e}^2$. This restriction on z implies that the firm is able to offer a contract to the work that incents the worker to choose the firm over the worker's outside option.

[28] These constraints bind in the monopsonistic equilibrium. In this subsection, we will express them as equalities.

revert to a lower effort level, $e^* = \beta$, and receive wage

$$w(e^*) = z + \frac{1}{2}\beta + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\overline{e}^2. \tag{7}$$

Note that this is worse than the worker would have done, had she originally chosen her outside option, because

$$\begin{aligned} -\frac{1}{2}\left(e^*\right)^2 + w(e^*) &= -\frac{1}{2}\beta^2 + z + \frac{1}{2}\beta + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\overline{e}^2 \\ &< z + \frac{1}{2}\beta\left(1 - \beta\right) + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\beta^2 \\ &= z. \end{aligned}$$

The inequality follows from the assumption $\overline{e} > \beta$. We now consider the payoff to the firm. Output is β and wages are $z + \frac{1}{2}\beta + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\overline{e}^2$. So the firm's profits are

$$\begin{aligned} \pi &= \beta - \left[z + \frac{1}{2}\beta + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\overline{e}^2\right] \\ &= \frac{1}{2}\beta - z + \frac{1}{2}\left(\frac{1}{\beta} - 1\right)\overline{e}^2 \\ &\geq 0 \end{aligned}$$

The last inequality follows from a restriction on z (see footnote 27).

4.3 Naive agents in a competitive labor market

The fundamental logic of these results doesn't change when the present-biased agent interacts with firms that compete with one another.[29] Returning to the illustrative example, two key properties also emerge in competitive equilibrium: (i) the worker is recruited expecting to work at effort level $\widehat{e} = \overline{e}$ and then switches to a less intensive effort level $e^* = \beta$, and (ii) the worker may end up with less than her outside option. In competitive equilibrium I have the same setup/timing, but now there is more than one firm (with Bertrand competition), so the firms make zero profit in equilibrium.

This equilibrium causes worker payoffs to rise (weakly), but the analysis from the previous

[29] However, the participation constraint now may not bind.

subsection doesn't change. To distinguish the previous monopsony wage mapping from the current competitive equilibrium wage mapping, I let $w^c(e)$ represent the competitive equilibrium wage mapping. In competitive equilibrium firms need to disgorge all of their rents in the second-stage of the relationship. So $w^c(e^*) = \beta$. Then incentive compatibility implies that:

$$-\frac{1}{2}(e^*)^2 + \beta w^c(e^*) = -\frac{1}{2}\beta^2 + \beta^2 = -\frac{1}{2}\overline{e}^2 + \beta w^c(\overline{e}).$$

Accordingly,

$$w^c(\overline{e}) = \frac{1}{2}\beta + \frac{1}{2\beta}\overline{e}^2.$$

We can compare the realized competitive equilibrium wage to the realized wage in the monopsony equilibrium. From equation (7) the monopsony realized wage is

$$w(e^*) = z + \frac{1}{2}\beta + \frac{1}{2}\left(1 - \frac{1}{\beta}\right)\overline{e}^2 \leq \beta = w^c(e^*).$$

So $w^c(e^*)$ has weakly risen relative to the realized wage in the monopsonist case.[30] See Laibson and Yariv (2006) for a similar analysis, which shows that firms do not make excess profits in competitive equilibrium, even when agents are present-biased and naive.

4.4 Exploitative contracts in practice

Such exploitative contracts frequently do arise, especially in product markets. For example, Dellavigna and Malmendier (2004, 2006) and Ausubel and Shui (2005) convincingly argue that exploitative contracts arise in the following service and goods markets: consumer credit (e.g., credit cards), gambling, health clubs, life insurance, mail order businesses, mobile phones, vacation time-sharing. A rich literature has developed to study the many ways that

[30] A similar inequality applies to the anticipated wage in competitive equilibrium. From equation (6) the anticipated monopsony wage is

$$\begin{aligned} w(\widehat{e}) &= z + \frac{1}{2}\overline{e}^2 \\ &\leq \frac{1}{2}\beta + \frac{1}{2}\left(\frac{1}{\beta} - 1\right)\overline{e}^2 + \frac{1}{2}\overline{e}^2 \\ &= \frac{1}{2}\beta + \frac{1}{2\beta}\overline{e}^2 \\ &= w^c(\widehat{e}). \end{aligned}$$

firms interact with naive agents, including the role for regulation.[31]

However, exploitative equilibria of the sort characterized in the previous subsections seem counterfactual. *Workers* rarely complain that they ended up working less hard then they thought they would work and that this 'underworking' is exploitative. Rather, workers complain about being overworked. So the exploitative equilibrium in the previous subsection seems largely counterfactual as applied to labor markets.

5 The commitment puzzle

Until now, I have emphasized the perverse properties of equilibria with naive present-biased agents. These equilibria feature misperceptions (e.g., bait and switch) and exploitation. Such perverse equilibria show up in many product markets (e.g., Dellavigna and Malmendier 2004, 2006), but there are other domains, especially labor markets, where such exploitation seems to be the exception rather than the rule.

As discussed at the end of Subsection 4.2, the exploitative equilibria are characterized by the expectation of high effort and high pay, followed by an unanticipated propensity for workers to switch to low effort and low pay. However, there are many firms where the work environment is designed to coax high effort out of workers instead of tricking them into slacking off. Many work environments are carefully designed to help workers with self-control problems reach a high level of productivity. These work environments deploy forcing mechanisms like deadlines, meetings, reviews, progress reports, etc., to push workers to be more productive. Employers also provide many paternalistic benefits like highly subsidized health and retirement programs, although tax-arbitrage and screening provide another rationale for such subsidies.

Work environments are one of many private institutions that are characterized by bundled forcing mechanisms. As I reviewed above, colleges feature myriad mechanisms that paternalistically encourage student effort. Even some private financial arrangements have bundled commitment features – e.g., mortgages with a fixed repayment stream and mandatory principal repayment. Commitment features are also bundled into the structure/operation of many social organizations: e.g., religious groups, social clubs, and marriage.

[31] For example, see Ausubel (1991), Hall (1997), Eliaz and Spiegler (2004, 2006), Gabaix and Laibson (2006), Gottlieb (2008), Bucks and Pence (2008), Stango and Zinman (2009), Grubb (2009, 2015), Heidhues and Koszegi (2010, 2015, 2017), Armstrong and Vickers (2012), Gerardi, Goette, and Meier (2013), Warren and Wood (2014), Grubb and Osborne (2015), Gurun, Matvos, and Seru (2016), Ru and Schoar (2016), and Johnen (2017).

None of this would be puzzling if these forcing mechanisms were widely advertised. But, by and large, these forcing mechanisms aren't being advertised or marketed. Firms don't go out of their way to tell applicants that the firm is going to reduce the scope of procrastination with deadlines and progress reports. Likewise, colleges don't boast about all of the work they are going to coax out of their students. We fear commitment – in marriage and in many other domains – rather than seeking it out.

Herein lies the *commitment puzzle*. On one hand, lots of thriving institutions have bundled commitment features that appear to be specifically designed to help agents overcome their self-control problems. On the other hand, these institutions generally don't market these commitment features – i.e., the forcing mechanisms are shrouded.

If people are sophisticated (or only partially naive), they should appreciate commitment features and firms should be eager to market them. So sophistication implies that the shrouding of forcing mechanisms is a puzzle. Is it then possible to resolve the commitment puzzle by assuming that people are completely naive? Here too we encounter a problem. If people are naive, they should not appreciate the commitment features, and it will be more profitable for firms to exploit these agents than to help them overcome their self-control problems (see the preceding two subsections and the exploitative equilibria therein). If people are naive, they should not want commitment devices, and firms won't provide them in the settings we have described so far.

This appears to leave us in a conceptual cul de sac. What could resolve the commitment puzzle?

6 When do naive agents "choose" commitment?

Economists tend to assume that people evaluate contracts – e.g., a job offer – by building a complete probability-weighted prediction of their future actions under the specific terms of that contract. I'll refer to this as model-based forecasting. In simple settings, like the employment relationship discussed in this lecture (see Subsections 4.2 and 4.3), it is plausible that new employees might enter the employment relationship with a model-based forecast of their actions: $\widehat{e}$. The illustrative employment setting that I have described has no uncertainty and no dynamics (one deterministic state), and the action space is one-dimensional (effort, e). The contract is a simple univariate function mapping effort levels to wages: $w(e)$.

Real employment relationships are incomparably more complex. The action space and

the (dynamic) state space both have numerous dimensions, and the contingent compensation function has so many branches that nobody could even write it down, let alone implement a probability-weighted model of future states and actions. Few potential employees would ever try to gather *all* of the necessary information to calculate this probability-weighted model: onboarding programs, human resource manuals, travel expense reimbursement policies, workplace training programs, quarterly review processes, performance measurement systems, coworker assignment policies, promotion guidelines, benefits manuals, rules for medical leaves, mental health programs, etc. Gathering all of this information and then building a best-response function (to all states of nature) is a daunting task.

In complex settings, people rely instead on other types of information: reputation, ratings, and summary measures of experienced utility.[32] For example, college applicants can easily look up a school's rating compiled by *U.S. News and World Report* (or one of its competitors). That takes a minute or two. Reading a student handbook takes hours. Reading and evaluating hundreds of course syllabi would take weeks.

There are many proxy measures that provide estimates of what life would be like with a new employer: Glassdoor (online) ratings, reports of how many hours (how 'hard') existing employers work each week, answers to broad questions about job quality (e.g., "How do you like working here?"). Firms report that these reputational factors play an important role in successfully recruiting new workers.[33] Potential employees can (imperfectly) forecast their own future work satisfaction from such reports, without developing a model-based forecast of their future behavior. Likewise, existing employees might use their own experienced utility at their employer – rather than a model-based theory of their behavior – to forecast their future utility flows at their employer. Memories don't reflect *forward-looking* present bias. Instead, memories reflect an opposite weighting bias: recent events get more weight than events farther back in time (e.g., Malmendier and Nagel 2016).[34]

I want to contrast model-based forecasting (the nearly universal assumption that we make in economic analysis), with backward-looking, experientially grounded model-free forecasting. The terms *model-based* and *model-free* are taken from the neuroscience literature (e.g.,

[32] Experienced utility is a measure of subjective well-being during some period of time. Experienced utility is typically measured with questions about affect, mood, emotions, and/or life satisfaction.

[33] Burgess, Wade. 2016. "A Bad Reputation Costs a Company at Least 10% More Per Hire." *Harvard Business Review*. https://hbr.org/2016/03/a-bad-reputation-costs-company-at-least-10-more-per-hire.

[34] See Kahneman, Fredrickson, Schrieber, and Redelmeier (1993) for other memory distortions, including end effects (experiences at the end of a stream of experienced utilty get more weight than experiences in the rest of the stream).

Dayan and Niv 2008). If people make decisions based on ratings, reputation, memory, proxy measures, and experienced utility – in other words, model-free forecasting – then equilibrium will sharply diverge from the analysis that I've described in previous sections. The equilibrium analysis in Subsections 4.2 and 4.3 assumed that agents generate model-based forecasts of their future behavior. If people use model-free forecasting, equilibrium will instead align with the (constrained) efficient contracts that O'Donoghue and Rabin (1999b) characterize in their analysis.

6.1 Model-free equilibrium

I now study a variant of the effort model from above. I change the model slightly to focus on the key issues that I want to highlight. First, I make the action choices more complex and the production/payoff functions arbitrary, emphasizing the general structure of the setting. Second, I omit the choice *between* formal sector production and home production. Now everyone works in a firm and *also* works at home (e.g., house cleaning, meal preparation, yard work, home repairs, and/or child care). This change is made to highlight the differences in efficiency in formal sector production and informal sector (home) production. Third, I incorporate model-free choice and call the resulting framework a *model-free equilibrium*.

I model the labor market as a game among identical firms in a competitive labor market and a naive, present-biased worker (or workers).

I now describe the notation. The agent chooses a formal sector action, a_f, from a compact space: $a_f \in A_f$. Actions in the formal sector have immediate utility payoff for the formal sector employee of $-e^f(a_f)$. Actions in the formal sector produce real-valued output $F(a_f)$ for the firm. Payments from the formal sector employer have delayed utility benefits $u^f(w(a_f))$, where $w(a_f)$ is a formal sector (real-valued) compensation schedule. The firm's profits are

$$F(a_f) - w(a_f).$$

The agent also chooses a home production action from a compact space, $a_h \in A_h$. Actions at home have immediate utility payoff $-e^h(a_h)$. Payoffs at home have delayed utility benefits $u^h(a_h)$.[35]

The agent's total payoff (omitting present bias), combines her (separable) payoff from

[35]Some actions may have immediate payoffs. I focus on the leading case in which actions have delayed payoffs: e.g., grocery shopping at t yields a well-stocked kitchen at $t+1$.

formal sector production and informal sector (home) production:

$$U\left(a_f, a_h\right) \equiv \left[-e^f(a_f) + u^f(w(a_f))\right] + \left[-e^h(a_h) + u^h(a_h)\right]. \tag{8}$$

I assume that all of the payoff functions described above are continuous, so all of the maxima defined below are attained.[36]

The game has four periods that match the periods in the previous game (Subsection 4.2).

Period 1: Each firm i commits to an enforceable wage schedule mapping a_f to wage $w_i(a_f)$. Agents observe the realized/remembered utility flows associated with each of these wage schedules (by talking to peers, viewing ratings/rankings, gathering other proxy data, or using their own personal experience/memories in the formal sector). In other words, agents gather data that proxies for

$$U_i^f \equiv -e^f(a_{f,i}^*) + u^f(w_i(a_{f,i}^*)),$$

where $a_{f,i}^*$ is the equilibrium/empirical/experienced formal-sector behavior generated by wage schedule $w_i(a_f)$. Note that the model-free summary measure, U_i^f, places equal weight on $-e^f(a_{f,i}^*)$ and $u^f(w_i(a_{f,i}^*))$ because the data is gathered from backward-looking memories and evaluations.[37]

Period 2: The agent chooses a firm using the model free data collected in period 1. In other words, the agent uses model free data to choose the most appealing firm:

$$\arg\max_i \beta U_i^f = \arg\max_i U_i^f.$$

This choice is binding for the next period.

Period 3: Agent chooses actions a_f and a_h subject to a present-bias wedge that she had not previously anticipated (i.e., the agent is naive with respect to her present bias). The cost of effort is experienced in period 3. Formally, the worker chooses a_f and a_h

[36] This is an application of the Weierstrass extreme value theorem. Recall too that the action spaces are compact.

[37] If there were recency bias, then $u^f(w_i(a_{f,i}^*))$ would actually be weighted more than $-e^f(a_{f,i}^*)$ (and not vice versa).

in the maximization problem:

$$\max_{a_f,\, a_h} \left\{ \left[-e^f(a_f) + \beta u^f(w(a_f)) \right] + \left[-e^h(a_h) + \beta u^h(a_h) \right] \right\}.$$

Period 4: The agent experiences utility flows $u^f(w(a_f))$ and $u^h(a_h)$. These are the agent's payoffs, which result from the actions that she chose in period 3.

Before analyzing this model, it is important to note that the formal sector in this model would generically feature an inefficient equilibrium *if* agents (i.e., formal sector workers) used model-based forecasts. This is a corollary of the analysis in Subsection 4.3. Specifically, the equilibrium would not generally maximize social welfare (equation 8). Instead, the equilibrium would feature bait-and-switch tactics of the type analyzed in Subsection 4.3.

However, I am now assuming that agents do not use model-based forecasting. Instead, they use model-free forecasting, which exploits peer reports, rankings, memories, and other proxies for experienced utility (see the assumptions associated with period 1). Under the assumption of model-free forecasting, socially efficient actions are implemented in the formal sector (i.e., actions are not affected by present bias).[38] Specifically, in equilibrium in the formal sector, workers choose an action a_f^*, which satisfies:

$$a_f^* \in \arg\max_{a_f} \left[-e^f(a_f) + u^f(F(a_f)) \right].$$

Moreover, equilibrium in the formal sector is not (generically) characterized by bait-and-switch tactics.

This property derives from the fact that a firm in a competitive market will not attract a worker unless the firm delivers the (weakly) highest total payoff to the worker, as reflected by the payoff measure that the worker uses when she picks the firm (in period 2). Using this metric, the highest feasible payoff to the worker is

$$\max_{a_f} \left[-e^f(a_f) + u^f(F(a_f)) \right].$$

If we use this equilibrium concept to study the original functional forms of Section 2, we see that the firms induce their employees to exert effort $e = 1$ in period 3 by *not* giving

[38] The assumption $\delta = 1$ is being used to generate this efficiency result. With $\delta < 1$, this equilibrium would not be socially efficient because the agent isn't using δ-discounted flows (in Period 2) to choose among firms.

them the option to slack off and choose a lower effort level (i.e., $e = \beta$). Firms coax high effort out of their employees by making the low effort option unappealing. For example, firms can sufficiently sanction workers who exert low effort so that the firms obtain $e = 1$ in equilibrium. In this equilibrium, firms push their employees to work hard and pay them high wages that go along with this high effort. This is the opposite of the exploitative equilibrium in which firms effectively encourage their workers to slack off so they can pay them very little (recall Subsections 4.2 and 4.3).

To summarize, if prospective employees use model-free forecasts to choose among employers, equilibrium contracts will be characterized by efficient forcing mechanisms (i.e., commitment), even if agents are naive. In competitive equilibrium, workers will choose $a_f^* \in \arg\max_{a_f} \left[-e^f(a_f) + u^f(F(a_f))\right]$ and get paid $F(a_f^*)$.

6.2 Resolution of the commitment puzzle

This framework provides a potential resolution to the commitment puzzle. With model-free forecasting, firms provide forcing mechanisms that look like commitment devices. Because these forcing mechanisms are being provided to *naive* agents, firms do not highlight the fact that their incentive contracts will turn out to be strictly binding. The commitment is implicit. Commitment is supported in equilibrium because it produces high levels of experienced/remembered utility for workers, thereby enabling the firms to recruit more workers under that successful compensation/work schedule.

6.3 Inefficiency in home production

In the current section, I have focused on equilibrium outcomes in the formal sector. As I have argued, those outcomes will be socially efficient. But the forces that create this efficiency do not apply to home production. In the formal sector, the agent chooses an employer in period 2 and is stuck with that employer through period 4.[39] Because the worker can't instantaneously switch employers in period 3, the employer is able to force a high (efficient) level of costly effort from the worker in period 3 (by using an incentive mechanism that punishes low effort/output). This forcing mechanism is not present in home production. Accordingly, the worker chooses actions in home production that are influenced by present bias and accordingly not socially efficient.

[39] At the end of the 4th period, the entire game can repeat and the worker can choose a new firm in the next iteration of period 2.

To illustrate this point, assume that a worker has two production systems – formal and informal/home – that are identical in functional form. Using our earlier example again, suppose that these symmetric effort costs are respectively $-\frac{1}{2}e_f^2$ and $-\frac{1}{2}e_h^2$. Suppose also that firms' production functions are linear in e_f, so that $F(e_f) = e_f$. Finally, suppose that $u_f(w) = w$ and $u_h(e_h) = e_h$. In equilibrium, $e_f = 1$ and $e_h = \beta$.[40] The socially efficient level of production is only achieved in the formal sector. Home production is characterized by an inefficiently low level of production (unless some other forcing mechanism is present at home, a question that I will return to in Section 8).

7 Private paternalism

This lecture has discussed two channels by which private, for-profit firms choose policies that advance workers' interests by restricting workers' choices. The first channel of private paternalism is frequently discussed in the intertemporal choice literature: if workers are present-biased and *sophisticated*, and they use model-based reasoning, they will explicitly ask firms to limit their choices (e.g., Laibson 1997, Laibson 2015, Lin 2017). While this channel exists in principle and has been elicited in lab and field experiments (see Section 3), it does not seem to be a strong force in most real markets. Very few firms recruit workers or customers by boasting about the firm's choice-limiting policies/products/services.

There is also a second channel of private paternalism that will arise for both sophisticated and naive agents. This channel depends on two conditions: (i) agents/workers use model-free data to choose among prospective firms, and (ii) those agent/worker choices have some degree of temporary stickiness (so the worker who chooses a firm in period 2, stays at that firm in period 3). When these two conditions are met, firms will choose freedom-restricting policies that engender social efficiency. These forcing mechanisms will be shrouded if there are enough naive workers in the population. Naive workers will not (fully) understand why these restrictive employer policies are desirable. Using model-free learning, naive agents will be able to identify the firms that are associated with good outcomes for their employees, but the naive agents won't understand the causal mechanisms. Naive agents view freedom-restricting policies as intrinsically adverse (to the extent that they use at least some model-based reasoning). Accordingly, firms will have no reason to trumpet freedom-restricting policies, even though firms have an incentive to adopt such policies.

[40] Equilibrium formal sector wage is $w = 1$.

These two channels demonstrate that private paternalism – i.e., paternalism that is implemented by private institutions without government intervention – is a coherent concept. The second channel demonstrates that private paternalism may occur even if agents are naive. Moreover, the second channel explains why private paternalism is often shrouded. Under the second channel, firms will want to highlight employee satisfaction and gloss over the forcing mechanisms that make those employees productive and well-compensated.

7.1 Private paternalism vs. libertarian paternalism

From a linguistic perspective, private paternalism looks like liberterian paternalism (Thaler and Sunstein, 2003, 2008). However, unlike libertarian paternalism, private paternalism (as defined in this lecture) incorporates forcing mechanisms that are ex-post binding, although they are adopted on a voluntary ex-ante basis. Consider the model-free equilibrium (Subsection 6.1) and the specific illustrative example that I have been carrying through this lecture. Agents in period 3 would like to opt out of the contract that they accepted in period 2. They joined the firm in period 2 to work at effort level $e = 1$ for payment $w = 1$, but they would prefer to costlessly opt out (in period 3) and work at another firm at effort level $e = \beta$ for payment $w = \beta$. What prevents them from doing this? Some friction or cost that is large enough to temporarily bind them to their current employer: e.g., the cost of switching jobs. Hence, private paternalism is not libertarian, because agents would opt out (at stage 3) if the cost were small enough. It is this friction that keeps the equilibrium from unraveling and serves to bind them (unexpectedly, in this instance) to the mast.

8 The limits and virtues of private paternalism

Private paternalism has numerous limitations that prevent it from serving as a general counterweight to present bias. I review those limitations first. Then I conclude this section by describing the benefits of private paternalism and speculate about constructive ways of expanding its practical scope.

8.1 Limits of private paternalism

The model that I used to illustrate private paternalism makes many special simplifying assumptions. Some of these assumptions are necessary for the efficiency result that the

model generated.

First, I assumed that the agents' payoff functions in the formal sector are separable from those in the informal sector (home production). In general, this won't be true. For example, there are only a fixed number of hours in the day, so more hours spent working in the formal sector will crowd out hours spent in the informal sector. Effort may work the same way – an intense day at formal sector work may crowd out effort at home (even if hours at the two activities don't change). Dropping separability will break the clean result that workplace actions are socially efficient. Naturally, one can still solve for model-free equilibria in the absence of separability. A characterization of the non-separable case is an open problem.

Second, unobserved heterogeneity in present bias will break the social efficiency result – even in the formal sector – because screening is now necessary (e.g., O'Donoghue and Rabin 1999b, 2006, Galperti 2015, Lin 2017). Efficient screening will require repeated interactions (or historical data) if agents are naive about their present bias.

Third, the efficiency results will break down when benefits or costs occur at horizons that fall beyond the horizon of experienced utility measures. For example, coal miners may experience health problems that occur at long delays. Even mundane tasks, like delivering packages or working on an assembly line, may generate long-horizon, adverse consequences that are missed in short-horizon measures of experienced utility. Likewise, some employer-based benefits may have long-horizon consequences. Health investments, like a smoking cessation program, generate long-run benefits. Retirement benefits accrue for decades before they are paid out. Experienced utility needs to capture all of these long-run channels to generate social efficiency in the model-free equilibrium.

Fourth, even when the benefits and costs occur in the short-run, it is unlikely that experienced utility proxies will fully reflect all of the actual utility flows. The model assumed that agents have access to information sources that jointly provide an unbiased, comprehensive, and noiseless measure of experienced utility. But all of these properties are problematic. Glassdoor reviews are prone to selection bias and some employers have been accused of pressuring/paying employees to write positive reviews. Current employees may distort their reports to potential employees that are being recruited. Employers and employees may collude in providing information to prospective employees. Job panels (e.g., at career fairs) usually don't include disgruntled employees and supervisors typically monitor the reports of entry-level employees at these events. Marketing videos for corporate recruitment usually feature the most excited/positive current employees, or at least those who are willing to express such opinions. Finally, many subtle aspects of a job may be hard to capture in rankings

or ratings. Experienced utility may be particularly hard to gather in novel settings, where very few people have experience with the good/service being sold or job being offered. In these settings, agents may be more prone to rely on model-based forecasts (because they effectively have no experienced-based data on which to rely).

Fifth, the (costly) collection of data on experienced utility will also be inefficiently low. Because naive, present-biased agents overestimate the accuracy of their own model-based forecasts and under-estimate the relevance of other people's behavior as a proxy for their own behavior (Fedyk 2017), they will not fully realize the benefits of gathering experienced utility data. In addition, data on experienced utility is a public good, so it will be undersupplied even if agents did understand its value. For all of these reasons, experienced utility is not provided, though it would be inexpensive to collect in principle: e.g., gyms generally do not disclose the average frequency of attendance of their members, though they could report this data because many gyms now use swipe cards for entry (e.g., Dellavigna and Malmendier 2006). In the model that I have discussed, such information would be useful for naive, present-biased agents using model-free forecasting. However, those agents don't realize the value of this information and gyms do not have an incentive to provide it.

8.2 Home production

Private paternalism is particularly ineffective in the domain of home production. There is no 'firm' to provide enforcement. However, some domestic partnerships – e.g., marriages – operate in a way that generates an efficient forcing mechanism. For example, my spouse does an excellent job of forcing me to turn off my computer/phone/iPad at certain times of the day. Nevertheless, marriage is a crude tool for achieving the efficiency gains that I modeled in the formal sector. Partner selection isn't efficient for many different reasons. First, it is difficult to gather forecast-relevant, experienced-utility data on one's partner before settling down with them – would you ask your fiancée's ex-spouse for this information, and, if so, how would you interpret the report? By contrast, a firm has many current and ex-employees that can serve as credible informants. Even good partners aren't always present to act as a commitment mechanism – I go to McDonalds whenever my spouse isn't with me. Good partners aren't scalable, whereas a firm with a good commitment/forcing technology can scale its operations. In addition, choices of romantic partners are sometimes heavily influenced by considerations that have little bearing on long-run happiness. Finally, a large and growing fraction of the world's adults don't have partners at all.

8.3 Public paternalism

When private paternalism is likely to fail, there may be a case for public paternalism (i.e., paternalism initiated by some part of the government). As discussed above, private paternalism is unlikely to efficiently address institutional arrangements that have long-run consequences (e.g., retirement, health, and insurance). Experienced utility of a mid-life worker may fail to capture payoffs that occur later in life. Defined benefit or defined contribution pension plans generate payoffs that aren't viscerally or hedonically experienced until retirement.[41] Likewise, a firm's health investments in its workforce generates benefits that may take a lifetime to be fully realized. Experienced utility, which is captured by backward-looking memories, is unlikely to reflect these future payoffs. Hence, the model-free equilibrium that I discussed (Subsection 6.1) won't create efficient long-run investments. Accordingly, long-run investments present a natural case for public paternalism. Indeed, a substantial fraction of public paternalism is concentrated in such domains, including tax and regulatory incentives for retirement savings,[42] Social Security (forced retirement savings), and Medicare (forced savings during working life that funds health insurance in retirement).

Likewise, private paternalism is likely to be less relevant outside of formal sector employment relationships. Indeed, in the model-free equilibrium (Subsection 6.1), there were no forcing mechanisms in the home production sector. By contrast, in a firm, monitoring and forcing mechanisms (e.g., deadlines and progress reports) are natural features of the environment. In home production, monitors are often absent, and, even when some potential monitor is present, it is not clear that they have the necessary authority or appropriately aligned incentives. Hence, home production is less likely to be characterized by efficient private paternalism than formal-sector production. Accordingly, many types of public paternalism are targeted at behavior *outside* of firms: e.g., drug criminalization, cigarette and soda taxes, helmet and seatbelt laws, and, in most countries, mandatory (state-funded) unemployment insurance.

8.4 Virtues of private paternalism

I now discuss six benefits of private paternalism.

[41] This raises the thorny question of whether a placeholder for a future flow of utils – say a worker's rising account balance in a 401(k) – serves as an accurate forecasting proxy.

[42] For example, the tax treatment of IRA's and 401(k)'s, as well as regulatory incentives for firms to offer matching contributions in 401(k) plans.

First, private paternalism can enable agents to overcome present bias, even when they have naive beliefs. Private paternalism is optional at low frequencies (e.g., the choice of an employer), but binding at high frequencies (e.g., effort choices over the course of a work day), which is an ideal correction for agents with present bias.

Second, private paternalism can enable agents to overcome a host of biases *other* than present bias. The examples discussed in this paper (and in O'Donoghue and Rabin 1999b, 2006) focus on present bias, but model-free learning is a general-purpose technique for avoiding some types of misforecasting (and the costs that usually accompany misforecasting). For example, irrationally optimistic agents may be less easily exploited if they use model-free forecasts instead of using the overoptimistic structural theory that distorts their forecasts. "The average person loses money in a casino, so I shouldn't expect to do better than that (despite my structural theory that there exists an exploitable hot hand effect)."[43]

Third, private paternalism is not imposed by the government. Many scholars are skeptical of strict, government-controlled paternalism.[44] Private paternalism avoids further empowering a Leviathan, which may not have our best interests at heart or may not be nimble/smart enough to be optimally paternalistic.

Fourth, private paternalism allows for innovation and improvement, because it is generated and continuously reinvented by agents in the private sector and subject to competitive forces.

Fifth, private paternalism will appeal to both naive agents (if they use model-free forecasting) as well as sophisticated agents (whether they use model-free or model-based forecasting). Hence, private paternalism will be successful for a wide range of cognitive types.

Sixth, private paternalism will only succeed in gaining greater market share if it raises social welfare. Unlike government paternalism, which is imposed by fiat, private paternalism is choice-based and, if agents use (accurate) model-free forecasts, prone to maximize flows of experienced utility.

[43] This argument has its limits. Inferences about experienced utility must sometimes be made contingently to be useful. One wouldn't want to reason that the average kayaker on class V rapids doesn't get injured so I can kayak down class V rapids (despite my complete lack of experience as a kayaker).

[44] For example, see Thaler and Sunstein (2003, 2008), and Glaeser (2006).

9 Conclusion: the scope of private paternalism

In this lecture, I have characterized the scope of private paternalism in an economy with present-biased agents. In this setting, private paternalism arises through two channels: (i) agents who seek out commitment because they hold sophisticated beliefs about their present bias and use *model-based* forecasting, and (ii) agents (naive or sophisticated) who use *model-free* forecasting to choose firms that generate high experienced utility flows by preventing workers from making present-biased choices. When the second channel is active, and naive consumers are a significant part of the population, then private paternalism will be shrouded. Firms will deploy private paternalism, but they won't highlight it. This combination of private paternalism and shrouding resolves the commitment puzzle. Forcing mechanisms are commonplace in the labor market but nevertheless *shrouded* because they are being deployed to attract *naive* workers who like the resulting flows of high experienced utility, and don't recognize the paternalistic roots of that success.

I have argued that private paternalism arising from model-free learning will be more socially efficient in the following settings: (i) when payoffs are clearly tied to actions, so measures of experienced utility are more accurate; (ii) when payoffs aren't too delayed relative to the actions that generate those payoffs; (iii) when employment relationships are long-lived, so that long-run costs and benefits are integrated into available measures of experienced utility; (iv) and when productive activity occurs in the formal sector, where monitoring and forcing mechanims are easy to deploy, instead of the informal (home production) sector, where monitors are not always present and not always able or willing to implement socially efficient forcing mechanims.

This last point leads me to speculate that some of the difference between the quality of life in developing and developed economies lies in the relative role of the formal sector. In developing countries, the formal sector is relatively small, which implies that the scope for private paternalism is small. In this sense, the lack of private paternalism in the informal sector can serve as a poverty trap.

To gain intuition for this effect, think of the life of a typical university employee. Most universities provide health clinics, health insurance, retirement benefits, life insurance, disability insurance, medical leave, a structured work week, performance benchmarks, frequent feedback, deadlines, and many other mechanisms that keep us productive. A self-employed worker has little or none of these supporting structures. Accordingly, I believe that a substantial fraction of my productivity owes to the paternalistic environments in which I am

embedded, rather than my intrinsic productivity. I believe that if I moved to the self-employment sector, my productivity and life satisfaction would plummet, in part because my present bias would have free rein to distort my behavior.

Future research should empirically evaluate these conceptual conjectures. It should also identify the feasibility and consequences of expanding the scope of private paternalism. The role of private paternalism is enlarged (i) when we find ways to shift workers from the informal sector to the formal sector (or make the informal sector more structured so it can support private paternalism); (ii) when we make agents more sophisticated, so they will explicitly seek out commitment devices; and (iii) when we increase the quality of and access to data on model-free experienced utility (e.g., requiring that gyms report the exercise frequency of their members, or requiring that firms disclose standardized satisfaction surveys of their employees). The provision of data on experienced utility is a public good that could be subsidized or mandated.

References

Acland, Dan, and Matthew R. Levy. 2015. "Naiveté, Projection Bias, and Habit Formation in Gym Attendance." *Management Science*, 61(1): 146–60.

Akerlof, George A. 1991. "Procrastination and Obedience." *American Economic Review*, 81(2): 1–19.

Alsan, Marcella, John Beshears, Wendy S. Armstrong, James J. Choi, Brigitte C. Madrian, Minh Ly T. Nguyen, Carlos Del Rio, David Laibson, and Vincent C. Marconi. 2017. "A Commitment Contract to Achieve Virologic Suppression in Poorly Adherent Patients with HIV/AIDS." *AIDS*, 31(12): 1765–69.

Ariely, Dan, and Klaus Wertenbroch. 2002. "Procrastination, Deadlines, and Performance: Self-Control by Precommitment." *Psychological Science*, 13(3): 219–24.

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

Ashraf, Nava, Colin F. Camerer, and George Loewenstein. 2005. "Adam Smith, Behavioral Economist." *Journal of Economic Perspectives*, 19(3): 131–45.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *Quarterly Journal of Economics*, 121(2): 635–72.

Augenblick, Ned, and Matthew Rabin. Forthcoming. "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *Review of Economic Studies*.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working Over Time: Dynamic Inconsitency in Real Effort Tasks." *Quarterly Journal of Economics*, 130(3): 1067–115.

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

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and Jung Sakong. 2017. "Which Early Withdrawal Penalty - 10%, 20%, or ∞ - Attracts the Most Deposits to a Commitment Savings Account?" National Bureau of Economic Research Working Paper no. 21474.

Bisin, Alberto, and Kyle Hyndman. 2014. "Present-Bias, Procrastination and Deadlines in a Field Experiment." National Bureau of Economics Research Working Paper no. 19874.

Brown, George W. 1951. "Iterative Solutions of Games by Fictitious Play." In *Activity Analysis of Production and Allocation.* , ed. Tjalling C. Koopmans, 374–76. New York:Wiley.

Bryan, Gharad, Dean Karlan, and Scott Nelson. 2010. "Commitment Devices." *Annual Review of Economics*, 2(1): 671–98.

Bucks, Brian, and Karen Pence. 2008. "Do Borrowers Know Their Mortgage Terms?" *Journal of Urban Economics*, 64(2): 218–33.

Cadena, Ximena, Antoinette Schoar, Alexandra Cristea, and Héber Delgado-Medrano. 2011. "Fighting Procrastination in the Workplace: An Experiment." National Bureau of Economics Research Working Paper no. 16944.

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.

Choi, James J., David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2003. "Optimal Defaults." *American Economic Review*, 93(2): 180–5.

Cho, SungJin, and John Rust. 2017. "Precommitments for Financial Self-Control? Micro Evidence from the 2003 Korean Credit Crisis." *Journal of Political Economy*, 125(5): 1413–64.

Clark, Gregory. 1994. "Factory Discipline." *Journal of Economic History*, 54(1): 128–63.

Cohen, Jonathan D., Keith M. Ericson, David Laibson, and John M. White. 2016. "Measuring Time Preferences." National Bureau of Economics Research Working paper no. 22455.

Dayan, Peter, and Yael Niv. 2008. "Reinforcement Learning: The Good, The Bad and The Ugly." *Current Opinion in Neurobiology*, 18(2): 185–96.

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

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.

Eliaz, Kfir, and Ran Spiegler. 2004. "Contracting with Diversely Naive Agents." *Review of Economic Studies*, 73(22): 689–714.

Eliaz, Kfir, and Ran Spiegler. 2006. "Consumer Optimism and Price Discrimination." *Theoretical Economics*, 3(4): 459–97.

Ely, Jeffrey C., and Juuso Valimaki. 2003. "Bad Reputation." *Quarterly Journal of Economics*, 118(3): 785–814.

Fedyk, Anastassia. 2017. "Asymmetric Naiveté: Beliefs about Self-Control." Unpublished.

Fudenberg, Drew, and David K. Levine. 1998. *The Theory of Learning in Games.* Cambridge:MIT Press.

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. 2015. "Commitment, Flexibility, and Optimal Screening of Time Inconsistency." *Econometrica*, 83(4): 1425–65.

Gerardi, Kristopher, Lorenz Goette, and Stephan Meier. 2013. "Numerical Ability Predicts Mortgage Default." *Proceedings of the National Academy of Sciences*, 110(28): 11267–71.

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." *American Economic Journal: Applied Economics*, 2(4): 213–35.

Glaeser, Edward L. 2006. "Paternalism and Psychology." *The University of Chicago Law Review*, 73(1): 133–56.

Goda, Gopi Shah, Matthew R. Levy, Colleen F. Manchester, Aaron Sojourner, and Joshua Tasoff. 2015. "The Role of Time Preferences and Exponential-Growth Bias in Retirement Savings." National Bureau of Economics Research Working Paper no. 21482.

Gottlieb, Daniel. 2008. "Competition over Time-Inconsistent Consumers." *Journal of Public Economic Theory*, 10(4): 673–84.

Grubb, Michael D. 2009. "Selling to Overconfident Consumers." *American Economic Review*, 99(5): 1770–807.

Grubb, Michael D. 2015. "Consumer Inattention and Bill-Shock Regulation." *Review of Economic Studies*, 82(1): 219–57.

Grubb, Michael D., and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review*, 105(1): 234–71.

Gurun, Umit G., Gregor Matvos, and Amit Seru. 2016. "Advertising Expensive Mortgages." *Journal of Finance*, 71(5): 2371–416.

Halac, Marina, and Andrea Prat. 2016. "Managerial Attention and Worker Performance." *American Economic Review*, 106(10): 3104–32.

Hall, Robert E. 1997. "The Inkjet Aftermarket: An Economic Analysis." Unpublished.

Heidhues, Paul, and Botond Koszegi. 2010. "Exploiting Naivete about Self-Control in the Credit Market." *American Economic Review*, 100(5): 2279–303.

Heidhues, Paul, and Botond Koszegi. 2015. "On the Welfare Costs of Naiveté in the US Credit-Card Market." *Review of Industrial Organization*, 47(3): 341–54.

Heidhues, Paul, and Botond Koszegi. 2017. "Naïveté-Based Discrimination." *Quarterly Journal of Economics*, 132(2): 1019–54.

Holmstrom, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *Review of Economic Studies*, 66(1): 169–82.

Houser, Daniel, Daniel Schunk, Joachim Winter, and Erte Xiao. 2010. "Temptation and Commitment in the Laboratory." Unpublished.

Johnen, Johannes. 2017. "Dynamic Competition in Deceptive Markets." Unpublished.

Kahneman, Daniel, Barbara L. Fredrickson, Charles A. Schreiber, and Donald A. Redelmeier. 1993. "When More Pain Is Preferred to Less: Adding a Better End." *Psychological Science*, 4(6): 401–5.

Kanemoto, Yoshitsugu, and W. Bentley MacLeod. 1992. "Firm Reputation and Self-Enforcing Labor Contracts." *Journal of the Japanese and International Economies*, 6(2): 144–62.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2010. "Self-Control and the Development of Work Arrangements." *American Economic Review: Papers and Proceedings*, 100(2): 624–8.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. "Self-Control at Work." *Journal of Political Economy*, 123(6): 1227–77.

Kuchler, Theresa, and Michaela Pagel. 2017. "Sticking to Your Plan: The Role of Present Bias for Credit Card Paydown." Unpublished.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112(2): 443–77.

Laibson, David. 2015. "Why Don't Present-Biased Agents Make Commitments?" *American Economic Review*, 105(5): 267–72.

Laibson, David, and Leeat Yariv. 2006. "Safety in Markets: An Impossibility Theorem for Dutch Books." Unpublished.

Levin, Jonathan. 2003. "Relational Incentive Contracts." *American Economic Review*, 93(3): 835–57.

Levy, David T., Ron Borland, Eric N. Lindblom, Maciej L. Goniewicz, Rafael Meza, Theodore R. Holford, Zhe Yuan, Yuying Luo, Richard J. O'Connor, Raymond Niaura, and David B. Abrams. 2017. "Potential Deaths Averted in USA by Replacing Cigarettes with E-Cigarettes." *Tobacco Control*, 27: 18–25.

Lin, Hong. 2017. "Firm as a Commitment Device." Unpublished.

Loewenstein, George, and Drazen Prelec. 1992. "Anomalies in Intertemporal Choice: Evidence and an Interpretation." *Quarterly Journal of Economics*, 107(2): 573–97.

Malmendier, Ulrike, and Stefan Nagel. 2016. "Learning from Inflation Experiences." *Quarterly Journal of Economics*, 131(1): 53–87.

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. 2006. "Incentives and Self-Control." In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*. Vol. 2, , ed. Richard Blundell, Whitney K. Newey and Torsten Persson, Chapter 8, 215–45. Cambridge:Cambridge University Press.

Phelps, E. S., and R. A. Pollak. 1968. "On Second-Best National Saving and Game-Equilibrium Growth." *Review of Economic Studies*, 35(2): 185–99.

Royer, Heather, Mark F. Stehr, and Justin R. Sydnor. 2012. "Incentives, Commitments and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company." National Bureau of Economics Research Working Paper no. 18580.

Ru, Hong, and Antoinette Schoar. 2016. "Do Credit Card Companies Screen for Behavioral Biases?" National Bureau of Economics Research Working Paper no. 22360.

Schilbach, Frank. 2015. "Alcohol and Self-Control: A Field Experiment in India." Unpublished.

Shapiro, Fred R., ed. 2006. *The Yale Book of Quotations.* New Haven:Yale University Press.

Shui, Haiyan, and Lawrence M. Ausubel. 2005. "Time Inconsistency in the Credit Card Market." Unpublished.

Stango, Victor, and Jonathan Zinman. 2009. "What Do Consumers Really Pay on Their Checking and Credit Card Accounts? Explicit, Implicit, and Avoidable Costs." *American Economic Review: Papers and Proceedings*, 99(2): 424–9.

Strotz, Robert H. 1955. "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies*, 23(3): 165–80.

Thaler, Richard H., and Cass R. Sunstein. 2003. "Libertarian Paternalism." *American Economic Review*, 93(2): 175–9.

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

Warren, Patrick L., and Daniel H. Wood. 2014. "The Political Economy of Regulation in Markets with Naïve Consumers." *Journal of the European Economic Association*, 12(6): 1617–42.

Wertenbroch, Klaus. 1998. "Consumption Self-Control by Rationing Purchase Quantities of Virtue and Vice." *Marketing Science*, 17(4): 317–37.

Appendix: Solving for the exploitative equilibrium

Here I derive equilibrium for the case of a monopsony employer (Subsection 4.2). Following the discussion in the text, the firm's optimization problem is

$$\max_{e^*, w(e^*), \widehat{e}, w(\widehat{e})} e^* - w(e^*) \tag{9}$$

subject to the participation constraint (P) and incentive compatability constraint (IC):

$$-\frac{1}{2}\widehat{e}^2 + w(\widehat{e}) \geq z \tag{P}$$

$$-\frac{1}{2}\left(e^*\right)^2 + \beta w(e^*) \geq -\frac{1}{2}\widehat{e}^2 + \beta w(\widehat{e}). \tag{IC}$$

Note that the participation and incentive compatability constraints will both bind in equilibrium. If the participation constraint didn't bind, the firm could increase profit by lowering $w(\widehat{e})$ and $w(e^*)$, holding all equal. If the incentive compatibility constraint didn't bind, the firm could increase profit by lowering $w(e^*)$, holding all else equal. Henceforth, I will treat these as binding constraints. Rearrange the participation constraint to find

$$w(\widehat{e}) = \frac{1}{2}\widehat{e}^2 + z.$$

Now take this wage and substitute it into the incentive compatability constraint to find

$$\beta w(e^*) - \frac{1}{2}\left(e^*\right)^2 = \beta\left[\frac{1}{2}\widehat{e}^2 + z\right] - \frac{1}{2}\widehat{e}^2.$$

Rearranging this expression, I express the final wage, $w(e^*)$, as a function of e^* and $\widehat{e}$:

$$w(e^*) = \frac{1}{2\beta}\left(e^*\right)^2 + \left[\frac{1}{2}\widehat{e}^2 + z\right] - \frac{1}{2\beta}\widehat{e}^2.$$

Accordingly, I can rewrite the firm's objective as

$$\max_{e^*, \widehat{e}} e^* - \left\{\frac{1}{2\beta}\left(e^*\right)^2 + \left[\frac{1}{2}\widehat{e}^2 + z\right] - \frac{1}{2\beta}\widehat{e}^2\right\}.$$

This maximization generates two FOC's. The FOC for e^* is

$$1 - \frac{1}{\beta} e^* = 0,$$

which implies that $e^* = \beta$. The FOC for $\widehat{e}$ is

$$-\widehat{e} + \frac{1}{\beta}\widehat{e} = 0.$$

This first derivative with respec to $\widehat{e}$ is everywhere strictly positive because $\beta < 1$. So $\widehat{e}$ is set equal to its upper bound $\overline{e}$. So the anticipated wage is

$$w(\widehat{e}) = \frac{1}{2}\overline{e}^2 + z.$$

But the actual wage is

$$w(e^*) = \frac{1}{2\beta}\beta^2 + \left(\frac{1}{2}\overline{e}^2 + z\right) - \frac{1}{2\beta}\overline{e}^2.$$