Information Nudges and Self-Control*

Thomas Mariotti† Nikolaus Schweizer‡
Nora Szech§ Jonas von Wangenheim¶

November 2, 2018

Abstract

We study the optimal design of information nudges for present-biased consumers who have to make sequential consumption decisions without exact prior knowledge of their long-term consequences. For arbitrary distributions of risk, there exists a consumer-optimal information nudge that is of cutoff type, recommending consumption or abstinence according to the magnitude of the risk. Under a stronger bias for the present, the target group receiving a credible signal to abstain must be tightened. We compare this nudge with those favored by a health authority or a lobbyist. When some consumers are more strongly present-biased than others, a traffic-light nudge is optimal.

Keywords: Information Design, Information Nudges, Present-Biased Preferences, Self-Control.
JEL Classification: C73, D82.

*We thank Sandro Ambuehl, Kai Barron, Catherine Casamatta, Francesc Dilme, Laura Doval, Jannis Engel, Daniel Garrett, Bertrand Gobillard, Paul Heidhues, Yves Le Yaouanq, George Loewenstein, Stefano Lovo, Collin Raymond, Frank Rosar, Roland Strausz, Jean Tirole, and Takuro Yamashita for very valuable feedback. We also thank seminar audiences at Toulouse School of Economics and Universität Bonn, as well as conference participants at the 2018 Durham University Business School Conference on Mechanism and Institution Design, the 2018 EARIE Annual Conference, the 2018 EEA Annual Congress, the 2018 HeiKaMax Spring Workshop, the 2018 Verein für Socialpolitik Annual Conference, and the 2018 ZEW Workshop on Market Design for many useful discussions. Anke Greif-Winzrieth, Michelle Hörrmann, Nicola Hüholt, and Christine Knopf provided excellent research assistance. Jonas von Wangenheim acknowledges support by the German Research Foundation through CRC TRR 190.

†Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, CEPR, and CESifo. Email: thomas.mariotti@tse-fr.eu.

‡Tilburg School of Economics and Management, Tilburg University, Tilburg, The Netherlands. E-mail: n.f.f.schweizer@uvt.nl.

§Karlsruhe Institute of Technology (KIT), ECON Institute, Karlsruhe, Germany, Berlin Social Science Center (WZB), and CESifo. Email: nora.szech@kit.edu.

¶Freie Universität Berlin, Berlin, Germany. Email: jonas.wangenheim@fu-berlin.de.

1 Introduction

There has been a remarkable variety across space and time in attempts to alleviate the consumption of potentially harmful goods. A particularly drastic policy is to prohibit those goods altogether. This was done in the US in the 1920s with regard to alcohol. However, Prohibition did not prevent illegal consumption: data suggests that, while consumption first declined during Prohibition, it increased again after a few years, when the illegal market had adapted; consumption remained stable after Prohibition ended (Miron and Zwiebel (1991)). On top of being antiliberal and leading to the criminalization of many people, this extreme measure only achieved moderate results regarding drinking behavior (Hall (2010)). A similar case has more recently been made against drug prohibition (Miron and Zwiebel (1995)). The reason might be that prohibition does not credibly convey information about the actual hazards of consumption.

Nowadays, a more liberal and more informative approach is to use information nudges. For example, in many countries, cigarette packages now come with graphic information and text messages about the potential consequences of smoking. Consumers take those warnings as sources of information and react to such labels, at least to some extent (Hammond, Fong, McNeill, Borland, and Cummings (2005)). Similar findings have been reported regarding alcohol warning labels (MacKinnon, Pentz, and Stacy (1993)) and mandatory calorie posting in chain restaurants (Bollinger, Leslie, and Sorensen (2011)).

Yet empirical research also documents that consumers do not feel properly addressed. In a study with adolescents, McCool, Webb, Cameron, and Hoek (2012) report that many participants questioned whether the graphic labels "portrayed an authentic representation of the harm caused by smoking. Indeed, the majority perceived graphic warning labels as 'showing the worst case scenario' because, for example 'of course no-one's going to let their foot get that bad.'" A targeted and more credible information nudge may have considerably more potential. For example, warnings against drinking during pregnancy seem to have a significant impact on those concerned (Hankin, Firestone, Sloan, Ager, Goodman, Sokol, and Martier (1993)). Yet little is known about the optimal design of information nudges. This paper aims at filling this gap.

Our formal analysis relies on three ingredients: present-biased preferences, incomplete information, and Bayesian updating. Let us examine each of these ingredients in turn.

Present-Biased Preferences In our model, a decision maker has to make a sequence of consumption choices that may have harmful consequences in the future. The decision maker

is present-biased, in that he puts a disproportionate utility weight on the current period compared to all later periods in time (Ainslie (1975, 1992), Thaler (1981), Loewenstein and Prelec (1992)). In this context, his preferred course of action may look as follows: Cheat today, but abstain from tomorrow on. Under no commitment, however, this course of action is not feasible: once tomorrow is reached, the same logic applies so that cheating "today" combined with abstaining from "tomorrow" on looks most appealing—again! As a consequence, every day becomes a cheating day, and consumption never comes to an end. A decision maker aware of this misery may decide that quitting now is a smarter choice than engaging in harmful consumption forever. Yet this choice may never feel appealing enough. Thus, even though putting an end to harmful consumption now may dominate in terms of overall utility, consuming forever is the only feasible outcome in intrapersonal equilibrium, with possibly dreadful consequences.

Incomplete Information The decision maker has initially incomplete information about the harmful consequences of consumption. This can be because the likelihood of harmful consequences hinges on his individual risk type, which he need not know with precision. This could for instance arise if there is heterogeneity in risks across individuals; the decision maker then only has access to risk statistics at the aggregate population level, but does not know his exact position in this distribution, because it depends on a variety of risk factors he lacks the expertise to assess and combine. Alternatively, one could think of a population of decision makers facing an aggregate risk of unknown magnitude. In both interpretations, we will assume that a decision maker does not know the actual risk he is facing; yet we assume that the distribution of risks is common knowledge.

Bayesian Updating In this context, information nudges can help affecting a decision maker's incentives by modifying his information structure. Depending on the interpretation of risk adopted, such nudges can be designed at the individual level, as in a doctor/patient relationship, or at the population level, as in the case of tobacco or alcohol warnings. To avoid the negative effects of overstated consumption risks, we require that information nudges be credible.[1] We capture this requirement by assuming that the decision maker, when exposed to new information about the harmful consequences of consumption, updates his prior beliefs in a Bayesian way. This generates a tradeoff between the credibility of the nudge and its efficiency at deterring consumption whenever it is undesirable. Using tools

[1] Of course, other types of nudges deter consumption via emotional reactions such as disgust (Hammond, Fong, McDonald, Brown, and Cameron (2004)). We abstract from these different approaches in our analysis and focus on the impact of information.

from Bayesian persuasion (Brocas and Carrillo (2007), Rayo and Segal (2010), Kamenica and Gentzkow (2011)), we first characterize the optimal information structure from the decision maker's perspective prior to taking any consumption decision. We then compare this information structure with the optimal information structure from the perspective of a health authority aiming at minimizing the probability of consumption, as well as with the optimal information structure from the perspective of a lobbyist aiming at maximizing the probability of consumption.

We find that there always exists a decision-maker-optimal information structure that is of cutoff type. In the corresponding persuasion mechanism, the decision maker either learns that the risk he is facing is high or low, depending on whether it lies above or below a certain cutoff. The intuition is that cutoff mechanisms are good for efficiency purposes, because consumption eventually takes place only when the risk is low enough, and that they also have good incentive properties, because, under no commitment, abstention is incentive-compatible only when the risk is perceived by the decision maker to be high enough. When there is heterogeneity in individual risk types, these signals can be interpreted as recommendations warning against consumption for high-risk individuals within the target group of the information nudge.[2] When the risk is an aggregate one, credible information about the hazards of consumption is conveyed to the whole population. In either case, finding the optimal information structure is easy in the sense that it requires pinning down one single parameter. What makes it challenging is that the degree of self-control is crucial for the optimal design of the information structure.

The optimal cutoff structure outperforms perfect transparency because, via pooling, more types may actually find the strength to abstain from consuming once they have learned that they are of high risk. This contrasts with a decision maker with no bias for the present, for whom perfect transparency would be optimal (Blackwell (1953)). By tightening the target group, more drastic information can be credibly communicated, thereby counteracting impulses from the decision maker's bias for the present. Indeed, such tightening may explain why warnings against alcohol work best when they are targeted at the most vulnerable groups, such as pregnant women. Of course, in practice, many more types should better abstain (Gutjahr, Gmel, and Rehm (2001), Shield, Parry, and Rehm (2014)). Yet our analysis suggests that it may be optimal to warn only high-risk types in order to deter at least them successfully, sacrificing types of lower but still significant risk who will be trapped

[2]If the distribution of types has atoms, we find that the recommendation to the marginal risk type may involve randomization.

in harmful consumption. Key to this logic is that, except possibly for marginal types, the optimal information structure is coarse: it is more efficient to shield the maximum mass of types away from consumption by issuing a straight recommendation to abstain, rather than to issue mixed messages that would only partially protect inframarginal types.

The cutoff structure of optimal incentive-compatible persuasion mechanisms carries over when one changes the objective function to a health authority's or a lobbyist's. Yet, of course, the cutoffs are chosen differently. While the health authority prefers to make as many consumers as possible shy away from harmful consumption, the lobbyist prefers to lower willpower in as many consumers as possible by convincing them that the risk is not that high. In the latter case, many consumers who would wish for an information nudge that helped them abstaining, are instead trapped in harmful consumption. Naturally, the lobbyist chooses the cutoff of the persuasion mechanism such that as few consumers as possible receive a convincing signal to abstain, while the health authority does exactly the opposite. A policy maker who would not take consumers' self-control problems into consideration may even misinterpret information structures implemented by a lobbyist as health-concerned, when, indeed, the target group for a warning label may be chosen deliberately broadly in order to reduce the impact of the nudge.

In all three optimization problems, the possibilities for designing effective persuasion mechanisms are limited by incentive constraints. Thus the signal of being a high risk in the target group must be alarming enough to induce the decision maker to abstain. From a liberal perspective, it would be ideal to choose the corresponding cutoff in the persuasion mechanism in such a way that consumption is recommended if and only if it involves no harm. Yet, under weak conditions on the distribution of risks, this mechanism is incentive-compatible if and only if the decision maker's bias for the present is low enough. In all other cases, harmful consumption takes place with positive probability in equilibrium, and the decision-maker-optimal information nudge coincides with the health-authority-optimal one. We also show that a positive shift of the distribution of types in the hazard-rate order reduces the probability of consuming, reflecting that it is both more desirable and easier for the mechanism designer to discourage consumption. Interestingly, while a more severe bias for the present intuitively increases the probability that consumption takes place, it need not increase the probability that harmful consumption takes place.

Levels of self-control vary across consumers (Mischel (2014), Sutter, Kocher, Glätzle-Rützler, and Trautmann (2013)). For example, consumers with high self-control differ from consumers with low self-control when it comes to food choice, as has been shown in a study on

the potential impact of product labeling on health (Koenigstorfer, Groeppel-Klein, Kamm (2014)). Therefore, in an extension, we analyze the case in which decision makers may have high or low self-control. We find that in many cases, both types of decision makers can be optimally informed via the same information structure, which turns out to take the form of a traffic-light nudge. While the strongest, "red" warning signal is drastic enough to make both decision makers with high and with low self-control abstain, the intermediate "yellow" warning convinces at least decision makers with high self-control to end harmful consumption. This discrimination property may be a reason why traffic-light nudges are one if not the most frequently used nonnumerical information structure, in addition to their potential saliency.[3]

Related Literature

Our paper lies at the intersection of three strands of literature. First, our work is related to the literature on present-biased preferences and information acquisition pioneered by Carrillo and Mariotti (2000) and Bénabou and Tirole (2002); specifically, we take the basic model of Carrillo and Mariotti (2000) as our starting point. However, while the literature has so far emphasized situations in which present-biased decision makers manipulate the information-acquisition or the information-storing processes, we characterize information structures that are optimal, given a decision maker's bias for the present, from different perspectives. While the basic insight remains that gathering no information may outperform full transparency, our analysis demonstrates that an intermediate information structure is best, and can be interpreted as an information nudge acting as a credible warning signal to a specific target group. This solves two problems that may appear when the task of gathering information is performed by the decision maker himself. The first is the multiplicity of equilibria arising from the difficulty to coordinate one's selves on an intrapersonal equilibrium. The second is that gathering information oneself creates an additional risk by making different pieces of information available only sequentially; as a result, some types may end up trapped in harmful consumption, whereas they would completely abstain if they were instead exposed to the coarser decision-maker-optimal information nudge.

Second, the information-design problem we study connects our paper to the recent and very active literature on Bayesian persuasion initiated by the seminal papers of Brocas and Carrillo (2007), Rayo and Segal (2010), and Kamenica and Gentzkow (2011); see Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017) for a recent contribution with many references.

[3]Evidence on the latter is mixed, see VanEpps, Downs, and Loewenstein (2016).

What sets our paper apart from most of this literature is that our focus is on frictions in information demand that arise from intrapersonal, psychological conflicts rather than from sender-receiver conflicts of interest. Depending on the mechanism designer's objective, the optimal persuasion mechanism varies drastically. Our paper thereby contributes to a small but growing literature on the optimal disclosure of information to agents with psychological preferences. Lipnowski and Mathevet (2018) show that a tempted agent in the sense of Gul and Pesendorfer (2001) does not want to know what he is missing, and thus an optimal disclosure mechanism should limit his information about the value of his preferred choice, so as to reduce the cost of self-control. Schweizer and Szech (2018) study the optimal revelation of life-changing information, such as that provided by a medical test, to a patient with anticipatory utility. Closer to Bénabou and Tirole (2002), Habibi (2017) studies feedback mechanisms in a setting where a benevolent principal motivates an agent with present-biased preferences to exert unobservable effort by providing him with feedback. Different from the optimal persuasion mechanisms that we construct, the feedback mechanisms studied by Habibi (2017) are based on a noisy signal that depends on both the agent's type and effort, thus providing a moral-hazard counterpart to our analysis.

Popularized by Thaler and Sunstein (2008), the literature on nudging is growing fast and into multiple directions, with remarkable success also on a political level. Research on nudging has informed policy making in various countries, such as in the US, UK, Australia, Germany, and Japan. Also the UN, the OECD, and the World Bank have set up nudging units. While contributions such as Benkert and Netzer (2018) focus on nudging in the sense of influencing the framing of decision problems, our focus is on nudges in the form of an optimized release of information, so called information nudges.[4] Such nudges in the form of warning signals or labels have already received much attention in previous decades, notably in the marketing literature; see Argo and Main (2004) for an overview. We address the design of credible information nudges for populations of heterogeneous decision makers who are present-biased, and compare optimal information nudges from different policy perspectives. While the optimal nudge can always be represented as a warning signal to a target group, the size of the target group and the according signal can vary drastically according to the political goal. Policy makers unaware of or underestimating consumers' self-control problems risk to implement an information nudge that completely misses its goal. It may even maximize consumption when minimizing consumption is intended.

[4]Coffman, Featherstone, and Kessler (2015) study information nudges assuming agents have mean-variance preferences. They focus on the comparative statics of agents' decisions in reaction to different nudges. In contrast, our focus is on characterizing optimal information nudges.

The paper is organized as follows. Section 2 describes the model. Section 3 characterizes optimal information disclosure. Section 4 illustrates our findings in the polar cases of binary and nonatomic distributions of types. Section 5 considers alternative objective functions. Section 6 studies the case of a mixed population of agents, in which some suffer from more severe self-control problems than others. Section 7 concludes. Proofs not given in the text can be found in the Online Appendix.

2 The Model

As in Carrillo and Mariotti (2000), we focus on a time-inconsistent decision maker (he) who makes sequential consumption decisions under no commitment. Consumption is enjoyable in the short term but, depending on the decision maker's type, may be quite harmful in the long term. The novelty of the model is that the decision maker's information about his type is optimized by a mechanism designer (she).

2.1 Actions and Payoffs

The decision maker lives at dates 0, 1, 2, and 3. At dates $\tau = 0, 1$, he can consume, $x_\tau = 1$, or abstain, $x_\tau = 0$. Consuming at any date τ increases current utility by 1 but comes with probability θ at a cost C, incurred at date $\tau + 2$. Following Phelps and Pollak (1968) and Laibson (1997), the decision maker discounts future payoffs according to a quasi-hyperbolic discount function with parameters β and δ. That is, his vNM utility functions at dates 0 and 1 are given by

$$U_0(x_0, x_1, \theta) = x_0(1 - \beta\delta^2\theta C) + x_1\beta\delta(1 - \delta^2\theta C), \tag{1}$$

$$U_1(x_0, x_1, \theta) = -x_0\beta\delta\theta C + x_1(1 - \beta\delta^2\theta C), \tag{2}$$

where $\beta \in (0, 1)$ is the time-inconsistency parameter capturing the bias for the present relative to the future, while $\delta \in (0, 1]$ is the usual per-period discount factor. As $\beta < 1$, the decision maker at date 1 puts, relatively to his utility from consuming, less weight on the harm his consuming might cause at date 3 than he does at date 0. We assume that $\beta\delta^2 C > 1$, so that the decision maker would always abstain if he believed that the cost C were incurred with probability 1 upon consuming.

2.2 Information and Strategies

The prior beliefs of the decision maker about θ are represented by a distribution $\mathbf{P}$ with cumulative distribution function F over $[0, 1]$. We denote by $\underline{\theta}$ and $\overline{\theta}$ the infimum and the

supremum of the support Θ of $\mathbf{P}$, respectively.

Before making his first consumption decision at date 0, the decision maker is exposed to additional information about θ. This information is distilled by a mechanism designer who knows the value of θ and can commit to a persuasion mechanism issuing messages conditional on that value. The decision maker then updates his beliefs about θ in a Bayesian way whenever that is possible.

As in Strotz (1956), however, the decision maker is unable to commit to a course of action contingent on his updated beliefs. This restriction is binding, because the preferences induced by (1)–(2) along with these beliefs are time-inconsistent as $\beta < 1$. Following Peleg and Yaari (1973), the date-0 and date-1 selves of the decision maker act as independent decision units. The decision maker is sophisticated, so that his behavior is described by a subgame-perfect equilibrium of the resulting intrapersonal game.

In most of our analysis, the mechanism designer is benevolent in that her interests are aligned with those of the decision maker at date 0. Alternative objective functions for the mechanism designer are considered in Section 5.

2.3 Applications

Our model applies to situations in which a mechanism designer can determine how much information she wants to give out regarding a risk type θ. She can pool information by issuing a coarse signal. Yet she need to stick to the truth: that is, she cannot fool Bayesian decision makers by systematically lying to them. Depending on the application, the risk type may be that of a product a decision maker can choose, a characteristic of the decision maker himself, or a combination of the two.

In the first case, information structures are typically identical for a whole population. Think, for example, of information nudges on food and beverages in a supermarket, indicating how healthy a specific choice would be. If the information nudge is printed on the item itself, the mechanism designer decides if she wants to disclose the risk type θ of a product, or if she prefers to pool information about different products. For example, she could decide whether a snack is labeled as a healthy, green-light item or as an unhealthy, red-light item. More detailed information can be provided by a traffic-light nudge.

In the second case, the mechanism designer may be able to individually address different consumers, and thereby make use of more personalized signals. An example is information nudging in a supermarket via smart glasses or smartphones. Another case in point is medical advice. A doctor or a medical agency may have superior information about a patient's risk

type, and optimize the way it is communicated in order to affect his behavior.[5] In the latter case, the risk type θ is an individual characteristic of the patient. The doctor can disclose the patient's risk type perfectly, but she could also tell him that he belongs to a group of smaller or larger risk.

A key point in that respect is that, even if a decision maker has some private information, he may lack the ability to translate it into his individual risk type θ. This is essentially equivalent to having no private information at all, and, hence, room for information design opens up. For example, consider a decision maker deciding between consuming now or saving towards retirement. The probability θ then corresponds to his individual survival probability. Assume that initially, the decision maker only has access to survival probabilities at the aggregate population level. Then, although he may possess some information about his age, socioeconomic status, health and other factors, he need not know how to combine these factors to compute his individual survival probability.[6] The mechanism designer has access to the relevant computation model and can offer personalized information to the decision maker. Again, she may decide to pool risk types.

2.4 The Intrapersonal Game

As a preliminary step, we focus on the intrapersonal game played by the decision maker's date-0 and date-1 selves following the issue of some message by the mechanism designer. Owing to the binary character of consumption decisions and to the linearity of utilities in θ, equilibrium behavior in this intrapersonal game only depends on the decision maker's mean posterior belief $\hat{\boldsymbol{\theta}}$ about θ following this message. Letting

$$t^a \equiv \frac{1}{\beta\delta^2 C} \in (0,1), \tag{3}$$

our first result is a direct consequence of (1)–(2).

Lemma 1 *Given a mean posterior belief $\hat{\boldsymbol{\theta}}$ about θ, the intrapersonal game has a unique efficient subgame-perfect equilibrium, in which the decision maker's date-0 and date-1 selves both consume if $\hat{\boldsymbol{\theta}} < t^a$ and both abstain if $\hat{\boldsymbol{\theta}} \geq t^a$.*

Observe from (1) that, if $\beta t^a < \hat{\boldsymbol{\theta}} < t^a$, then the decision maker at date 0 would be strictly better off consuming at date 0 and abstaining at date 1. However, there is no way

[5] For economic studies in this context, see, for instance, Caplin and Leahy (2004), Köszegi (2003), and Schweizer and Szech (2018).

[6] As a stark example, Hurwitz and Sade (2017) find that, compared to nonsmokers, smokers more rarely prefer the lump-sum option when life insurance money is paid out; actually, they do not think that they have a shorter life expectancy than nonsmokers either.

he can reach this outcome under no commitment. Notice also that there is a discontinuity in the decision maker's date-0 equilibrium payoff at $\hat{\boldsymbol{\theta}} = t^a$. Indeed, letting

$$t^h \equiv \frac{1+\beta\delta}{1+\delta}\, t^a \in (0, t^a), \tag{4}$$

if $t^h < \hat{\boldsymbol{\theta}} < t^a$, then the decision maker at date 0 would be strictly better off abstaining at both dates than consuming at both dates, and the more so, the closer $\hat{\boldsymbol{\theta}}$ is to t^a. Yet, under no commitment, he cannot help doing so; we then say that *harmful consumption* takes place in equilibrium. The resulting discontinuity in the decision maker's date-0 equilibrium payoff arises from his bias for the present: in the limiting case $\beta = 1$, the gap between t^h and t^a vanishes, and the decision maker's date-0 equilibrium payoff is continuous in $\hat{\boldsymbol{\theta}}$; indeed, his payoff is then convex in $\hat{\boldsymbol{\theta}}$, reflecting that the value of information for a time-consistent decision maker is always nonnegative.

The equilibrium outcome described in Lemma 1 is unique if $\hat{\boldsymbol{\theta}} \neq t^a$. If $\hat{\boldsymbol{\theta}} = t^a$, then both the date-0 and the date-1 selves are indifferent between consuming and abstaining, whereas the date-0 self strictly prefers that the date-1 self abstains, and reciprocally. Because the date-0 self can do nothing to influence the behavior of the date-1 self, and reciprocally, there is a continuum of subgame-perfect equilibria in which both the date-0 and the date-1 self abstain with arbitrary probabilities in $[0, 1]$; yet, according to (1)–(2), the efficient equilibrium arises when they both abstain with probability 1. We focus on this equilibrium for three reasons.

1. First, as our goal is to characterize the best persuasion mechanism from the perspective of the decision maker at date 0, it is natural to select the continuation equilibrium that maximizes the payoff of the date-0 self, leaving the date-1 self indifferent.

2. Second, and more subtly, mean posterior beliefs $\hat{\boldsymbol{\theta}}$ equal to the cutoff t^a will play a key role in our analysis, and it is crucial for the existence of optimal persuasion mechanisms that the continuation equilibrium given such beliefs be efficient.

3. Third, no matter the selected continuation equilibrium, there exists for each $\varepsilon > 0$ an ε-optimal persuasion mechanism that induces posterior beliefs such that the above tie-breaking issue never arises.

For these reasons, we disregard equilibria of the intrapersonal game other than the efficient one and proceed as if the decision maker's behavior given any mean posterior belief about θ were uniquely determined. Figure 1 below illustrates the resulting decision maker's date-0 equilibrium payoff as a function of $\hat{\boldsymbol{\theta}}$.

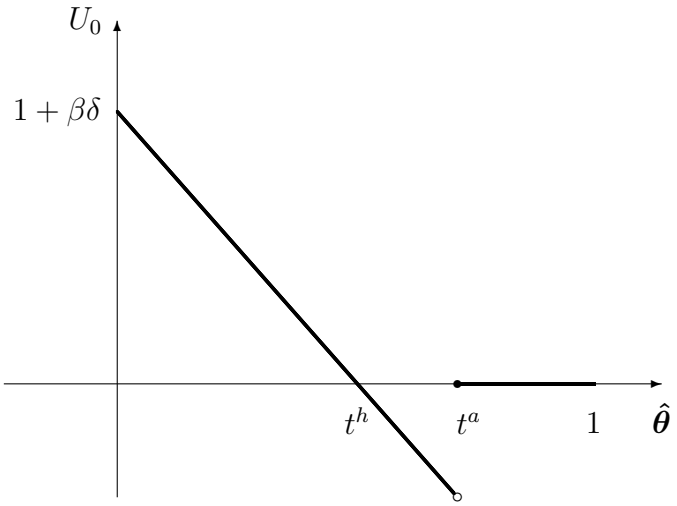

Figure 1: The decision maker's date-0 equilibrium payoff.

3 Optimal Information Disclosure

3.1 Suboptimality of Full Information Revelation

If the decision maker had no bias for the present or could commit to a course of action, full information would be optimal from the perspective of the decision maker at date 0 (Blackwell (1953)). As shown by Carrillo and Mariotti (2000), however, this is no longer the case if he suffers from a self-control problem. Intuitively, this follows from the nonconvexity of his equilibrium payoff as a function of his mean posterior belief, as illustrated in Figure 1. To see this point more formally, suppose

$$\mathbf{E}[\theta] \geq t^a \quad \text{and} \quad t^h < \mathbf{E}[\theta \,|\, \theta < t^a] < t^a.$$

The first inequality implies that, if the decision maker stayed with his prior, then he would abstain at both dates and thus obtain a zero payoff. Together with (1) and (3)–(4), the second inequality implies

$$\mathbf{E}[U_0(1,1,\theta) \,|\, \theta < t^a] < 0.$$

Hence, if the decision maker were to learn that $\theta < t^a$, then he would on average derive a negative payoff from consuming at both dates, an outcome which, according to Lemma 1, he could not prevent from happening under no commitment. Because learning that $\theta \geq t^a$ would in any case not affect his behavior relative to his prior, the decision maker thus strictly

prefers to stay with his prior and abstain at both dates, rather than learning the value of θ and possibly getting trapped in harmful consumption. Thus full transparency can destroy beneficial beliefs that help him overcome temptation.

3.2 Persuasion Mechanisms

The above argument shows that the value of becoming perfectly informed relative to staying ignorant can be negative from the perspective of the decision maker at date 0. However, this comparison is extreme, and does not shed light on the date-0 optimal information structure. We now tackle this issue, building on the Bayesian-persuasion models of Brocas and Carrillo (2007), Rayo and Segal (2010), and Kamenica and Gentzkow (2011).

Following Aumann (1964), there is no loss of generality in focusing on measurable direct persuasion mechanisms $x : \Theta \times \Omega \to \{0, 1\}$ issuing, for any type $\theta \in \Theta$ and for any element ω of some sample space Ω, a recommendation $x(\theta, \omega)$ to abstain (0) or to consume (1) at dates 0 and 1.[7] As in Aumann (1964), we can take Ω to be $[0, 1]$, endowed with Lebesgue measure $\boldsymbol{\lambda}$ over the Borel sets. To any measurable direct persuasion mechanism $x : \Theta \times \Omega \to \{0, 1\}$ corresponds a measurable mapping $\pi : \Theta \to [0, 1]$ that associates to each $\theta \in \Theta$ a probability

$$\pi(\theta) = \boldsymbol{\lambda}[\{\omega \in \Omega : x(\theta, \omega) = 1\}] \tag{5}$$

of issuing a recommendation to consume at dates 0 and 1. Conversely, it follows from Aumann (1964, Lemma F) that, for any measurable mapping $\pi : \Theta \to [0, 1]$, there exists a measurable direct persuasion mechanism $x : \Theta \times \Omega \to \{0, 1\}$ such that (5) holds for all $\theta \in \Theta$. In line with Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017), we will mostly work with this equivalent and more convenient probabilistic representation of persuasion mechanisms, but we will occasionally rely on the original formulation.

3.3 Incentive Compatibility and Optimality

An important difference between our setting and standard models of Bayesian persuasion is that the decision maker cannot implement an optimal course of action conditional on his information; rather, his behavior results from a subgame-perfect equilibrium of the game played by his date-0 and date-1 selves. In particular, abstaining is not a default option for the decision maker, because there are information states in which he would be strictly better off abstaining but cannot help consuming. As a result, we cannot write his incentive-compatibility constraints in the usual way.

[7] No other recommendation would be followed by the decision maker, because his equilibrium behavior in any information state is uniquely determined, in the sense explained in Section 2.4.

As the decision maker has no private information, we just need to focus on his willingness to comply with the recommendations made to him by the mechanism designer. Consider first a mechanism π under which both the recommendations to consume and to abstain are sent with positive probability, which allows for straightforward applications of Bayes' rule. By Lemma 1, complying with the recommendation to consume is consistent with a continuation equilibrium if and only if $\mathbf{E}[\theta \,|\, x(\theta,\omega)=1] < t^a$, that is,

$$\frac{\mathbf{E}[\theta\pi(\theta)]}{\mathbf{E}[\pi(\theta)]} < t^a. \tag{6}$$

Similarly, complying with the recommendation to abstain is consistent with a continuation equilibrium if and only if $\mathbf{E}[\theta \,|\, x(\theta,\omega)=0] \geq t^a$, that is,

$$\frac{\mathbf{E}[\theta[1-\pi(\theta)]]}{\mathbf{E}[1-\pi(\theta)]} \geq t^a. \tag{7}$$

More generally, the left-hand side of constraint (6) is not well defined if $\pi = 0$ $\mathbf{P}$-almost surely, and similarly the left-hand side of constraint (7) is not well defined if $\pi = 1$ $\mathbf{P}$-almost surely. We adopt the convention that the undefined constraint is then emptily satisfied. A mechanism π is *incentive-compatible* if it satisfies (6)–(7).

Given the expression (1) for $U_0(1,1,\theta)$, the optimal-design problem can then, up to a multiplicative constant $(1+\delta)/t^a$, be formulated as

$$\max\{t^h\mathbf{E}[\pi(\theta)] - \mathbf{E}[\theta\pi(\theta)] : \pi \text{ is incentive-compatible}\}. \tag{8}$$

Observe that the objective function in (8) as well as the constraints (6)–(7) are all linear in π, which greatly simplifies the analysis. It is worth noticing that there is something slightly unusual about problem (8), namely, that the inequality in (6) is strict. This, again, reflects the fact that the behavior of the decision maker results from the equilibrium of a game, and not from a standard optimization problem.

3.4 A Characterization of Optimal Persuasion Mechanisms

We now characterize optimal persuasion mechanisms. This requires very little structure on the decision maker's prior beliefs: the distribution $\mathbf{P}$ may be discrete, continuous, or mixed. The only restriction we impose is that the support Θ of $\mathbf{P}$ be sufficiently spread out.

Assumption 1 $\mathbf{P}[\theta \leq t^h] > 0$ *and* $\mathbf{P}[\theta > t^a] > 0$.

Because $t^h < t^a$, this, in particular, implies $\underline{\theta} < t^a < \overline{\theta}$.

3.4.1 Cutoff Mechanisms

A *cutoff mechanism* recommends consuming (abstaining) if θ is below (above) a cutoff, with possible randomization between the two recommendations at the cutoff. Formally,

$$\pi(\theta) = 1_{\{\theta<t\}} + a1_{\{\theta=t\}}$$

for some pair $(t, a) \in \Theta \times [0, 1]$. The first step of our characterization consists in showing that there is no loss of generality in focusing on such mechanisms. The intuition is that cutoff mechanisms are good for efficiency purposes, because they recommend consuming for values of θ such that consumption is the most valued by the decision maker, and that they also have good incentive properties, because they recommend abstaining when the news about θ is the most alarming.

To see why, we first restrict attention to the mechanisms π that recommend consuming with some given probability $\gamma \in [0, 1]$, that is, $\mathbf{E}[\pi(\theta)] = \gamma$. Let us momentarily abstract from incentive considerations and consider, among these mechanisms, one that maximizes the objective in (8) or, equivalently, that solves

$$\min \{\mathbf{E}[\theta\pi(\theta)] : \mathbf{E}[\pi(\theta)] = \gamma\}. \tag{9}$$

That is, subject to the constraint that consuming be recommended with probability γ, we want to find a mechanism that minimizes the expected harm from consumption. Given this objective, it is optimal to concentrate the mass γ of consumption recommendations on small values of θ. If the distribution $\mathbf{P}$ is nonatomic, then a solution to (9) takes the value 1 in an interval starting at zero until enough probability mass has accumulated, that is, until the γ-quantile of F is reached, after which it takes the value 0. If $\mathbf{P}$ has atoms, then the γ-quantile of F may well lie within an atom; in that case, a solution to (9) may necessitate randomization, but only at this atom. Formally, the following result holds.

Lemma 2 *The unique solution to (9) is, up to a* $\mathbf{P}$*-null set, the cutoff mechanism*

$$\pi^*_\gamma(\theta) = 1_{\{\theta<t_\gamma\}} + \frac{\gamma - F(t_\gamma^-)}{F(t_\gamma) - F(t_\gamma^-)} 1_{\{\theta=t_\gamma\}} \tag{10}$$

for $t_\gamma \equiv \inf\{\theta : F(\theta) > \gamma\}$*, with* $\frac{0}{0} = 1$ *and* $\inf \emptyset = \infty$ *by convention.*

If the distribution $\mathbf{P}$ is nonatomic, then the second term on the right-hand side of (10) is irrelevant. Conversely, if this term is positive, then t_γ is an atom of $\mathbf{P}$ and the mechanism π^*_γ involves randomization at t_γ unless $\gamma \in \{F(t_\gamma^-), F(t_\gamma)\}$.

Returning to incentive considerations, observe that if some mechanism π such that $\mathbf{E}[\pi(\theta)] = \gamma$ is incentive-compatible, then so is π^*_γ. This is clear if $\gamma \in (0,1)$, for π^*_γ minimizes the left-hand side of (6) and maximizes the left-hand side of (7) among the mechanisms π such that $\mathbf{E}[\pi(\theta)] = \gamma$. This is also trivially true if $\gamma \in \{0,1\}$, for then $\pi = \pi^*_\gamma$ up to a $\mathbf{P}$-null set. Because, by Lemma 2, π^*_γ uniquely minimizes the expected harm from consumption among the mechanisms π such that $\mathbf{E}[\pi(\theta)] = \gamma$, this shows that we can confine ourselves to the class of incentive-compatible cutoff mechanisms.

From a practical viewpoint, what it needs in order to implement a given cutoff mechanism is the identification of the target group that will receive an effective warning to abstain. Our analysis shows that the target group will always be coherent. We now determine the optimal target group for an effective information nudge.

3.4.2 Optimization

The cutoff mechanism π^*_γ is incentive-compatible if

$$\frac{\mathbf{E}[\theta\pi^*_\gamma(\theta)]}{\mathbf{E}[\pi^*_\gamma(\theta)]} < t^a, \tag{11}$$

$$\frac{\mathbf{E}[\theta[1-\pi^*_\gamma(\theta)]]}{\mathbf{E}[1-\pi^*_\gamma(\theta)]} \geq t^a, \tag{12}$$

with the same convention as for (6)–(7) if $\gamma \in \{0,1\}$. The optimal-design problem (8) can then be restated as

$$\max\{t^h\mathbf{E}[\pi^*_\gamma(\theta)] - \mathbf{E}[\theta\pi^*_\gamma(\theta)] : \gamma \text{ satisfies (11)–(12)}\}. \tag{13}$$

We now provide an explicit characterization of optimal persuasion mechanisms, proving in particular that there always exists a solution to (13).

The Unconstrained-Optimal Mechanism To characterize the solution to (13), let us again momentarily abstract from incentive considerations and consider, among all values of $\gamma \in [0,1]$, the largest one that maximizes the objective in (13). The corresponding unconstrained-optimal mechanism is the indicator function of the range where the net benefit $t^h - \theta$ from consuming is nonnegative,

$$\pi^*_{\gamma^u}(\theta) = 1_{\{\theta \leq t^h\}},$$

so that $\gamma^u = F(t^h) > 0$. In particular, $\pi^*_{\gamma^u}$ does not involve randomization.[8]

[8] If $\mathbf{P}[\theta = t^h] > 0$, there exists a continuum of unconstrained-optimal cutoff mechanisms indexed by $\gamma \in [F(t^{h-}), F(t^h)]$. We choose the largest one because it is most likely to satisfy constraint (12).

If $\pi^*_{\gamma^u}$ satisfies (11)–(12), then it is also incentive-compatible and, therefore, solves the initial optimal-design problem (8). Under Assumption 1, this amounts to

$$\mathbf{E}[\theta \,|\, \theta \leq t^h] < t^a, \tag{14}$$

$$\mathbf{E}[\theta \,|\, \theta > t^h] \geq t^a. \tag{15}$$

Because $t^h < t^a$, (14) is automatically satisfied. Hence the following result holds.

Proposition 1 *If (15) holds, then the optimal incentive-compatible persuasion mechanism is the unconstrained-optimal mechanism* $\pi^*_{\gamma^u}$.

Harmful Consumption If (15) does not hold, then the unconstrained-optimal mechanism $\pi^*_{\gamma^u}$ is no longer incentive-compatible. This implies that harmful consumption can no longer be avoided for all values of $\theta > t^h$ or, equivalently, that consuming is optimally recommended with probability $\gamma^c > \gamma^u$ in the constrained-optimal mechanism. Because the net benefit $t^h - \theta$ from consuming only switches sign once, the objective in (13) is first nondecreasing and then nonincreasing in γ. It is then optimal to have γ^c as close as possible to γ^u, while preserving (12). The following result thus holds.

Proposition 2 *If (15) does not hold, then the optimal incentive-compatible persuasion mechanism is* $\pi^*_{\gamma^c}$, *where*

$$\frac{\mathbf{E}[\theta[1 - \pi^*_{\gamma^c}(\theta)]]}{\mathbf{E}[1 - \pi^*_{\gamma^c}(\theta)]} = t^a \tag{16}$$

implicitly defines the probability γ^c *of consuming.*

The proof of Proposition 2 relies on two observations. First, when γ increases from 0 to 1, the left-hand side of (12) strictly and continuously increases from $\mathbf{E}[\theta]$ to $\overline{\theta}$. Now, $\mathbf{E}[\theta] < \mathbf{E}[\theta \,|\, \theta > t^h] < t^a$ if (15) does not hold, while $\overline{\theta} > t^a$ under Assumption 1. Thus there exists a single value of γ such that the incentive constraint (12) following the recommendation to abstain is just satisfied as an equality, that is, (16) holds. Second, the resulting cutoff mechanism $\pi^*_{\gamma^c}$ also satisfies the incentive constraint (11) following the recommendation to consume, for the corresponding mean posterior belief about θ is below $\mathbf{E}[\theta]$ and thus, a fortiori, below t^a if (15) does not hold. Thus our candidate optimal mechanism $\pi^*_{\gamma^c}$ is incentive-compatible, which achieves the characterization.

The key insight of Proposition 2 is that, following the recommendation to abstain, the decision maker is actually on the verge of falling into the harmful-consumption trap, as his

mean posterior belief about θ is just at the critical level t^a and is thus just high enough to induce him to abstain. This reflects that the probability $1-\gamma^c$ of issuing a recommendation to abstain is chosen in the mechanism $\pi^*_{\gamma^c}$ to alarm the decision maker in an optimal way: any higher value would undermine the credibility of the mechanism, whereas any lower value would make the recommendation to abstain inefficiently alarming. An optimal balance is thus achieved between the credibility and the efficiency of the mechanism.

Finally, we give a more explicit characterization of γ^c in the case where (15) does not hold. If the equation

$$\mathbf{E}[\theta \,|\, \theta > t] = t^a \tag{17}$$

has a solution $t = t^c$, then $\gamma^c = F(t^c)$ and $\pi^*_{\gamma^c} = 1_{\{\theta \leq t^c\}}$. If $\mathbf{P}$ has atoms, however, then such a solution need not exist because the mapping $t \mapsto \mathbf{E}[\theta \mid \theta > t]$ is discontinuous at the atoms of $\mathbf{P}$. In that case, the optimal incentive-compatible mechanism may necessitate randomization to achieve an equality in (12). Let us then define t^c as the supremum of the set of cutoffs that are too small to satisfy (17),

$$t^c \equiv \sup\{t \in [0, t^a] : \mathbf{E}[\theta \,|\, \theta > t] < t^a\}, \tag{18}$$

which is well defined under Assumption 1. Because $\mathbf{E}[\theta \,|\, \theta > t]$ is right-continuous in t, it follows that either (17) is satisfied by t^c or (17) has no solution. In the latter case, we have $\mathbf{P}[\theta = t^c] > 0$, $\mathbf{E}[\theta \,|\, \theta \geq t^c] \leq t^a$, and $\mathbf{E}[\theta \,|\, \theta > t^c] > t^a$. If the second of these inequalities is an equality, then it is optimal to recommend to abstain for sure at $\theta = t^c$ and $\pi^*_{\gamma^c} = 1_{\{\theta < t^c\}}$. If this inequality is strict, recommending to abstain for sure at $\theta = t^c$ would undermine the credibility of the mechanism, while recommending to consume for sure at $\theta = t^c$ would make the recommendation to abstain inefficiently alarming. Then γ^c is implicitly defined by (10) and (16) with $t_{\gamma^c} = \inf\{\theta : F(\theta) > \gamma^c\} = t^c$, reflecting how randomization allows us to interpolate through possible discontinuities of F.

4 Illustrations

We now illustrate our findings in the polar cases of binary and nonatomic distributions. We pay particular attention to the comparative statics of the optimal incentive-compatible persuasion mechanism with respect to changes in the distribution of types and the severity of the decision maker's bias for the present; one key question, notably, is how such changes affect the probability of consuming, and to which extent this is harmful from the decision maker's perspective at date 0.

4.1 The Binary Case

Suppose first that θ can only take two values $\underline{\theta}$ and $\overline{\theta}$ such that $t^h < \underline{\theta} < t^a$ and $\overline{\theta} > t^a$.[9] Hence, according to Lemma 1, the decision maker consumes at both dates and obtains a negative payoff if θ is revealed to be $\underline{\theta}$, and abstains at both dates and obtains a zero payoff if θ is revealed to be $\overline{\theta}$. To characterize the optimal incentive-compatible persuasion mechanism in this binary case, we can use Kamenica and Gentzkow's (2011) standard concavification argument, working directly in terms of the prior belief $\underline{p} = \mathbf{P}[\theta = \underline{\theta}]$ as in Aumann and Maschler (1995). Letting

$$\underline{p}^a \equiv \frac{\overline{\theta} - t^a}{\overline{\theta} - \underline{\theta}} \in (0,1), \tag{19}$$

the date-0 expected payoff of the decision maker is, up to a multiplicative constant $(1+\delta)/t^a$,

$$V_0(\underline{p}) \equiv [t^h - \underline{p}\underline{\theta} - (1-\underline{p})\overline{\theta}]1_{\{\underline{p} > \underline{p}^a\}},$$

which is negative for $\underline{p} > \underline{p}^a$; notice the downward discontinuity of V_0 at $\underline{p}^a$, reflecting the discontinuity in the decision maker's date-0 equilibrium payoff at $\hat{\boldsymbol{\theta}} = \underline{p}^a\underline{\theta} + (1-\underline{p}^a)\overline{\theta} = t^a$. The concavification $\operatorname{cav} V_0$ of V_0 coincides with V_0 over $[0, \underline{p}^a]$, where it is flat and equal to zero, and is affine and decreasing over $(\underline{p}^a, 1]$. Figure 2 below illustrates this construction.

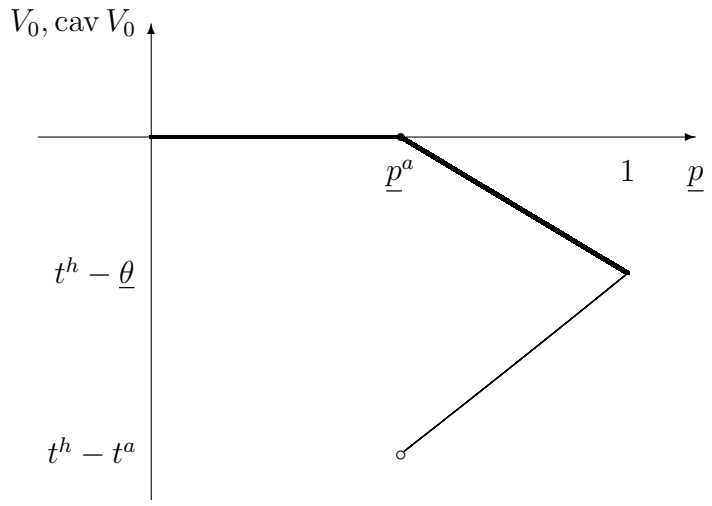

Figure 2: Concavifying the decision maker's date-0 equilibrium payoff.

Two regimes then emerge. When $\underline{p} \leq \underline{p}^a$, that is, when the decision maker would abstain

[9] The first of these two inequalities is not consistent with the first half of Assumption 1. However, a careful reading of Section 3.4 reveals that we only used the latter when discussing the unconstrained-optimal mechanism, which is not required in the present binary case.

absent further information, there is no reason to disclose such information, as doing so can never benefit the decision maker and may actually hurt him. The unconstrained-optimal mechanism is incentive-compatible and prescribes that no information be disclosed to the decision maker, who thus does not engage in harmful consumption.

By contrast, when $\underline{p} > \underline{p}^a$, that is, when the decision maker would consume absent further information, the optimal incentive-compatible mechanism involves randomization and, with positive probability, harmful consumption. Specifically, if $\theta = \overline{\theta}$, then the recommendation to consume is issued with probability $\pi^*_{\gamma^c}(\overline{\theta}) = 0$, while, if $\theta = \underline{\theta}$, then the recommendation to consume is issued with probability $\pi^*_{\gamma^c}(\underline{\theta}) \in (0, 1)$, where

$$\frac{\underline{p}[1 - \pi^*_{\gamma^c}(\underline{\theta})]}{\underline{p}[1 - \pi^*_{\gamma^c}(\underline{\theta})] + 1 - \underline{p}} = \underline{p}^a. \tag{20}$$

Therefore, the recommendation to consume reveals that $\theta = \underline{\theta}$, which triggers consumption as $\underline{\theta} < t^a$. By contrast, the recommendation to abstain does not fully disclose θ to the decision maker: according to (19)–(20), the decision maker's mean posterior belief about θ following the recommendation to abstain is equal to t^a, that is, the minimum level consistent with him abstaining. In this way, this recommendation is used in the most efficient way, while remaining credible.

It is interesting to point out how the optimal incentive-compatible persuasion mechanism responds to an increase in the severity of the decision maker's bias for the present, that is, a decrease in β. First, according to (3), the cutoff t^a increases, reflecting that the decision maker engages in potentially harmful consumption for a larger value of the mean posterior belief $\hat{\boldsymbol{\theta}}$. According to (19), this, in turn, decreases the cutoff belief $\underline{p}^a$ that $\theta = \underline{\theta}$ below which the decision maker abstains absent any further information: he must thus be more pessimistic to abstain. Finally, according to (20), if $p > \underline{p}^a$, the probability of issuing the recommendation to abstain conditional on $\theta = \underline{\theta}$ must decrease to preserve its credibility. Overall, harmful consumption is more likely to take place, the more severe the decision maker's bias for the present. The optimal information nudge then mostly targets high-risk individuals to the detriment of low-risk individuals, who would still prefer a warning that would deter them from consuming. An example of such selective nudging is alcohol warnings that target pregnant women—instead of the whole population of consumers who should better drink less.

4.2 The Nonatomic Case

Suppose next that $\mathbf{P}$ is nonatomic, with full support over $[0, 1]$. Then the optimal incentive-

compatible persuasion mechanism recommends abstinence when $\theta > t^* \equiv \max\{t^h, t^c\}$, where t^c is defined by (18). The most interesting scenario arises when (15) does not hold, so that the optimal incentive-compatible mechanism $\pi^*_{\gamma^c}$ involves harmful consumption and $t^c > t^h$ is the unique solution to (17). As a result, one also has $t^c < t^a$: therefore, there are types close to but below t^a, for which harmful consumption would necessarily take place under complete information, but which are completely neutralized under $\pi^*_{\gamma^c}$. That $t^c > t^h$ reflects that consumption must take place for types for which consumption is slightly harmful to preserve the credibility of the mechanism when it recommends abstinence for types for which consumption is more harmful, but would nevertheless take place if these types were disclosed. Notice, incidentally, that there are multiple ways of implementing the cutoff mechanism $\pi^*_{\gamma^c}$: indeed, consumption for types $\theta \geq t^c$ can indifferently be triggered by fully disclosing these types, or by sending the message that $\theta \leq t^c$. Thus, the optimal information nudge does not have to be simple—but it can be. What is crucial is the composition of the target group that receives a warning against consumption.

4.2.1 Sampling versus Information Design

It is instructive to compare these results to those obtained by Carrillo and Mariotti (2000), assuming that the decision maker can sample costless information about θ at date 0 before making his consumption decisions.[10] Then the decision maker never finds it optimal to consume without the benefit of full information about θ. Indeed, because, at any stage of the sampling process, his posterior beliefs have full support over $[0, 1]$ and thus put a strictly positive weight on the abstinence interval $[t^a, 1]$, he is strictly better off, before engaging in consumption, acquiring information that will either confirm his consumption decision or lead him to rationally abstain. In the present model, by contrast, the posterior belief of the decision maker following a recommendation to consume is $\mathbf{P}[\cdot \,|\, \theta \leq t^c]$, the support of which does not intersect $[t^a, 1]$ as $t^c < t^a$; as noted above, the decision maker is then indifferent about acquiring additional information about θ.

A common feature of the two models is that abstinence can be only sustained for mean posterior beliefs $\hat{\boldsymbol{\theta}} \geq t^a$; this inequality is typically strict when the decision maker samples information himself, while it is an equality in the optimal incentive-compatible persuasion mechanism. This, in turn, reflects that ignorance is achieved in different ways in the two models. In the sampling model, when the decision maker has a current posterior belief with mean $\hat{\boldsymbol{\theta}}$ slightly above t^a and with low variance, it is typically optimal for him to stop

[10] An infinitely-lived decision maker may also proceed to such sampling at dates $\tau = 1, 2, \ldots$ without affecting the results.

sampling. Indeed, conditional on $\theta < t^a$, it is likely that θ will be close to t^a; there is then a nonnegligible risk that the decision maker will eventually learn this and be trapped in harmful consumption. By contrast, in the present information-design model, types $\theta < t^a$ close to t^a are completely neutralized as they are pooled with types $\theta \geq t^a$. Thus, although the rationale for strategic ignorance is the same in the two models, and although the decision maker's beliefs follow a martingale in both cases, sampling creates an additional risk by making pieces of information available only sequentially; this creates a further motive for information avoidance, inducing the decision maker to be more cautious in his collection of information. By contrast, the release of signals in the information-design model is optimized by a mechanism designer, contingent on the value of θ; thus everything happens as if all sampling was done ex ante and different pieces of information were batched together to be optimally presented to the decision maker.

Overall, no decision-maker type should better avoid the optimized information nudge. In contrast, the only "shortcoming" of the optimal nudge may be that the target group is smaller than some types may wish for. However, a tightening of the target group is necessary in order to preserve credibility and efficiently mitigate self-control problems.

4.2.2 Comparative Statics

Distributions The characterization (17) of the cutoff t^c leads to unambiguous comparative statics in terms of the distribution $\mathbf{P}$. Suppose indeed that $\overline{\mathbf{P}}$ dominates $\underline{\mathbf{P}}$ in the hazard-rate order, that is, $(1-\overline{F})/(1-\underline{F})$ is increasing over $[0,1)$. By the full support assumption, the conditional distributions $\overline{\mathbf{P}}[\cdot\,|\,\theta > t]$ and $\underline{\mathbf{P}}[\cdot\,|\,\theta > t]$ are well defined for all $t \in [0,1)$, and the assumption that $\overline{\mathbf{P}}$ dominates $\underline{\mathbf{P}}$ in the hazard-rate order is equivalent to the condition that, for each $t \in [0,1)$, $\overline{\mathbf{P}}[\cdot\,|\,\theta > t]$ first-order stochastically dominates $\underline{\mathbf{P}}[\cdot\,|\,\theta > t]$ (Shaked and Shanthikumar (2007, Section 1.B.1)). This, in turn, implies that $\overline{\mathbf{E}}[\theta\,|\,\theta > t] > \underline{\mathbf{E}}[\theta\,|\,\theta > t]$ for any such t. It then follows from (17) that the cutoff t^c is strictly less under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$, $\bar{t}^c < \underline{t}^c$. Thus, if the optimal incentive-compatible persuasion mechanism when θ is distributed according to $\underline{\mathbf{P}}$ involves no consumption for some type, then neither does the optimal incentive-compatible persuasion mechanism when θ is distributed according to $\overline{\mathbf{P}}$. The intuition is that, for any cutoff $t \in [0,1)$, the announcement that $\theta > t$ is more efficient at discouraging consumption under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$. Hence it is credible to set the cutoff t^c at a lower value under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$, which allows the mechanism designer to neutralize a larger set of types for which consumption would be harmful. Because such types are more likely under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$ by first-order stochastic dominance, the following result holds.

Corollary 1 *If the distribution* $\overline{\mathbf{P}}$ *dominates the distribution* $\underline{\mathbf{P}}$ *in the hazard-rate order, the probability of consuming is strictly lower under* $\overline{\mathbf{P}}$ *than under* $\underline{\mathbf{P}}$.

Intuitively, two effects are reinforcing each other: it is more desirable to discourage consumption under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$, and it is also an easier task for the mechanism designer.

Bias for the Present We now turn to the comparative statics with respect to the severity of the decision maker's bias for the present. To see how a change in β affects the probability of consuming, it is helpful to start with a closer examination of the condition (15) under which the unconstrained-optimal mechanism $\pi^*_{\gamma^u}$ associated to the cutoff t^h is incentive-compatible. Using (3)–(4), this condition can be explicitly written as

$$\mathbf{E}\left[\theta \,|\, \theta > \frac{1+\beta\delta}{(1+\delta)\beta\delta^2 C}\right] \geq \frac{1}{\beta\delta^2 C}. \tag{21}$$

As a time-consistent decision maker never engages into harmful consumption, a natural guess is that the optimal incentive-compatible persuasion mechanism involves no harmful consumption and, hence, coincides with π^u when the decision maker's bias for the present is not too severe. This intuition is confirmed by the observation that, because the distribution $\mathbf{P}$ has full support, a sufficient condition for (21) to hold is that β be close enough to 1. This condition turns out to be necessary when the distribution $\mathbf{P}$ satisfies a weakening of the monotone-hazard-rate property. We will henceforth assume that $\mathbf{P}$ has a continuous density f over $[0,1]$ that is positive over $(0,1)$. The appropriate regularity concept for distributions can then be formulated as follows.

Definition 1 *The distribution* $\mathbf{P}$ *is* λ*-regular for some* $\lambda \geq 0$ *if*

$$r_\lambda(t) \equiv \frac{f(t)}{[1-F(t)]^\lambda} \tag{22}$$

is strictly increasing in $t \in [0,1)$.

It is clear from (22) that a lower value of λ corresponds to a more stringent restriction on the distribution $\mathbf{P}$; thus 0-regularity means that the density f is strictly increasing, 1-regularity is the strict monotone-hazard-rate property, and 2-regularity is equivalent to strict Myerson-regularity.[11] The following result holds.

Corollary 2 *If the distribution* $\mathbf{P}$ *is* $[2-1/(1+\delta)]$*-regular, then the unconstrained-optimal persuasion mechanism* $\pi^*_{\gamma^u}$ *is incentive-compatible if and only if* $\beta \geq \beta^u$, *where* β^u *is the unique value of* $\beta \in (1/(\delta^2 C), 1)$ *that achieves equality in (21).*

[11] See Ewerhart (2013) for this last equivalence and Schweizer and Szech (2017) for a systematic exploration.

Thus, for a fixed $[2-1/(1+\delta)]$-regular distribution $\mathbf{P}$, if the optimal incentive-compatible persuasion mechanism for a decision maker with time-inconsistency parameter $\underline{\beta}$ involves no harmful consumption, then neither does the optimal incentive-compatible persuasion mechanism for a decision maker with time-inconsistency parameter $\overline{\beta} > \underline{\beta}$. That is, harmful consumption takes place if and only if the decision maker's bias for the present is severe enough. Some regularity of the distribution of θ is necessary for obtaining such a clear-cut result. What is needed is a bound on the derivative with respect to t of the upper-tail conditional expectation $\mathbf{E}[\theta \,|\, \theta > t]$.[12] We will henceforth assume that the distribution $\mathbf{P}$ is $[2 - 1/(1 + \delta)]$-regular.

Observe that the cutoff $t^* = \max\{t^h, t^c\}$ for θ above which abstinence is recommended is strictly decreasing in $\beta \in (1/(\delta^2 C), 1)$. Indeed, if $\beta \in [\beta^u, 1)$, then $t^h \geq t^c$, and this directly follows from (3)–(4); if $\beta \in (1/(\delta^2 C), \beta^u)$, then $t^c > t^h$, and this directly follows from (3) and (18). Thus, if the optimal incentive-compatible persuasion mechanism for a decision maker with time-inconsistency parameter $\underline{\beta}$ involves abstinence for a given value of θ, then so does the optimal incentive-compatible persuasion mechanism for a decision maker with time-inconsistency parameter $\overline{\beta} > \underline{\beta}$. That is, a more severe bias for the present increases the probability that consumption takes place.

What about *harmful* consumption? There are two effects at play here. On the one hand, from the above reasoning, there are values of θ such that the decision maker would be trapped in harmful consumption under $\underline{\beta}$ but abstain under $\overline{\beta}$; on the other hand, according to (3)–(4), the lower bound t^h of the harmful-consumption interval $[t^h, t^a]$ is lower under $\overline{\beta}$ than under $\underline{\beta}$, because the decision maker attaches greater importance to future costs of consuming. Hence, any statement about how harmful consumption varies with β under the optimal incentive-compatible persuasion mechanism is necessarily of a probabilistic nature. The following result is a first step in that direction. It shows that, under a strengthening of the strict monotone-hazard-rate property, harmful consumption is more likely to take place, the more severe the decision maker's bias for the present is.

Corollary 3 *If the distribution* $\mathbf{P}$ *satisfies the strict monotone-hazard-rate property and its density* f *does not decrease too fast, in the sense that, for all* t *and* t',

$$t > t' \quad \textit{implies} \quad f(t) > \frac{1}{1+\delta} f(t'), \tag{23}$$

then the probability $F(t^c) - F(t^h)$ *that harmful consumption takes place under the optimal incentive-compatible persuasion mechanism is strictly decreasing in* $\beta \in (1/(\delta^2 C), \beta^u)$.

[12] Remark A.1 in the Online Appendix provides an example that violates this bound; the unconstrained-optimal persuasion mechanism is then incentive-compatible over two disjoint intervals of values for β.

Condition (23) is satisfied, for instance, if $\mathbf{P}$ is the uniform distribution. However, it is not satisfied, for instance, if $\mathbf{P}$ is a Beta(a, b) distribution with $a, b > 1$, which satisfies the monotone-hazard-rate property (Bagnoli and Bergstrom (2005)), but not condition (23) as then $f(1) = 0$. The following result shows that Corollary 3 does not extend to this case.

Corollary 4 *If the distribution* $\mathbf{P}$ *satisfies the strict monotone-hazard-rate property and its density* f *is nonincreasing in a left-neighborhood of* $t = 1$ *or strictly positive at* $t = 1$*, and if*

$$f(1) < \frac{1}{2(1+\delta)} f\left(\frac{1 + 1/(\delta C)}{1+\delta}\right), \tag{24}$$

then the probability $F(t^c) - F(t^h)$ *that harmful consumption takes place under the optimal incentive-compatible persuasion mechanism is strictly increasing in* β *in a right-neighborhood of* $\beta = 1/(\delta^2 C)$*.*

Thus, whenever the decision maker's bias for the present is already severe, a decrease in this bias can actually lead to an increase in the probability of harmful consumption. This contrasts with the binary case, where the probability of harmful consumption is decreasing in β. Condition (24), which is satisfied as soon as $f(1) = 0$, can be intuitively interpreted as follows. If initially $\beta \approx 1/(\delta^2 C)$, then almost all types consume under the optimal incentive-compatible persuasion mechanism, that is, $t^c \approx 1$. If β increases by $\mathrm{d}\beta$, then the cutoff t^c above which abstinence is recommended decreases by some amount $\mathrm{d}t^c$, so that a mass of types approximately equal to $f(1)\,\mathrm{d}t^c$ can be neutralized. At the same time, however, the cutoff t^h above which consumption is harmful,

$$t^h = \frac{\mathbf{E}[\theta \,|\, \theta > t^c] + 1/(\delta C)}{1+\delta} \approx \frac{1 + 1/(\delta C)}{1+\delta},$$

decreases by an amount $\mathrm{d}t^h$ approximately equal to $1/(1+\delta)\,\{\mathrm{d}\mathbf{E}[\theta \,|\, \theta > t]/\mathrm{d}t\}|_{t=1}\,\mathrm{d}t^c$, which is at least $1/[2(1+\delta)]\,\mathrm{d}t^c$ under the weak conditions we impose on f.[13] The mass of new types thus trapped in harmful consumption is approximately equal to $1/[2(1+\delta)]f(t^h)\,\mathrm{d}t^c$, which exceeds the mass $f(1)\,\mathrm{d}t^c$ of neutralized types if (24) is satisfied.

5 Alternative Objective Functions

So far, we have focused on benevolent persuasion mechanisms that maximize the decision maker's date-0 utility. We now contrast this optimal liberal policy with the optimal policies

[13] The intuition for the factor 2 is easy to grasp when $f(1) > 0$. Indeed, in that case, the distribution of θ conditional on $\theta > t$ is approximately uniform when t is close to 1 as f is continuous, and hence a marginal increase $\mathrm{d}t$ in t increases $\mathbf{E}[\theta \,|\, \theta > t]$ by approximately $\mathrm{d}[\frac{1}{2}(t+1)] = \frac{1}{2}\,\mathrm{d}t$.

of other interest groups. For instance, a lobbyist might have an interest in implementing an information nudge that convinces as many people as possible to consume. By contrast, a health authority focusing on the long-run health effects of harmful consumption and ignoring its short-term enjoyable aspects might want to use an information nudge that deters as many people as possible from consuming. Motivated by these two polar cases, we consider the problems of finding incentive-compatible mechanisms that maximize or minimize the expected probability of consuming, solving, respectively,[14]

$$\max\{\mathbf{E}[\pi(\theta)] : \pi \text{ is incentive-compatible}\}, \tag{25}$$

$$\min\{\mathbf{E}[\pi(\theta)] : \pi \text{ is incentive-compatible}\}. \tag{26}$$

Recall that, if some mechanism π such that $\mathbf{E}[\pi(\theta)] = \gamma$ is incentive-compatible, then so is the cutoff mechanism π^*_γ. We can thus again focus on cutoff mechanisms. The maximizer in (25) wants to choose the largest incentive-compatible γ, while the minimizer in (26) wants to choose the smallest one. Hence a lobbyist wants the target group such that the warning loses its impact to be as large as possible; by contrast, a health authority wants to send a convincing warning and therefore needs to tighten the target group. If the bias for the present is severe enough, this may imply that many types who would rather abstain cannot be warned. In the following, we study how this sacrifice needs to be done.

Observe that, depending on the parameters of the model, one of the two problems (25)–(26) is always trivial. Indeed, if $\mathbf{E}[\theta] < t^a$, then the decision maker consumes absent further information. The uninformative persuasion mechanism associated to $\gamma = 1$ thus solves (25) in this case, so that, from a lobbyist's perspective, there is no need for an information nudge. Conversely, if $\mathbf{E}[\theta] \geq t^a$, then the decision maker abstains absent further information. The uninformative persuasion mechanism associated to $\gamma = 0$ thus solves (26) in this case, so that, from a health authority's perspective, there is no need for an information nudge. However, a lobbyist would like to spread information, in order to seduce low-risk types into harmful consumption, as we analyze in detail below.

Overall, depending on the value of β, the same nudge either minimizes or maximizes the probability of consumption. Specifically, notice that there always exists $\beta^m \in (0,1)$ such that $t^a(\beta^m) = \mathbf{E}[\theta]$. The mechanism that minimizes the probability of consumption for $\beta \geq \beta^m$ maximizes it for $\beta < \beta^m$. Therefore, a misspecification of β can lead to an information nudge with consequences opposite to those initially intended: a miscalibrated health authority may

[14]Notice that this formulation implicitly assumes that θ is unobserved by the mechanism designer. One possible interpretation is that θ corresponds to a consumer's individual disposition for being harmed by consumption which is independently distributed across consumers.

think that a lobbyist's policy is ideal from a health perspective when, indeed, exactly the opposite is the case. A health authority must thus be careful not to overestimate β, that is, not to underestimate agents' bias for the present. The converse holds for a lobbyist who must be careful not to overestimate agents' bias for the present.

There remains to study (25) for $\mathbf{E}[\theta] \geq t^a$ and (26) for $\mathbf{E}[\theta] < t^a$. Consider first the latter problem. The left-hand side of (11) strictly and continuously increases in γ from $\underline{\theta}$ to $\mathbf{E}[\theta]$, while the left-hand side of (12) strictly and continuously increases in γ from $\mathbf{E}[\theta]$ to $\overline{\theta}$. By Assumption 1, $\underline{\theta} < t^a < \overline{\theta}$. Thus (11) is satisfied for all $\gamma \in [0, 1]$. To satisfy (12), γ has to be chosen sufficiently large. By continuity, there exists a single value $\gamma^{\min}$ of γ in $(0, 1)$ such that (12) holds with equality,

$$\frac{\mathbf{E}[\theta[1 - \pi^*_{\gamma^{\min}}(\theta)]]}{\mathbf{E}[1 - \pi^*_{\gamma^{\min}}(\theta)]} = t^a. \tag{27}$$

Thus $\gamma^{\min}$ is the smallest value of γ that is consistent with an incentive-compatible persuasion mechanism. As a result, $\pi^*_{\gamma^{\min}}$ solves (26). A key observation that follows from (16) and (27) is that $\gamma^{\min}$ coincides with γ^c, the probability of consuming in the decision-maker-optimal incentive-compatible mechanism in the case where the unconstrained-optimal mechanism $\pi^*_{\gamma^u}$ is not incentive-compatible. Hence we can reinterpret the mechanism characterized in Propositions 1–2 as follows: if possible, implement the unconstrained-optimal mechanism; otherwise, implement the mechanism that minimizes the probability of consuming. The decision maker's interests are, therefore, aligned with those of a health authority aiming at minimizing the probability of consuming if his bias for the present is severe enough.

Figure 3 illustrates the relation between the optimal liberal policy and the consumption-minimizing policy as functions of the time-inconsistency parameter β. Types are distributed with quadratic density $f(\theta) = 12(\theta - \frac{1}{2})^2$, and we set $\delta = 0.9$ and $C = 1.5$. As the figure demonstrates, for $\beta \gg \beta^u \approx 0.867$, the proportion of consumers who abstain under the consumption-minimizing policy (dotted line) is much higher than the proportion of those who abstain under the optimal liberal policy (dashed line). In this case, the optimal liberal policy coincides with the unconstrained-optimal policy. When the bias for the present is more severe, that is, for β between $1/(\delta^2 C) \approx 0.8230$ and β^u, stronger warnings are necessary in order to successfully deter consumption. The abstinence probabilities of the optimal liberal policy and of the consumption-minimizing policy coincide (solid black line). As β further decreases towards $1/(\delta^2 C)$, the probability of harmful consumption (solid grey line) increases substantially because the target group receiving a credible warning needs to be increasingly tightened. For β near $1/(\delta^2 C)$, more than 20% of consumers are in the harmful-consumption

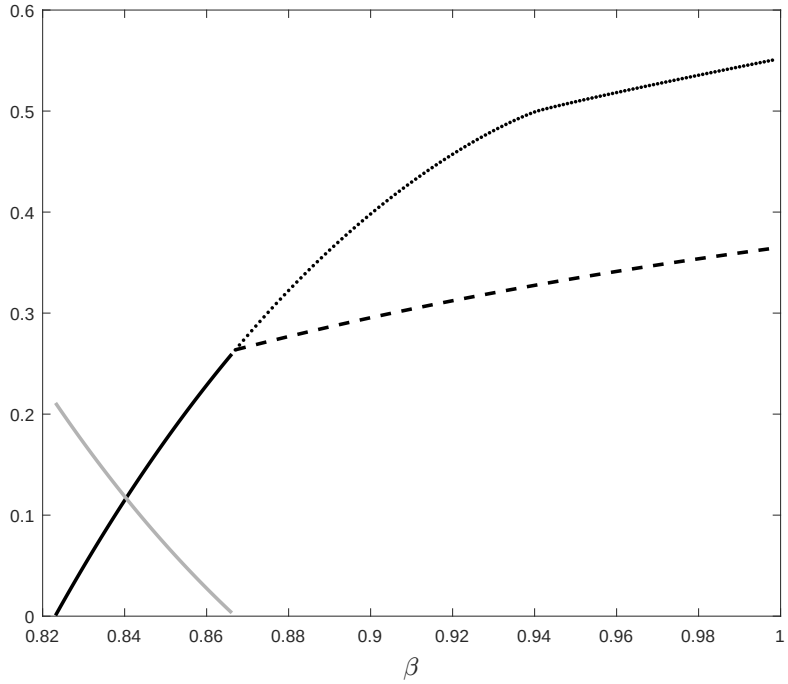

Figure 3: Probabilities of abstinence and of harmful consumption as functions of β.

trap, and only a tiny fraction of types can be convinced to abstain under such a severe bias for the present.

Let us now turn to the lobbyist's perspective. Problem (25) for $\mathbf{E}[\theta] \geq t^a$ is almost the mirror image of problem (26) for $\mathbf{E}[\theta] < t^a$. In this case, (12) trivially holds, while γ has to be chosen sufficiently small to ensure that (11) is satisfied. We can characterize a threshold $\gamma^{\max}$ via equality in (11),

$$\frac{\mathbf{E}[\theta \pi^*_{\gamma^{\max}}(\theta)]}{\mathbf{E}[\pi^*_{\gamma^{\max}}(\theta)]} = t^a. \tag{28}$$

However, due the fact that (6) is a strict inequality constraint, the mechanism $\pi^*_{\gamma^{\max}}$ is not incentive-compatible if the decision maker's behavior is described by the efficient subgame-perfect equilibrium of the intrapersonal game characterized in Lemma 1. As a result, (25) does not possess a solution. Instead, the best the lobbyist can do is to implement a cutoff mechanism with $\gamma = \gamma^{\max} - \varepsilon$ for some small $\varepsilon > 0$. Alternatively, we may assume that, in the intrapersonal game, self-0 and self-1 coordinate on the least efficient subgame-perfect equilibrium, in which they both consume for a mean posterior belief $\hat{\boldsymbol{\theta}} = t^a$. In that case, it is possible to implement a cutoff mechanism with $\gamma = \gamma^{\max}$ which maximizes the expected probability of consuming. This mechanism cynically takes advantage of the decision maker's self-control problem by issuing the recommendation to consume in such a way that, upon

receiving it, the decision maker ends up, in terms of his date-0 utility, at the lowest point of the harmful-consumption trap. In analogy with (18), the corresponding cutoff for θ can be characterized as follows. If the equation

$$\mathbf{E}[\theta \,|\, \theta \leq t] = t^a \tag{29}$$

has a solution $t = t^{\max}$, then $\gamma^{\max} = F(t^{\max})$ and $\pi^*_{\gamma^{\max}} = 1_{\{\theta \leq t^{\max}\}}$. For now familiar reasons, if $\mathbf{P}$ has atoms, such a solution need not exist. In that case, the optimal mechanism may necessitate randomization to achieve an equality in constraint (11), as required by (28). In analogy with (18), let

$$t^{\max} \equiv \inf\{t \in [t^a, 1] : \mathbf{E}[\theta \,|\, \theta \leq t] \geq t^a\}, \tag{30}$$

which is well defined by Assumption 1 as $\mathbf{E}[\theta] \geq t^a$. Because $\mathbf{E}[\theta \,|\, \theta \leq t]$ is right-continuous in t, it follows that either (29) is satisfied by $t^{\max}$ or (29) has no solution. In the latter case, we have $\mathbf{P}[\theta = t^{\max}] > 0$, $\mathbf{E}[\theta \mid \theta < t^{\max}] \leq t^a$, and $\mathbf{E}[\theta \mid \theta \leq t^{\max}] > t^a$. If the second of these inequalities is an equality, then it is optimal to recommend to abstain for sure at $\theta = t^{\max}$ and $\pi^*_{\gamma^{\max}} = 1_{\{\theta < t^{\max}\}}$. If this inequality is strict, recommending to abstain for sure at $\theta = t^{\max}$ would make the recommendation to abstain inefficiently alarming, thereby preventing the lobbyist from inducing consumption with probability $\gamma^{\max}$, while, according to the third inequality, recommending to consume for sure at $\theta = t^{\max}$ would undermine the credibility of the mechanism. Randomization at the atom $t^{\max}$ is then required, in line with Lemma 2. In any case, it is easy to check from (30) that $\mathbf{P}[(t^a, t^{\max}]] > 0$, so that, if $\mathbf{E}[\theta] > t^a$, there are types who are trapped in harmful consumption who, had they not been exposed to further information, would have abstained. This shows how a present-biased decision maker can fall prey to an opportunistic information design.

For example, nutritionists argue that by issuing warnings for specific high-risk groups only, many foods may still feel appropriate for people of lower risk type.[15] These people then continue to consume not so healthy foods that they may otherwise have started to call into question. Examples include an abundant consumption of fatty cheese and meat products which can possibly deteriorate health, and should better be replaced by healthier choices such as vegetables and fruits. This is likely not only true for people with specifically high risk of stroke or heart disease, but for everybody.[16] Thus the release of a warning

[15] Compare, for instance, Fuhrman (2011).

[16] See, for instance, advice by the Mayo Clinic for a heart healthy diet, www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702 as well as dietary recommendations by the Australian Heart Foundation to those who had to suffer from a heart attack, www.heartfoundation.org.au/after-my-heart-attack/heart-attack-recovery/diets-and-meals.

to a high-risk group can at the same time function as a justification to continue harmful consumption for those of lower risk, belonging to the nontarget group. Policy makers need to be aware of this problem which arises because of present-biased preferences.

6 Traffic-Light Nudges

Not all individuals suffer from the same self-control problems. On the one hand, people seem to differ in their overall self-control capacities (Mischel (2014), Sutter, Kocher, Glätzle-Rützler, and Trautmann (2013)). Studies suggest that the genetic profile plays a significant role for whether or not a person becomes addicted to harmful behaviors (Davis and Loxton (2013)). Moreover, parenting seems to affect the development of self-control in children (Finkenauer, Engels, and Baumeister (2005)). On the other hand, the specific context can matter a lot. While smoking may be very tempting for some consumers, others may find it easy to resist cigarettes, yet lose their self-control when it comes to chocolate or candy. Also, self-control relies on levels of glucose available, so that a hungry individual may display comparatively little self-control (Gailliot and Baumeister (2007)).

To address these issues, we analyze optimal information nudges in a mixed population, a share $p_L \in (0,1)$ of which has low self-control and the remaining share p_H has high self-control, with corresponding time-inconsistency parameters $0 < \beta_L < \beta_H \leq 1$.[17] The vNM utility functions at dates 0 and 1 for type $i = L, H$ are given by

$$U_{i,0}(x_0, x_1, \theta) = x_0(1 - \beta_i\delta^2\theta C) + x_1\beta_i\delta(1 - \delta^2\theta C), \tag{31}$$

$$U_{i,1}(x_0, x_1, \theta) = -x_0\beta_i\delta C + x_1(1 - \beta_i\delta^2\theta C). \tag{32}$$

Whether a specific decision maker is of type L or H is unknown to the mechanism designer. Her goal is to maximize social welfare at date 0. In the following, we focus on the case where each decision maker is offered the same information structure. For simplicity, we will assume that $\mathbf{P}$ has a continuous density f over $[0,1]$ that is positive over $(0,1)$ and, whenever needed, that $\mathbf{P}$ is $[2 - 1/(1+\delta)]$-regular.

It is clear from (31)–(32) that, for any mean posterior belief $\hat{\boldsymbol{\theta}}$, type L consumes whenever type H does. Therefore, we can focus on measurable direct persuasion mechanisms $x : \Theta \times \Omega \to \{0, L, LH\}$ issuing a recommendation for both types to abstain (0), for only type L to consume (L), or for both types to consume (LH). In analogy with (5), the probability of issuing recommendation $j = 0, L, LH$ is

$$\pi_j(\theta) = \boldsymbol{\lambda}[\{\omega \in \Omega : x(\theta, \omega) = j\}]. \tag{33}$$

[17] In particular, we allow for the case $\beta_H = 1$ where type H is not present-biased.

As in Section 3.2, we can identify x with $\pi \equiv (\pi_0, \pi_L, \pi_{LH})$. For each type i, we denote by t_i^a, t_i^h, t_i^c, and $t_i^* \equiv \max\{t_i^h, t_i^c\}$ the relevant cutoffs defined in Sections 2–3.

6.1 The No-Externality Case

We first analyze under which circumstances the two types exert no externality on each other. For each type i, the optimal incentive-compatible persuasion mechanism characterized in Propositions 1–2 recommends abstinence if $\theta > t_i^*$, and we have $t_H^* < t_L^*$. The same outcome can be achieved in a mixed population if and only if the mechanism

$$(\pi_0^*, \pi_L^*, \pi_{LH}^*)(\theta) \equiv (1_{\{\theta > t_L^*\}}, 1_{\{t_H^* < \theta \leq t_L^*\}}, 1_{\{\theta \leq t_H^*\}}) \tag{34}$$

is incentive-compatible. This is the case if and only if, upon receiving recommendation L, type H is willing to abstain, that is,

$$\mathbf{E}[\theta \,|\, t_H^* < \theta \leq t_L^*] \geq t_H^a. \tag{35}$$

For β_H close enough to 1, we have $t_H^* = t_H^h \approx t_H^a$, and the incentive-compatibility constraint (35) is slack. By contrast, for β_H close enough to β_L, we have $(t_H^*, t_H^a) \approx (t_L^*, t_L^a)$ and the incentive-compatibility constraint (35) is violated as $t_L^* < t_L^a$. The following result formalizes the idea that the two types exert no externality on each other if and only if β_H is large enough relative to β_L, so that a single traffic-light nudge can replicate the outcome of the individually optimal information nudges.

Proposition 3 *If the distribution* $\mathbf{P}$ *is* $[2 - 1/(1+\delta)]$*-regular, then, for each* $\beta_L > 1/(\delta^2 C)$*, there exists a threshold* $\hat{\beta}_H(\beta_L) \in (\beta_L, 1)$ *such that the mechanism (34) is incentive-compatible if and only if* $\beta_H \geq \hat{\beta}_H(\beta_L)$*. Moreover, the threshold* $\hat{\beta}_H(\beta_L)$ *is strictly greater than* β^u *and is strictly increasing in* β_L*.*

6.2 The Externality Case

We now analyze the case where the individually optimal incentive-compatible persuasion mechanisms are not simultaneously implementable, so that the two types exert an externality on each other. A mechanism (π_0, π_L, π_{LH}) is incentive-compatible if and only if

$$\frac{\mathbf{E}[\theta \pi_0(\theta)]}{\mathbf{E}[\pi_0(\theta)]} \geq t_L^a, \tag{36}$$

$$\frac{\mathbf{E}[\theta \pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} < t_L^a, \tag{37}$$

$$\frac{\mathbf{E}[\theta \pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} \geq t_H^a, \tag{38}$$

$$\frac{\mathbf{E}[\theta\pi_{LH}(\theta)]}{\mathbf{E}[\pi_{LH}(\theta)]} < t_H^a. \tag{39}$$

Letting $\Pi_L \equiv \pi_L + \pi_{LH}$ and $\Pi_H \equiv \pi_{LH}$ be the respective probabilities of consuming for type L and type H, the optimal-design problem can then, up to a multiplicative constant $(1+\delta)\delta^2 C$, be formulated as[18]

$$\max\left\{\sum_i p_i\beta_i\{t_i^h\mathbf{E}[\Pi_i(\theta)] - \mathbf{E}[\theta\Pi_i(\theta)]\} : \pi \text{ is incentive-compatible}\right\}. \tag{40}$$

For simplicity, we will henceforth focus on the case where types L and H differ enough in their levels of self-control, so that the intervals $[t_L^h, t_L^a]$ and $[t_H^h, t_H^a]$ do not overlap.

Assumption 2 $t_H^a < t_L^h$.

Thus, conditional on the same posterior belief $\hat{\boldsymbol{\theta}} \in (t_H^a, t_L^h)$, type L at date 0 favors a higher consumption rate than type H at date 1. By (3)–(4), Assumption 2 is equivalent to

$$\beta_H > \tilde{\beta}_H(\beta_L) \equiv \frac{(1+\delta)\beta_L}{1+\beta_L\delta} \in (0,1),$$

so that β_H is large enough relative to β_L. This lower bound is consistent with $\beta_H < \hat{\beta}_H(\beta_L)$, in which case, according to Proposition 3, we are indeed in the externality case. To see this, suppose, for instance, that $\beta_L \in [\beta^u, 1)$, so that $t_L^* = t_L^h$ by Corollary 2. Then, for $\beta_H = \tilde{\beta}_H(\beta_L)$, we have $t_L^h = t_H^a > t_H^*$, and constraint (35) is violated as

$$\mathbf{E}[\theta|t_H^* < \theta \leq t_L^*] = \mathbf{E}[\theta|t_H^* < \theta \leq t_H^a] < t_H^a.$$

Hence, for $\beta_L \in [\beta^u, 1)$ and $\beta_H = \tilde{\beta}_H(\beta_L)$, the mechanism (34) is not incentive-compatible and the threshold $\hat{\beta}_H(\beta_L)$ in Proposition 3 satisfies $\hat{\beta}_H(\beta_L) > \tilde{\beta}_H(\beta_L)$. The following result shows that, under Assumption 2, a two-cutoff mechanism is optimal.

Proposition 4 *There exists a pair of cutoffs* $0 < t_{LH}^{**} \leq t_L^{**} \leq 1$ *such that*

$$(\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})(\theta) \equiv (1_{\{\theta > t_L^{**}\}}, 1_{\{t_{LH}^{**} < \theta \leq t_L^{**}\}}, 1_{\{\theta \leq t_{LH}^{**}\}}) \tag{41}$$

is an optimal incentive-compatible mechanism.

When $t_{LH}^{**} < t_L^{**}$, the optimal mechanism for simultaneously targeting types L and H can thus be implemented via a monotone traffic-light nudge; as we will see in Lemma 3,

[18]Notice that the populations share p_L and p_H can also be interpreted as Pareto weights in the mechanism designer's social-welfare function.

this is always the case under Assumption 2. High-risk consumers with $\theta > t_L^{**}$ receive a warning to abstain, regardless of their level of self-control. This signal corresponds to a red traffic light; all consumers will find their risk high enough to abstain. For intermediate-risk consumers with $t_{LH}^{**} < \theta \leq t_L^{**}$, those with high self-control receive a warning to abstain, while those with low self-control receive a recommendation to consume. This signal corresponds to a yellow traffic light; while consumers with high self-control will find their risk high enough to abstain, consumers with low self-control will consume. Low-risk consumers with $\theta \leq t_{LH}^{**}$ receive a recommendation to consume, regardless of their level of self-control. This signal corresponds to a green traffic light; all consumers will find their risk low enough to consume. Thus a traffic-light nudge can optimally reach consumers with low self-control without sacrificing consumers with high self-control.[19] Koenigstorfer, Groeppel-Klein, and Kamm (2014) confirm this prediction in an empirical study, comparing consumers with high and low levels of self-control.

Proposition 4 generalizes the optimality of cutoff mechanisms to the more realistic case of heterogenous β's. As in the proof of Lemma 2 for the homogeneous case, the intuition is based on a comparison of all mechanisms that assign the same probabilities to the different recommendations. As before, using a cutoff t_{LH}^{**} to distinguish between "green" and "yellow" is good for both efficiency and incentive-compatibility purposes. For the optimal decision whether to display "yellow" or "red" there arises, however, a novel tradeoff. On the one hand, pooling the highest risk types into "red" rather than "yellow" is good for efficiency purposes as this signal induces consumers to abstain regardless of their level of self-control. On the other hand, pooling the highest risk types into "yellow" rather than "red" is good for incentive-compatibility purposes as this relaxes the incentive constraint (38). In the Online Appendix, we prove that, under Assumption 2, the first effect dominates. We also show that, when Assumption 2 does not hold, the second effect may dominate, and a nonmonotone traffic-light nudge of the form

$$(\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})(\theta) \equiv (1_{\{\underline{t}_L^{**} < \theta \leq \bar{t}_L^{**}\}}, 1_{\{t_{LH}^{**} < \theta \leq \underline{t}_L^{**}\}} + 1_{\{\theta > \bar{t}_L^{**}\}}, 1_{\{\theta \leq t_{LH}^{**}\}}) \tag{42}$$

may be optimal.

Several studies document that traffic-light labels work. For example, they are used to promote healthy food choices, see Hawley, Roberto, Bragg, Liu, Schwartz, and Brownell (2013), Thorndike, Riis, Sonnenberg, and Levy (2014), and the references therein. Relying

[19] Another useful aspect of traffic-light nudges may be their easy-to-grasp connotation. A red signal may be an especially salient warning. Indeed, the empirical literature is mixed on whether traffic-light labels render the provision of information more effective or not, see VanEpps, Downs, and Loewenstein (2016) for a discussion. Yet, of course, this aspect is beyond the analysis of this paper.

on nationally representative data from six European nations, Reisch and Sunstein (2016) demonstrate that there is also broad support in the population for the introduction of such information nudges in order to support healthy eating habits and fight obesity.

Our next result explicitly characterizes the optimal cutoffs $(t^{**}_{LH}, t^{**}_{L})$.

Lemma 3 *Suppose that (35) does not hold, so that the individually optimal mechanisms with cutoffs t^*_H and t^*_L are not simultaneously implementable, and let $\hat{t}_{LH}(t^*_L)$ be implicitly defined by*

$$\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t^*_L) < \theta \leq t^*_L] = t^a_H. \tag{43}$$

Then the optimal cutoffs $(t^{**}_{LH}, t^{**}_{L})$ in (41) are given by

1. *$(\hat{t}_{LH}(t^*_L), t^*_L)$ if and only if*

$$\frac{p_H \beta_H}{p_L \beta_L} \, \frac{\hat{t}_{LH}(t^*_L) - t^h_H}{t^*_L - t^h_L} \leq \frac{t^a_H - \hat{t}_{LH}(t^*_L)}{t^*_L - t^a_H},$$

2. *$(t^*_H, 1)$ if and only if*

$$\frac{p_H \beta_H}{p_L \beta_L} \, \frac{t^*_H - t^h_H}{1 - t^h_L} \geq \frac{t^a_H - t^*_H}{1 - t^a_H},$$

3. *the unique solution to*

$$\mathbf{E}[\theta \,|\, t^{**}_{LH} < \theta \leq t^{**}_{L}] = t^a_H \;\; \textit{and} \;\; \frac{p_H \beta_H}{p_L \beta_L} \, \frac{t^{**}_{LH} - t^h_H}{t^{**}_{L} - t^h_L} = \frac{t^a_H - t^{**}_{LH}}{t^{**}_{L} - t^a_H} \tag{44}$$

otherwise.

The characterization in Case 3 exactly reflects the tradeoff faced by the mechanism designer when she attempts to simultaneously persuade both types. Pooling marginally more risks into "yellow" rather than in "green" by decreasing t^{**}_{LH} comes at a benefit proportional to $p_H \beta_H (t^{**}_{LH} - t^h_H)$ due to higher abstinence of type H. Yet there is also the marginal cost of tightening type H's incentive constraint from below, which is proportional to $t^a_H - t^{**}_{LH}$. Similarly, pooling marginally more risks into "red" rather than in "yellow" by decreasing t^{**}_{L} comes at a benefit proportional to $p_L \beta_L (t^{**}_{L} - t^h_L)$ due to higher abstinence of type L. Yet there is also the marginal cost of tightening type H's incentive constraint from above, which is proportional to $t^{**}_{L} - t^a_H$. In an interior solution, we obtain the standard result that the marginal rate of substitution equals the marginal cost ratio, where the cost is here measured in terms of tightening type H's incentive constraint. Case 1 corresponds to a

corner solution in which the marginal rate of substitution of decreases in t_{LH} for decreases in t_L is everywhere less than the marginal cost ratio, so that type L faces his individually optimal mechanism with cutoff t_L^*. Similarly, Case 2 corresponds to a corner solution in which the designer entirely gives up on inducing abstinence for type L in order to achieve the maximum possible abstinence probability for type H.

We conclude this section with comparative statics with respect to the population share, which notably determines which of Cases 1–3 in Lemma 3 arises.

Corollary 5 *Suppose that (35) does not hold, so that the individually optimal mechanisms with cutoffs t_H^* and t_L^* are not simultaneously implementable. Then there exist thresholds $0 \leq \underline{p} < \overline{p} \leq 1$ such that*

1. *for $p_H \in [0, \underline{p}]$, the optimal mechanism implements the individually optimal cutoff t_L^* for type L and the cutoff for type H is determined by (43),*

2. *for $p_H \in [\overline{p}, 1]$, the optimal mechanism implements the individually optimal cutoff t_H^* for type H, while type L always consumes,*

3. *for $p_H \in (\underline{p}, \overline{p})$, the optimal mechanism implements the interior solution determined by (44). Consumption of type H is strictly decreasing in p_H, while consumption of type L is strictly increasing in p_H.*

Moreover, $\underline{p} = 0$ if and only if the individually unconstrained-optimal mechanism for type L is incentive-compatible in the sense of Proposition 1, and similarly $\overline{p} = 1$ if and only if the individually unconstrained-optimal mechanism for type L is incentive-compatible in the sense of Proposition 1.

7 Concluding Remarks

In this paper, we studied the optimal design of credible information nudges for populations of heterogeneous consumers with present-biased preferences. We found that the implementation of optimal information structures is easy in the sense that they are of cutoff type: an optimal information nudge should focus on a specific target group, and present a signal that is credible to this target group.

Yet the design of optimal information nudges is challenging in the sense that the bias for the present plays a crucial role: depending on how drastic it is, the target group needs to be adapted. Populations with a heavy bias need a much more drastic signal in order to

avoid harmful consumption. From a liberal designer's perspective, this means that fewer consumers can receive a credible signal to abstain. If consumers have different biases for the present, the traffic-light structure of the optimal nudge addresses this problem by releasing a specifically strong, red warning in addition to a milder, yellow warning.

A lobbyist aiming at high consumption rates will provide an information nudge of no impact, or, worse, one that tempts people into consumption who would otherwise abstain. If policy makers overlook or underestimate consumers' self-control problems, such a nudge may seem health-concerned when in fact exactly the opposite is the case. It is thus a necessity for policy makers to figure in effects of self-control when it comes to the design and evaluation of powerful information nudges to limit harmful consumption.

References

[1] Ainslie, G. (1975): "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, 82(4), 463–496.

[2] Ainslie, G. (1992): *Picoeconomics*: *The Strategic Interaction of Successive Motivational States within the Person*, Cambridge, UK: Cambridge University Press.

[3] Argo, J.J., and K.J. Main (2004): "Meta-Analyses of the Effectiveness of Warning Labels," *Journal of Public Policy Marketing*, 23(2), 193–208.

[4] Aumann, R.J. (1964): "Mixed and Behavior Strategies in Infinite Extensive Games," in *Advances in Game Theory*, Annals of Mathematics Study, Vol. 52, ed. by M. Dresher, L.S. Shapley, and A.W. Tucker. Princeton: Princeton University Press, 627–650.

[5] Aumann, R.J., and M.B. Maschler (1995): *Repeated Games with Incomplete Information*, Cambridge, MA: MIT Press.

[6] Bagnoli, M., and T. Bergstrom (2005): "Log-Concave Probability and Its Applications," *Economic Theory*, 26(2), 445–469.

[7] Bénabou, R. and J. Tirole (2002): "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117(3), 871–915.

[8] Benkert, J.-M., and N. Netzer (2018): "Informational Requirements of Nudging," *Journal of Political Economy*, forthcoming.

[9] Blackwell, D. (1953): "Equivalent Comparisons of Experiments," *Annals of Mathematical Statistics*, 24(2), 265–272.

[10] Bollinger, B., P. Leslie, and A. Sorensen (2011): "Calorie Posting in Chain Restaurants," *American Economic Journal*: *Economic Policy*, 3(1), 91–128.

[11] Brocas, I., and J.D. Carrillo (2007): "Influence through Ignorance," *RAND Journal of Economics*, 38(4), 931–947.

[12] Caplin, A., and J. Leahy (2004): "The Supply of Information by a Concerned Expert," *Economic Journal*, 114(497), 487–505.

[13] Carrillo, J.D., and T. Mariotti (2000): "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies*, 67(3), 529–544.

[14] Coffman, L., C.R. Featherstone, and J.B. Kessler (2015): "A Model of Information Nudges", Working Paper, Wharton School, University of Pennsylvania.

[15] Davis, C., and N.J. Loxton (2013): "Addictive Behaviors and Addiction-Prone Personality Traits: Associations with a Dopamine Multilocus Genetic Profile," *Addictive Behaviors*, 38(7), 2306–2312.

[16] Ewerhart, C. (2013): "Regular Type Distributions in Mechanism Design and ρ-Concavity," *Economic Theory*, 53(3), 591–603.

[17] Finkenauer, C., R. Engels, and R. Baumeister (2005): "Parenting Behaviour and Adolescent Behavioural and Emotional Problems: The Role of Self-Control," *International Journal of Behavioral Development*, 29(1), 58–69.

[18] Fuhrman, J. (2011): *Eat to Live*, New York: Little Brown.

[19] Gailliot, M.T., and R.F. Baumeister (2007): "The Physiology of Willpower: Linking Blood Glucose to Self-Control," *Personality and Social Psychology Review*, 11(4), 303–327.

[20] Gul, F., and W. Pesendorfer (2001): "Temptation and Self-Control," *Econometrica*, 69(6), 1403–1435.

[21] Gutjahr, E., G. Gmel, and J. Rehm (2001): "Relation between Average Alcohol Consumption and Disease: An Overview," *European Addiction Research*, 7(3), 117–127.

[22] Habibi, A. (2017): "Motivation and Information Design," Unpublished Manuscript, Department of Economics, University College London.

[23] Hall, W. (2010): "What Are the Policy Lessons of National Alcohol Prohibition in the United States, 1920–1933?" *Addiction*, 105(7), 1164–1173.

[24] Hammond, D., G.T. Fong, P.W. McDonald, K.S. Brown, and R. Cameron (2004): "Graphic Canadian Cigarette Warning Labels and Adverse Outcomes: Evidence from Canadian Smokers," *American Journal of Public Health*, 94(8), 1442–1445.

[25] Hammond, D., G.T. Fong, A. McNeill, R. Borland, and K.M. Cummings (2005): "Effectiveness of Cigarette Warning Labels in Informing Smokers about the Risks of Smoking: Findings from the International Tobacco Control (ITC) Four Country Survey," *Tobacco Control*, 15(Suppl III), iii19–iii25.

[26] Hankin, J.R., I.J. Firestone, J.J. Sloan, J.W. Ager, A.C. Goodman, R.J. Sokol, and S.S. Martier (1993): "The Impact of the Alcohol Warning Label on Drinking During Pregnancy," *Journal of Public Policy and Marketing*, 12(1), 10–18.

[27] Hawley K.L., C.A. Roberto, M.A. Bragg, P.J. Liu, M.B. Schwartz, and K.D. Brownell (2013): "The Science of Front-of-Package Food Labels," *Public Health Nutrition*, 16(3), 430–439.

[28] Hurwitz, A., and O. Sade (2017): "An Investigation of Time Preferences, Life Expectancy and Annuity versus Lump-Sum Choices—Can Smoking Harm Long-Term Saving Decisions?" Unpublished Manuscript, Department of Finance, Hebrew University.

[29] Kamenica, E., and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101(6), 2590–2615.

[30] Koenigstorfer, J., A. Groeppel-Klein, and F. Kamm (2014): "Healthful Food Decision Making in Response to Traffic Light Color-Coded Nutrition Labeling," *Journal of Public Policy and Marketing*, 33(1), 65–77.

[31] Kolotilin, A., T. Mylovanov, A. Zapechelnyuk, and M. Li (2017): "Persuasion of a Privately Informed Receiver," *Econometrica*, 85(6), 1949–1964.

[32] Köszegi, B. (2003): "Health Anxiety and Patient Behavior," *Journal of Health Economics*, 22(6), 1073–1084.

[33] Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112(2), 443–477.

[34] Lipnowski, E., and L. Mathevet (2018): "Disclosure to a Psychological Audience," *American Economic Journal*: *Microeconomics*, 10(4), 67–93.

[35] Loewenstein, G., and D. Prelec (1992): "Anomalies in Intertemporal Choice: Evidence and an Interpretation," *Quarterly Journal of Economics*, 107(2), 573–597.

[36] MacKinnon, D.P., M.A. Pentz, and A.W. Stacy (1993): "The Alcohol Warning Label and Adolescents: The First Year," *American Journal of Public Health*, 83(4), 585–587.

[37] McCool, J., L. Webb, L.D. Cameron, and J. Hoek (2012): "Graphic Warning Labels on Plain Cigarette Packs: Will They Make a Difference to Adolescents?" *Social Science and Medicine*, 74(8), 1269–1273.

[38] Miron, J.A., and J. Zwiebel (1991): "Alcohol Consumption During Prohibition," *American Economic Review*, 81(2), 242–247.

[39] Miron, J.A., and J. Zwiebel (1995): "The Economic Case against Drug Prohibition," *Journal of Economic Perspectives*, 9(4), 175–192.

[40] Mischel, W. (2014): *The Marshmallow Test*: *Understanding Self-Control and how to Master It*, New York: Little Brown.

[41] Peleg, B., and M.E. Yaari (1973): "On the Existence of a Consistent Course of Action when Tastes are Changing," *Review of Economic Studies*, 40(3), 391–401.

[42] Phelps, E.S., and R.A. Pollak (1968): "On Second-Best National Saving and Game-Equilibrium Growth," *Review of Economic Studies*, 35(2), 185–199.

[43] Rayo, L., and I.R. Segal (2010): "Optimal Information Disclosure," *Journal of Political Economy*, 118(5), 949–987.

[44] Reisch, L.A., and C.R. Sunstein (2016): "Do Europeans Like Nudges?," *Judgment and Decision Making*, 11(4), 310–325.

[45] Schweizer, N., and N. Szech (2017): "The Quantitative View of Myerson Regularity," Unpublished Manuscript, Department of Economics, Karlsruhe Institute of Technology.

[46] Schweizer, N., and N. Szech (2018): "Optimal Revelation of Life-Changing Information," *Management Science*, forthcoming.

[47] Shaked, M., and J.G. Shanthikumar (2007): *Stochastic Orders*, New York: Springer.

[48] Shield, K.D., C. Parry, and J. Rehm (2014): "Chronic Diseases and Conditions Related to Alcohol Use," *Alcohol Research*: *Current Reviews*, 35(2), 155–171.

[49] Strotz, R.H. (1956): "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23(3), 165–180.

[50] Sutter, M., M.G. Kocher, D. Glätzle-Rützler, and S.T. Trautmann (2013): "Impatience and Uncertainty: Experimental Decisions Predict Adolescents' Field Behavior," *American Economic Review*, 103(1), 510–531.

[51] Thaler, R. (1981): "Some Empirical Evidence on Dynamic Inconsistency," *Economics Letters*, 8(3), 201–207.

[52] Thaler, R., and C.R. Sunstein (2008): *Nudge*: *Improving Decisions about Health*, *Wealth*, *and Happiness*, New Haven: Yale University Press

[53] Thorndike, A.N., J. Riis, L.M. Sonnenberg, and D.E. Levy (2014): "Traffic-Light Labels and Choice Architecture Promoting Healthy Food Choices," *American Journal of Preventive Medicine*, 46(2), 143–149.

[54] VanEpps, E.M., J.S. Downs, and G. Loewenstein (2016): "Calorie Label Formats: Using Numeric and Traffic Light Calorie Labels to Reduce Lunch Calories," *Journal of Public Policy and Marketing*, 35(1), 26–36.

Technical Appendix (For Online Publication)

Proof of Corollary 2. For future reference, we more generally show the result for any left-truncation $\mathbf{P}_b \equiv \mathbf{P}[\cdot \,|\, \theta \leq b]$ of $\mathbf{P}$, with cumulative distribution function F_b and probability density function f_b over the support $[0, b]$, where $1/(\delta^2 C) < b \leq 1$. Corollary 2 corresponds to the special case $b = 1$. An important observation is that λ-regularity is preserved by left-truncation.

Lemma A.1 *Suppose that, for some $\lambda \geq 0$, the distribution $\mathbf{P}$ is λ-regular. Then, for each $b \in (0, 1)$, the distribution $\mathbf{P}_b$ is λ-regular.*

Proof. For each $t \in [0, b)$, we have

$$r_{b,\lambda}(t) \equiv \frac{f_b(t)}{[1 - F_b(t)]^\lambda} \propto \frac{f(t)}{[F(b) - F(t)]^\lambda} = r_\lambda(t) \left[\frac{1 - F(t)}{F(b) - F(t)}\right]^\lambda, \tag{A.1}$$

so that $r_{b,\lambda}(t)$ is the product of two strictly positive and strictly increasing functions of t. The result follows. ■

Now, fix some $b \in (1/(\delta^2 C), 1)$ and, for each $\beta \in (1/(b\delta^2 C), 1)$, define

$$\phi_b(\beta) \equiv \mathbf{E}_b\left[\theta \,|\, \theta > \frac{1 + \beta\delta}{(1 + \delta)\beta\delta^2 C}\right] - \frac{1}{\beta\delta^2 C}. \tag{A.2}$$

We show that there exists a unique solution β_b^u to $\phi_b(\beta) = 0$ and that $\phi_b(\beta) \geq 0$ if and only if $\beta \geq \beta_b^u$. This, in particular, implies Corollary 2, with $\beta^u \equiv \beta_1^u$. Because f is continuous, so is ϕ_b. Hence, by the intermediate value theorem, we only need to check that $\phi_b(1/(b\delta^2 C)) < 0$, that $\phi_b(1) > 0$, and that ϕ_b is strictly increasing. As for the first two statements, we have

$$\phi_b\left(\frac{1}{b\delta^2 C}\right) = \mathbf{E}_b\left[\theta \,|\, \theta > \frac{1 + b\delta C}{(1 + \delta)\delta C}\right] - b \quad \text{and} \quad \phi_b(1) = \mathbf{E}_b\left[\theta \,|\, \theta > \frac{1}{\delta^2 C}\right] - \frac{1}{\delta^2 C},$$

and the result follows from $b\delta^2 C > 1$ and the fact that $\mathbf{P}_b$ has full support over $[0, b]$. As for the third statement, notice that, letting $\xi \equiv 1/(\beta\delta^2 C)$ and changing variables accordingly, it is equivalent to the claim that

$$\mathbf{E}_b\left[\theta \,|\, \theta > \frac{\xi + 1/(\delta C)}{1 + \delta}\right] - \xi$$

is strictly decreasing in $\xi \in (1/(\delta^2 C), b)$. That this is the case if $\mathbf{P}$ is $[2 - 1/(1 + \delta)]$-regular is a consequence of the following lemma, which generalizes the standard observation that the mean residual life is strictly decreasing in the age when $\mathbf{P}$ satisfies the monotone-hazard-rate property (see, for instance, Bryson and Siddiqui (1969)).

Lemma A.2 *Suppose that, for some $\lambda \in (0,2)$, the distribution $\mathbf{P}$ is λ-regular. Then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}_b[\theta \,|\, \theta > t]\right\} < \frac{1}{2-\lambda} \tag{A.3}$$

for all $b \in (0,1)$ and $t \in [0,b)$.

Proof. By Lemma A.1, $\mathbf{P}_b$ is λ-regular over its support $[0,b]$, so that $r_{b,\lambda}$ as defined by (A.1) is strictly increasing. For each $t \in [0,b)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}_b[\theta \,|\, \theta > t]\right\} = \frac{f_b(t)}{1-F_b(t)}\left\{\mathbf{E}_b[\theta \,|\, \theta > t] - t\right\}. \tag{A.4}$$

Suppose, by way of contradiction, that (A.3) does not hold for some $t \in [0,b)$. Then, according to (A.4) and to the strict monotonicity of $r_{b,\lambda}$, we have

$$\frac{f_b(\theta)}{[1-F_b(\theta)]^{\lambda}} > \frac{[1-F_b(t)]^{1-\lambda}}{(2-\lambda)\{\mathbf{E}_b[\theta \,|\, \theta > t] - t\}}$$

for all $\theta \in (t,b)$ and, therefore,

$$\int_t^b [1-F_b(\theta)]^{1-\lambda} f_b(\theta)\,\mathrm{d}\theta > \frac{[1-F_b(t)]^{1-\lambda}}{(2-\lambda)\{\mathbf{E}_b[\theta \,|\, \theta > t] - t\}} \int_t^b [1-F_b(\theta)]\,\mathrm{d}\theta. \tag{A.5}$$

As the integral on the left-hand side of (A.5) equals $[1-F_b(t)]^{2-\lambda}/(2-\lambda)$, rearranging yields

$$\mathbf{E}_b[\theta \,|\, \theta > t] - t > \frac{1}{1-F_b(t)} \int_t^b [1-F_b(\theta)]\,\mathrm{d}\theta. \tag{A.6}$$

Integrating by parts, we have

$$\int_t^b [1-F_b(\theta)]\,\mathrm{d}\theta = \int_t^b \theta f_b(\theta)\,\mathrm{d}\theta - t[1-F_b(t)].$$

Substituting in (A.6) and rearranging then yields

$$\mathbf{E}_b[\theta \,|\, \theta > t] - t > \mathbf{E}_b[\theta \,|\, \theta > t] - t,$$

a contradiction. The result follows. ■

By Lemma A.2, taking $\lambda = 2 - 1/(1+\delta)$ so that $1/(2-\lambda) = 1+\delta$ then implies

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\left\{\mathbf{E}_b\left[\theta \,|\, \theta > \frac{\xi + 1/(\delta C)}{1+\delta}\right] - \xi\right\} < 0$$

for all $\xi \in (1/(\delta^2 C), b)$. Hence the result. ■

Remark A.1. Figure A.1 below shows that some regularity of $\mathbf{P}$ is necessary for a clear-cut result like Corollary 2. The figure shows $t^a(\beta)$, $\mathbf{E}[\theta \,|\, \theta > t^h(\beta)]$ and $t^h(\beta)$ for $C = 2$ and $\delta = 1$, when $\mathbf{P}$ is a mixture of three uniform distributions with density

$$f(t) = 0.1 \cdot 1_{\{t\in[0,1]\}} + 0.45 \cdot 1_{\{t\in[0.69,0.71]\}} + 0.45 \cdot 1_{\{t\in[0.94,0.96]\}}. \tag{A.7}$$

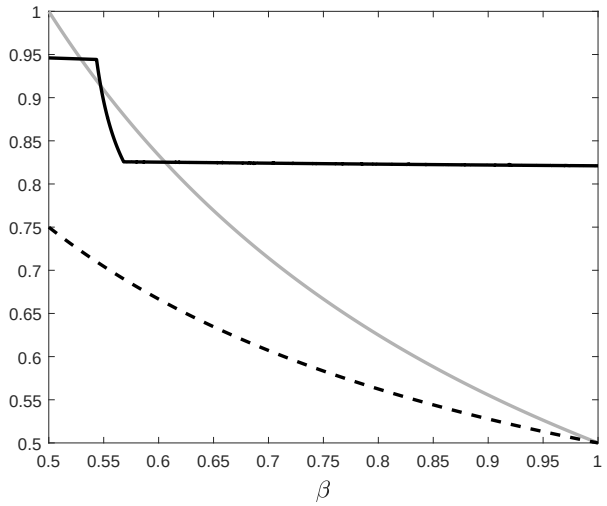

Figure A.1: $t^a(\beta)$ (in gray), $\mathbf{E}[\theta \,|\, \theta > t^h(\beta)]$ (in black) and $t^h(\beta)$ (dashed) as functions of β for θ distributed as in (A.7).

Most of the probability mass is thus concentrated in two small intervals around 0.7 and 0.95. As β increases, $t^h(\beta)$ decreases. Now, as $t^h(\beta)$ passes through the interval $[0.69, 0.71]$, which carries almost half the probability mass, we observe a steep drop in $\mathbf{E}[\theta \,|\, \theta > t^h(\beta)]$ from the interval $[0.94, 0.96]$ to values approximately in the middle between 0.7 and 0.95. Before the drop in the upper-tail conditional expectation, the signal that θ is above $t^h(\beta)$ is threatening enough to prevent harmful consumption. After the drop, this is no longer the case, and the optimal incentive-compatible mechanism can no longer prevent harmful consumption. Yet at some point as β increases further, the unconstrained-optimal mechanism becomes incentive-compatible again. Lemma A.2 shows that λ-regularity puts a bound on the derivative of the upper-tail conditional expectation function, thus preventing the type of behavior observed in Figure A.1.

Proof of Corollary 3. According to (3)–(4) and (18), we can rewrite the probability of harmful consumption as

$$F(t^c) - F(t^h) = F(t^c) - F\left(\frac{\mathbf{E}[\theta \,|\, \theta > t^c] + 1/(\delta C)}{1+\delta}\right).$$

As observed in the main text, t^c is strictly decreasing in $\beta \in (1/(\delta^2 C), \beta^u)$. Hence it is sufficient to show that

$$H(t) \equiv F(t) - F\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + 1/(\delta C)}{1+\delta}\right) \tag{A.8}$$

is strictly increasing in $t \in (t^u, 1)$, where

$$t^u \equiv \frac{1+\beta^u \delta}{(1+\delta)\beta^u \delta^2 C}.$$

Notice for future reference that, for each $t \in (t^u, 1)$,

$$t > \frac{\mathbf{E}[\theta \,|\, \theta > t] + 1/(\delta C)}{1+\delta} \tag{A.9}$$

because, as β^u is the unique value of $\beta \in (1/(\delta^2 C), 1)$ that achieves equality in (21), (A.9) becomes an equality at $t = t^u$ and because, as $\mathbf{P}$ satisfies the strict monotone-hazard-rate property, the mapping $t \mapsto (1+\delta)t - \mathbf{E}[\theta \,|\, \theta > t]$ is strictly increasing over $[0, 1)$. Then, for each $t \in (t^u, 1)$,

$$\begin{aligned} H'(t) &= f(t) - \frac{1}{1+\delta} f\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + 1/(\delta C)}{1+\delta}\right) \frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} \\ &\geq f(t) - \frac{1}{1+\delta} f\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + 1/(\delta C)}{1+\delta}\right) \\ &> 0, \end{aligned} \tag{A.10}$$

where the first inequality again follows from the strict monotone-hazard-rate property, and the second inequality follows from (23) and (A.9). Hence the result. ■

Proof of Corollary 4. Defining H as in (A.8), we have

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\left[F(t^c) - F(t^a)\right] > 0$$

in a strict right-neighborhood of $\beta = 1/(\delta^2 C)$ if and only if

$$H' < 0$$

in a strict left-neighborhood of $t = 1$ or, equivalently,

$$f(1) - \frac{1}{1+\delta} f\left(\frac{1 + 1/(\delta C)}{1+\delta}\right) \liminf_{t \to 1} \frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} < 0, \tag{A.11}$$

according to (A.10). We need to show that (24) implies (A.11) if $f(1) > 0$ or, if $f(1) = 0$, if f is nonincreasing in a left-neighborhood of $t = 1$.[1] That is, we need to show that, under these assumptions,

$$\liminf_{t \to 1} \frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} \geq \frac{1}{2}. \tag{A.12}$$

Suppose, by way of contradiction, that there exists a sequence $(t_n)_{n \in \mathbb{N}}$ in $(0, 1)$ converging to 1 such that, for some $\varepsilon > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\}\Big|_{t = t_n} < \frac{1-\varepsilon}{2}$$

[1] Notice that, in the latter case, condition (24) is automatically satisfied.

for all n. Then, according to (A.4), we have

$$f(t_n)\left\{\int_{t_n}^{1}\theta f(\theta)\,\mathrm{d}\theta - t_n[1-F(t_n)]\right\} - \frac{1-\varepsilon}{2}\,[1-F(t_n)]^2 < 0 \tag{A.13}$$

for all n. Consider then the function

$$I(t) = f(t)\left\{\int_{t}^{1}\theta f(\theta)\,\mathrm{d}\theta - t[1-F(t)]\right\} - \frac{1-\varepsilon}{2}\,[1-F(t)]^2.$$

We clearly have $I(1) = 0$. We now show that, under the stated assumptions on f, I is strictly decreasing in a left-neighborhood of $t = 1$, which, given (A.13), yields the desired contradiction as the sequence $(t_n)_{n\in\mathbb{N}}$ converges to 1. As I is continuous, it is sufficient to show that its right upper Dini derivative D^+I is strictly negative in a strict left-neighborhood of $t = 1$ (Giorgi and Komlósi (1992, Theorem 1.14)). Because f is continuous, the mapping $t \mapsto \int_t^1 \theta f(\theta)\,\mathrm{d}\theta - t[1-F(t)]$ is continuously differentiable. A simple calculation then shows that, for each $t \in (0,1)$,

$$\mathrm{D}^+I(t) = [1-F(t)](\mathrm{D}^+f(t)\{\mathbf{E}[\theta\,|\,\theta > t] - t\} - \varepsilon f(t)).$$

Now, recall that f is strictly positive over $(0,1)$. Thus, if $f(1) > 0$, then D^+I is strictly negative in a strict left-neighborhood of $t = 1$ because the mean residual life $\mathbf{E}[\theta\,|\,\theta > t] - t$ converges to zero as t goes to 1; similarly, if $f(1) = 0$, then, because the mean residual life $\mathbf{E}[\theta \mid \theta > t] - t$ is strictly positive for all $t \in [0,1)$, the same conclusion obtains if f is nonincreasing and hence its right upper Dini derivative D^+f is nonpositive in a strict left-neighborhood of $t = 1$. Hence the result. ■

Proof of Proposition 3. For each $\beta_H \in (\beta_L, 1)$, we denote by $t^a(\beta_H)$, $t^h(\beta_H)$, $t^c(\beta_H)$, and $t^*(\beta_H) \equiv \max\{t^h(\beta_H), t^c(\beta_H)\}$ the relevant cutoffs defined in Sections 2–3. It follows from (3)–(4) that t^a and t^h are continuous. As for t^c and t^*, notice that, for each $\beta_H \in (\beta_L, 1)$, the assumption that $\mathbf{P}$ has a continuous density f allows us to rewrite (18) as

$$\frac{\int_{t^c(\beta_H)}^{1}\theta f(\theta)\,\mathrm{d}\theta}{1-F(t^c(\beta_H))} = \frac{1}{\beta_H\delta^2 C}, \tag{A.14}$$

which implies, using again the assumption that f is continuous, that t^c and t^* are continuous as well. Now, for each $\beta_H \in (\beta_L, 1)$, define

$$\varphi_{t_L^*}(\beta_H) \equiv \mathbf{E}[\theta\,|\,t^*(\beta_H) < \theta \le t_L^*] - t^a(\beta_H) = \frac{\int_{t^*(\beta_H)}^{t_L^*}\theta f(\theta)\,\mathrm{d}\theta}{F(t_L^*) - F(t^*(\beta_H))} - \frac{1}{\beta_H\delta^2 C}. \tag{A.15}$$

Because f and t^* are continuous, so is $\varphi_{t_L^*}$. Hence, by the intermediate value theorem, we only need to check that $\varphi_{t_L^*}(\beta_L^+) < 0$, that $\varphi_{t_L^*}(1) > 0$, and that $\varphi_{t_L^*}$ crosses zero only once.

As for the first two statements, we have

$$\varphi_{t_L^*}(\beta_L^+) = t_L^* - t^a(\beta_L) \quad \text{and} \quad \varphi_{t_L^*}(1) = \mathbf{E}[\theta \,|\, t^*(1) < \theta \leq t_L^*] - t^a(1),$$

and the result follows from $t_L^* < t^a(\beta_L)$, $t^*(1) = t^a(1) = t^h(1) < t^h(\beta_L) \leq t_L^*$, and the fact that $\mathbf{P}$ has full support over $[0,1]$. As for the third statement, we distinguish two cases.

Case 1 If $\beta_L < \beta_H < \beta^u$, with β^u defined as in Corollary 2, then the unconstrained-optimal mechanism for type H is not incentive-compatible and, therefore, $t^*(\beta_H) = t^c(\beta_H) > t^h(\beta_H)$. In this case, from (A.14)–(A.15), we have

$$\varphi_{t_L^*}(\beta_H) = \frac{\int_{t^c(\beta_H)}^{t_L^*} \theta f(\theta)\,\mathrm{d}\theta}{F(t_L^*) - F(t^c(\beta_H))} - \frac{\int_{t^c(\beta_H)}^{1} \theta f(\theta)\,\mathrm{d}\theta}{1 - F(t^c(\beta_H))} < 0$$

as $t_L^* < 1$ and $\mathbf{P}$ has full support over $[0,1]$. It follows that $\varphi_{t_L^*}$ cannot cross zero over (β_L, β^u).

Case 2 If $\beta_H \geq \max\{\beta_L, \beta^u\}$, then the unconstrained-optimal mechanism for type H is incentive-compatible and, therefore, $t^*(\beta_H) = t^h(\beta_H)$. In this case, we have

$$\varphi_{t_L^*}(\beta_H) = \mathbf{E}\left[\theta \,|\, t_L^* \geq \theta > \frac{1+\beta_H\delta}{(1+\delta)\beta_H\delta^2 C}\right] - \frac{1}{\beta_H\delta^2 C} = \phi_{t_L^*}(\beta_H),$$

where $\phi_{t_L^*}(\beta_H)$ is as defined in (A.2) with $b = t_L^*$. As shown in the proof of Corollary 2, if $\mathbf{P}$ is $[2 - 1/(1+\delta)]$-regular, then $\phi_{t_L^*}$ is strictly increasing and vanishes at a single point $\beta_{t_L^*}^u$, which defines the desired threshold $\hat{\beta}_H(\beta_L)$. That $\hat{\beta}_H(\beta_L) > \beta^u$ was shown in Case 1. That $\hat{\beta}_H(\beta_L)$ is strictly increasing in β_L follows from the fact that $t_L^* = t^*(\beta_L)$ and, hence, $\phi_{t_L^*}$ are strictly decreasing in β_L. Hence the result. ■

Proof of Proposition 4. A useful preliminary observation is that, because the mechanism designer always prefers a higher abstinence rate than the decision maker, we can neglect constraints (37) and (39) in our quest for an optimal incentive-compatible mechanism.

Lemma A.3 *Any solution to the relaxed problem*

$$\max\left\{\sum_i p_i\beta_i\{t_i^h \mathbf{E}[\Pi_i(\theta)] - \mathbf{E}[\theta\Pi_i(\theta)]\} : \pi \text{ satisfies (36) and (38)}\right\} \tag{A.16}$$

is a solution to problem (40).

Proof. We show that any solution to (A.16) satisfies (37) and (39), and thus is a solution to (40). We accordingly distinguish two cases.

Case 1 Suppose, by way of contradiction, that a solution (π_0, π_L, π_{LH}) to (A.16) violates (37). Then type L would prefer to abstain whenever the recommendation is L. Because the utility from consumption is weakly lower for the mechanism designer than for type L, the former prefers abstinence for type L in this case, and a fortiori for type H as $t_H^a < t_L^a$. Thus the mechanism $(\pi_0 + \pi_L, 0, \pi_{LH})$ would satisfy (36) and (38) and improve upon the solution to (A.16), a contradiction.

Case 2 Suppose, by way of contradiction, that a solution (π_0, π_L, π_{LH}) to (A.16) violates (39). Then type H would prefer to abstain whenever the recommendation is LH. Because the utility from consumption is weakly lower for the mechanism designer than for type H, the former prefers abstinence for type H in this case. Thus the mechanism $(\pi_0, \pi_L + \pi_{LH}, 0)$ would satisfy (36) and (38) and improve upon the solution to (A.16), once again a contradiction. The result follows. ■

Among all mechanisms $\pi = (\pi_0, \pi_L, \pi_{LH})$ that issue recommendation LH with some probability γ_{LH}, the mechanisms with

$$\pi_{LH}(\theta) = 1_{\{\theta < t_{\gamma_{LH}}\}}$$

for $t_{\gamma_{LH}} \equiv F^{-1}(\gamma_{LH})$ are the best for efficiency purposes as they minimize the expected harm from consumption for a given probability of joint consumption. The following lemma shows that they are also best at satisfying the incentive constraints (36) and (38), as they issue recommendations to abstain to higher risk types than any other mechanism with the same probabilities of consumption recommendations that also satisfies these constraints.

Lemma A.4 *For any mechanism $\pi = (\pi_0, \pi_L, \pi_{LH})$ that satisfies (36) and (38), there exists a mechanism $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_L, \tilde{\pi}_{LH})$ that also satisfies (36), (38), and such that*

$$\mathbf{E}[\tilde{\pi}_j(\theta)] = \mathbf{E}[\pi_j(\theta)], \quad j = 0, L, LH, \tag{A.17}$$

$$\tilde{\pi}_{LH}(\theta) = 1_{\{\theta < t_{\gamma_{LH}}\}} \tag{A.18}$$

for $\gamma_{LH} \equiv \mathbf{E}[\pi_{LH}(\theta)]$ and $t_{\gamma_{LH}} \equiv F^{-1}(\gamma_{LH})$. Moreover, $\tilde{\pi}$ achieves a weakly higher value in (A.16) than π, and strictly so if π does not satisfy (A.18) on a $\mathbf{P}$-nonnull set.

Proof. We go back to the initial formulation of the optimal-design problem, in terms of direct persuasion mechanisms. Specifically, let $x : \Theta \times \Omega \to \{0, L, LH\}$ be the direct persuasion mechanism associated to π, and, for each $j \in \{0, L, LH\}$, let

$$\gamma_j(t_{\gamma_{LH}}) \equiv \mathbf{P} \otimes \boldsymbol{\lambda}[\{(\theta, \omega) \in \Theta \times \Omega : x(\theta, \omega) = j \wedge \theta < t_{\gamma_{LH}}\}]$$

be the probability that x issues recommendation j and $\theta < t_{\gamma_{LH}}$. Define a new direct persuasion mechanism

$$\tilde{x}(\theta,\omega) \equiv \begin{cases} LH & \text{if} \quad \theta \leq t_{\gamma_{LH}}, \\ L & \text{if} \quad \theta > t_{\gamma_{LH}} \wedge \Big(x(\theta,\omega) = L \vee \Big(x(\theta,\omega) = LH \wedge \omega < \frac{\gamma_L(t_{\gamma_{LH}})}{\gamma_0(t_{\gamma_{LH}})+\gamma_L(t_{\gamma_{LH}})}\Big)\Big), \\ 0 & \text{if} \quad \theta > t_{\gamma_{LH}} \wedge \Big(x(\theta,\omega) = 0 \vee \Big(x(\theta,\omega) = LH \wedge \omega \geq \frac{\gamma_L(t_{\gamma_{LH}})}{\gamma_0(t_{\gamma_{LH}})+\gamma_L(t_{\gamma_{LH}})}\Big)\Big), \end{cases}$$

and let $\tilde{\pi} \equiv (\tilde{\pi}_0, \tilde{\pi}_L, \tilde{\pi}_{LH})$ be the corresponding mechanism. The direct persuasion mechanism $\tilde{x}$ is constructed such that recommendation probabilities are the same as under the direct mechanism x, but consumption is recommended to both types if and only if $\theta \leq t_{\gamma_{LH}}$. Hence (A.17)–(A.18) hold by construction. Moreover, $\tilde{\pi}$ satisfies the incentive constraints (36) and (38), as it gives recommendations to abstain to higher risk types than π. Finally, $\tilde{\pi}$ weakly improves efficiency upon π, as it induces the same expected consumption levels with a lower expected harm from consumption, and strictly so if π does not satisfy (A.18) on a $\mathbf{P}$-nonnull set. The result follows. ■

Lemma A.4 implies that any solution $\pi^{**} = (\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})$ to (A.16) is such that, for some cutoff t_{LH}^{**},

$$\pi_{LH}^{**}(\theta) = 1_{\{\theta \leq t_{LH}^{**}\}}$$

up to a $\mathbf{P}$-null set. For any such mechanism, type H consumes if and only if $\theta \leq t_{LH}^{**}$. Thus his consumption behavior is already fully determined. Hence, given an optimal cutoff t_{LH}^{**}, problem (A.16) reduces to finding a measurable function $\pi_L^{**} : [0,1] \to [0,1]$ that vanishes over $[0, t_{LH}^{**}]$ and solves

$$\max\{t_L^h \mathbf{E}[\pi_L(\theta)] - \mathbf{E}[\theta\pi_L(\theta)] : \pi \text{ satisfies (36) and (38)}\}. \tag{A.19}$$

As in Section 3.3, the left-hand side of constraint (36) is not well defined if $\pi_0 = 0$ $\mathbf{P}$-almost surely over $(t_{LH}^{**}, 1)$, and similarly the left-hand side of constraint (38) is not well defined if $\pi_0 = 0$ $\mathbf{P}$-almost surely over $(t_{LH}^{**}, 1)$. We adopt the convention that the undefined constraint is emptily satisfied, which allows us to linearize the constraints (36) and (38). We start with an existence result.

Lemma A.5 *Problems (A.19), (A.16), and (40) have a solution.*

Proof. Our convention on the constraints (36) and (38) allows us to rewrite (A.19) as

$$\begin{aligned} \max\{t_L^h \mathbf{E}[\pi_L(\theta)] - \mathbf{E}[\theta\pi_L(\theta)] : \mathbf{E}[\theta[1-\pi_L(\theta)]] \geq t_L^a \mathbf{E}[1-\pi_L(\theta)] \\ \text{and } \mathbf{E}[\theta\pi_L(\theta)] \geq t_H^a \mathbf{E}[\pi_L(\theta)]\}, \end{aligned} \tag{A.20}$$

where the maximum is taken over the set

$$S \equiv \{\pi_L \in L_\infty(\mathbf{P}) : \pi_L(\theta) \in [0,1] \text{ for all } \theta \in [0,1] \text{ and } \pi_L(\theta) = 0 \text{ for all } \theta \in [0, t_{LH}^{**}]\}.$$

Notice that S is a closed subset of the unit ball $B_{L_\infty(\mathbf{P})}$ of $L_\infty(\mathbf{P})$ when the latter set is endowed with the weak* topology $\sigma(L_\infty(\mathbf{P}), L_1(\mathbf{P}))$, which we will henceforth assume without further mention. By the Banach–Alaoglu compactness theorem (Aliprantis and Border (2006, Theorem 6.21)), S is thus compact in that topology, and so is by duality the set S' of the functions in S that satisfy the constraints in (A.20); notice furthermore that S' is nonempty as it contains

$$\pi_L(\theta) = 1_{\{t_H^a < \theta \leq t_L^a\}} 1_{\{\theta > t_{LH}^{**}\}}.$$

Because S' is a nonempty compact set and the objective in (A.20) is continuous in π_L by duality, (A.20) and hence (A.19) have a solution.

To complete the proof, observe that, by Lemma A.3, we only need to show that (A.16) has a solution. Treating t_{LH}^{**} as a parameter, Berge maximum theorem (Aliprantis and Border (2006, Theorem 17.31)) implies that the solutions to (A.19) as t_{LH}^{**} varies are described by an upper hemicontinuous correspondence $\varpi_L^{**} : [0,1] \twoheadrightarrow B_{L_\infty(\mathbf{P})}$ with nonempty compact values. Thus, by Lemma A.4, (A.16) reduces to maximizing a continuous function of $(t_{LH}^{**}, \pi_L^{**})$ over $\{(t_{LH}^{**}, \pi_L^{**}) : t_{LH}^{**} \in [0,1] \text{ and } \pi_L^{**} \in \varpi_L^{**}(t_{LH}^{**})\}$, which is a compact set by the closed graph theorem (Aliprantis and Border (2006, Theorem 17.11)). The result follows. ■

We are now ready to characterize the solutions to (A.19).

Lemma A.6 *Problem (A.19) has a solution of the form (42). If Assumption 2 holds, then* $\bar{t}_L^{**} = 1$ *and, up to a* $\mathbf{P}$*-null set, any solution to problem (A.19) is of the form (41).*

Proof. We distinguish two cases.

Case 1 If constraint (38) is slack at the optimum, then (A.19) reduces to finding an optimal mechanism for type L alone, as described in Section 3. Propositions 1–2 imply that this mechanism is given by

$$\Pi_L^{**}(\theta) = 1_{\{\theta \leq t_L^*\}},$$

so that

$$\pi_L^{**}(\theta) = 1_{\{t_{LH}^{**} < \theta \leq t_L^*\}}.$$

Hence we must have $t_{LH}^{**} = t_H^*$. We thus fall back on the mechanism (34), which is incentive-

compatible if and only the no-externality condition (35) holds.

Case 2 If constraint (38) is binding at the optimum, that is, according to Case 1, if the no-externality condition (35) does not hold, then

$$\frac{\mathbf{E}[\theta\pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} = t_H^a. \tag{A.21}$$

Plugging (A.21) into the objective of (A.19), the problem becomes

$$\max\{(t_L^h - t_H^a)\mathbf{E}[\pi_L(\theta)] : \pi \text{ satisfies (36) and (A.21)}\}. \tag{A.22}$$

Our convention on the constraints (36) and (38) allows us to replace expectations in (A.22) by integrals, yielding the equivalent problem

$$\max\left\{(t_L^h - t_H^a)\int_{t_{LH}^{**}}^1 \pi_L(\theta)f(\theta)\,\mathrm{d}\theta : \int_{t_{LH}^{**}}^1 \theta[1-\pi_L(\theta)]f(\theta)\,\mathrm{d}\theta \geq t_L^a\int_{t_{LH}^{**}}^1 [1-\pi_L(\theta)]f(\theta)\,\mathrm{d}\theta \right.$$
$$\left. \text{and } \int_{t_{LH}^{**}}^1 \theta\pi_L(\theta)f(\theta)\,\mathrm{d}\theta = t_H^a\int_{t_{LH}^{**}}^1 \pi_L(\theta)f(\theta)\,\mathrm{d}\theta\right\},$$

where the maximum is taken over the set S defined in the proof of Lemma A.5. Because S is convex, and the objective as well as the constraints are affine in π_L, this equivalent problem is convex. Therefore, by the Kuhn–Tucker theorem (Clarke (2013, Theorem 9.4)), for any solution π_L^{**} to this problem, which is by construction a solution to (A.22) and (A.19), there exists a vector of Lagrange multipliers $(\eta^{**}, \lambda^{**}, \mu^{**})$ such that we have:

- Nontriviality:

$$(\eta^{**}, \lambda^{**}, \mu^{**}) \neq (0,0,0). \tag{A.23}$$

- Positivity:

$$\eta^{**} \in \{0,1\} \quad \text{and} \quad \lambda^{**} \in \mathbb{R}_+. \tag{A.24}$$

- Lagrangian maximization:

$$\pi_L^{**} \in \arg\max\left\{\int_{t_{LH}^{**}}^1 h^{**}(\theta)\pi_L(\theta)f(\theta)\,\mathrm{d}\theta : \pi_L \in S\right\}, \tag{A.25}$$

  where h^* is the affine function defined by

$$h^{**}(\theta) \equiv \eta^{**}(t_L^h - t_H^a) + \lambda^{**}t_L^a + \mu^{**}t_H^a - (\lambda^{**} + \mu^{**})\theta.$$

- Complementary slackness:

$$\lambda^{**}\left\{\int_{t_{LH}^{**}}^{1}\theta[1-\pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta - t_L^a\int_{t_{LH}^{**}}^{1}[1-\pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta\right\}=0. \tag{A.26}$$

- Equality constraint:

$$\int_{t_{LH}^{**}}^{1}\theta\pi_L^{**}(\theta)f(\theta)\,\mathrm{d}\theta = t_H^a\int_{t_{LH}^{**}}^{1}\pi_L^{**}(\theta)f(\theta)\,\mathrm{d}\theta. \tag{A.27}$$

We distinguish four subcases.

Subcase 2.1 If $h^{**}(\theta)>0$ for all $\theta\in(t_{LH}^{**},1)$, then the objective in (A.25) is uniquely (up to a **P**-null set) maximized over S by

$$\pi_L^{**}(\theta)=1_{\{\theta\geq t_{LH}^{**}\}},$$

which corresponds to a cutoff $t_L^{**}=1$ in (41). Notice that (A.26) is automatically satisfied and that (A.27) becomes

$$\mathbf{E}[\theta\,|\,\theta>t_{LH}^{**}]=t_H^a.$$

Hence we must have $t_{LH}^{**}=t_H^*$. That is, type L always consumes and type H is facing his individually optimal incentive-compatible mechanism.

Subcase 2.2 If $h^{**}(\theta)<0$ for all $\theta\in(t_{LH}^{**},1)$, then the objective in (A.25) is uniquely (up to a **P**-null set) maximized over S by

$$\pi_L^{**}(\theta)=0,$$

which corresponds to a cutoff $t_L^{**}=t_{LH}^{**}$ in (41). Notice that (A.27) is automatically satisfied, and that (A.26) becomes

$$\lambda^{**}\{\mathbf{E}[\theta\,|\,\theta>t_{LH}^{**}]-t_L^a\}=0.$$

Hence we must have $t_{LH}^{**}=t_L^*$ if $\lambda^{**}>0$. That is, both types are facing the individually optimal incentive-compatible mechanism for type L.

Subcase 2.3 Suppose that h^{**} changes sign over $(t_{LH}^{**},1)$—so that we have, in particular, $\lambda^{**}+\mu^{**}\neq 0$—at

$$t_L^{**}\equiv\frac{\eta^{**}(t_L^h-t_H^a)+\lambda^{**}t_L^a+\mu^{**}t_H^a}{\lambda^{**}+\mu^{**}}.$$

We claim that $\lambda^{**}+\mu^{**}>0$. Indeed, if $\lambda^{**}+\mu^{**}<0$, then the objective in (A.25) is uniquely (up to a **P**-null set) maximized over S by

$$\pi_L^{**}(\theta)=1_{\{\theta\geq t_L^{**}\}}, \tag{A.28}$$

so that

$$\pi_0^{**}(\theta)=1_{\{t_{LH}^{**}<\theta<t_L^{**}\}}. \tag{A.29}$$

Now, given (A.29), (36) requires

$$\mathbf{E}[\theta\,|\,t_{LH}^{**}<\theta<t_L^{**}]\geq t_L^a. \tag{A.30}$$

However, we know from Lemma A.3 that any solution to (A.16) and, hence, to (A.19) and (A.22), is also a solution to (40). In particular, given (A.28), (37) requires

$$\mathbf{E}[\theta\,|\,\theta\geq t_L^{**}]<t_L^a. \tag{A.31}$$

Because (A.30)–(A.31) contradict each other, we obtain $\lambda^{**}+\mu^{**}>0$, as claimed, and the objective in (A.25) is uniquely (up to a **P**-null set) maximized over S by

$$\pi_L^{**}(\theta)=1_{\{t_{LH}^{**}<\theta\leq t_L^{**}\}},$$

once again in line with (41).

Subcase 2.4 Suppose finally that h^{**} is identically zero over $(t_{LH}^{**},1)$—so that we have, in particular, $\lambda^{**}+\mu^{**}=0$. Then

$$\eta^{**}(t_L^h-t_H^a)+\lambda^{**}(t_L^a-t_H^a)=0.$$

Because $t_L^a>t_H^a$, we have $\eta^{**}=1$ by (A.24); otherwise, by (A.24) again, $\eta^{**}=\lambda^{**}=\mu^{**}=0$, which violates (A.23). Applying (A.24) yet again, we obtain $t_H^a\geq t_L^h$, with equality if and only if $\lambda^{**}=0$. By Subcases 2.1–3, this completes the proof in case Assumption 2 holds. Suppose then that Assumption 2 does not hold, and consider first the case $t_H^a>t_L^h$. Then $\lambda^{**}>0$ and, by (A.26), any solution π_L^{**} to (A.19) must satisfy (A.27) and

$$\int_{t_{LH}^{**}}^1\theta[1-\pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta=t_L^a\int_{t_{LH}^{**}}^1[1-\pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta. \tag{A.32}$$

Notice that, because Lemma A.5 guarantees that a solution π_L^{**} to (A.19) exists, there exists a solution to (A.27) and (A.32). Conversely, because h^{**} is identically zero over $(t_{LH}^{**},1)$, any solution to (A.27) and (A.32) is a solution to the maximization condition (A.25) and

hence to (A.19) as this is a convex problem and $\eta^{**} > 0$ (Clarke (2013, Exercise 9.7)). Let us then fix a solution π_L^{**} to (A.27) and (A.32). We focus with no loss of generality on the case where π_L^{**} is not equal to 1 or to 0, $\mathbf{P}$-almost surely over $(t_{LH}^{**}, 1)$; otherwise, we are back to Subcases 2.1 or 2.2 as above. That is, we focus on the case where both constraints (36) and (38) in (A.16) are well defined and binding. In particular, we must have

$$t_{LH}^{**} < t_H^a < \mathbf{E}[\theta \,|\, \theta > t_{LH}^{**}] < t_L^a. \tag{A.33}$$

Summing (A.27) and (A.32) and rearranging, we obtain that any solution to (A.27) and (A.32) satisfies

$$\int_{t_{LH}^{**}}^1 [1 - \pi_L^{**}(\theta)] f(\theta)\,\mathrm{d}\theta = \rho \equiv \frac{\mathbf{E}[\theta \,|\, \theta > t_{LH}^{**}] - t_H^a}{t_L^a - t_H^a}\,[1 - F(t_{LH}^{**})] < 1 - F(t_{LH}^{**}). \tag{A.34}$$

We claim that, in line with (42), there exists a solution to (A.27) and (A.32) of the form

$$\pi_L^{**}(\theta) = 1_{\{t_{LH}^{**} < \theta \leq \underline{t}_L^{**}\}} + 1_{\{\theta > \bar{t}_L^{**}\}}$$

for some cutoffs $\bar{t}_L^{**} > \underline{t}_L^{**} > t_{LH}^{**}$. To prove this claim, we show that the system in $(\underline{t}, \bar{t})$

$$\int_{\underline{t}}^{\bar{t}} \theta f(\theta)\,\mathrm{d}\theta = t_L^a [F(\bar{t}) - F(\underline{t})] \tag{A.35}$$

$$\int_{t_{LH}^{**}}^{\underline{t}} \theta f(\theta)\,\mathrm{d}\theta + \int_{\bar{t}}^1 \theta f(\theta)\,\mathrm{d}\theta = t_H^a [F(\underline{t}) - F(t_{LH}^{**}) + 1 - F(\bar{t})], \tag{A.36}$$

has a unique solution. As above, summing (A.35)–(A.36) yields

$$F(\bar{t}) - F(\underline{t}) = \rho, \tag{A.37}$$

and hence (A.35) rewrites as

$$\psi(\underline{t}) \equiv \frac{\int_{\underline{t}}^{F^{-1}(F(\underline{t})+\rho)} \theta f(\theta)\,\mathrm{d}\theta}{\rho} = \mathbf{E}[\theta \,|\, \underline{t} < \theta \leq F^{-1}(F(\underline{t}) + \rho)] = t_L^a,$$

which we must solve for $\underline{t} \in (t_{LH}^{**}, F^{-1}(1-\rho)]$. By the intermediate value theorem, we only need to check that $\psi(t_{LH}^{**}) < t_L^a$, that ψ is strictly increasing over $(t_{LH}^{**}, F^{-1}(1-\rho)]$, and that $\psi(F^{-1}(1-\rho)) \geq t_L^a$. The first statement follows from

$$\psi(t_{LH}^{**}) = \mathbf{E}[\theta \,|\, t_{LH}^{**} < \theta \leq F^{-1}(F(t_{LH}^{**}) + \rho)] < \mathbf{E}[\theta \,|\, \theta > t_{LH}^{**}] < t_L^a,$$

where the first inequality follows from the fact that $F(t_{LH}^{**}) + \rho < 1$ by (A.37) and that $\mathbf{P}$ has full support over $[0, 1]$, and the second inequality follows from (A.33). The second statement

follows from a straightforward computation,

$$\psi'(\underline{t}) = \frac{f(\underline{t})[F^{-1}(F(\underline{t}) + \rho) - \underline{t}]}{\rho} > 0.$$

The third statement amounts to

$$\frac{\int_{F^{-1}(1-\rho)}^{1} \theta f(\theta)\,\mathrm{d}\theta}{\rho} \geq t_L^a. \tag{A.38}$$

But we know that there exists a solution to (A.27) and (A.32), which satisfies

$$t_L^a = \frac{\int_{t_{LH}^{**}}^{1} \theta[1 - \pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta}{\int_{t_{LH}^{**}}^{1} [1 - \pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta} = \frac{\int_{t_{LH}^{**}}^{1} \theta[1 - \pi_L^{**}(\theta)]f(\theta)\,\mathrm{d}\theta}{\rho}$$

by (A.34), and clearly

$$\int_{F^{-1}(1-\rho)}^{1} \theta f(\theta)\,\mathrm{d}\theta = \max\left\{\int_{t_{LH}^{**}}^{1} \theta[1 - \pi_L(\theta)]f(\theta)\,\mathrm{d}\theta : \int_{t_{LH}^{**}}^{1} [1 - \pi_L(\theta)]f(\theta)\,\mathrm{d}\theta = \rho\right\},$$

which yields the desired inequality (A.38). The claim follows. In case (A.38) holds as an equality, we have $\bar{t}_L^{**} = 1$, and π_L^{**} has the same form as in Subcase 2.3.

The proof for the limiting case $t_H^a = t_L^h$ or, equivalently, $\beta_H = \tilde{\beta}_H(\beta_L)$, relies on a simple continuity argument. From the proof of Lemma A.5, for each $\beta_H \geq \tilde{\beta}_H(\beta_L)$, any solution to (A.16) for β_H can be represented by a pair $(t_{LH}^{**}(\beta_H), \pi_L^{**}(\beta_H)) \in [0,1] \times B_{L_\infty(\mathbf{P})}$. Consider a strictly decreasing sequence $(\beta_{H,n})_{n\in\mathbb{N}}$ converging to $\tilde{\beta}_H(\beta_L)$. By Berge maximum theorem (Aliprantis and Border (2006, Theorem 17.31)) along with the fact that $B_{L_\infty(\mathbf{P})}$ is metrizable as $L_1(\mathbf{P})$ is separable (Aliprantis and Border (2006, Theorems 6.30 and 13.16)), any sequence $((t_{LH}^{**}(\beta_{H,n}), \pi_L^{**}(\beta_{H,n})))_{n\in\mathbb{N}}$ of solutions to (A.16) for each term of the sequence $(\beta_{H,n})_{n\in\mathbb{N}}$ has a subsequence that converges in $[0,1] \times B_{L_\infty(\mathbf{P})}$ to a solution $(t_{LH}^{**}(\tilde{\beta}_H(\beta_L)), \pi_L^{**}(\tilde{\beta}_H(\beta_L)))$ to (A.16) for $\tilde{\beta}_H(\beta_L)$. We can with no loss of generality assume that this sequence converges. For each $n \in \mathbb{N}$, we have $\beta_{H,n} > \tilde{\beta}_H(\beta_L)$ and hence

$$\pi_L^{**}(\beta_{H,n})(\theta) = 1_{\{t_{LH}^{**}(\beta_{H,n}) < \theta \leq t_L^{**}(\beta_{H,n})\}} \tag{A.39}$$

by Subcases 2.1–3. Therefore,

$$\begin{aligned}
\int \pi_L^{**}(\tilde{\beta}_H(\beta_L))(\theta)\,\mathbf{P}(\mathrm{d}\theta) &= \lim_{n\to\infty} \int \pi_L^{**}(\beta_{H,n})(\theta)\,\mathbf{P}(\mathrm{d}\theta) \\
&= \lim_{n\to\infty} F(t_L^{**}(\beta_{H,n})) - F(t_{LH}^{**}(\beta_{H,n})) \\
&= \lim_{n\to\infty} F(t_L^{**}(\beta_{H,n})) - F(t_{LH}^{**}(\tilde{\beta}_H(\beta_L))),
\end{aligned} \tag{A.40}$$

where the first equality follows from the fact that the sequence $(\pi_L^{**}(\beta_{H,n}))_{n\in\mathbb{N}}$ converges to $\pi_L^{**}(\tilde{\beta}_H(\beta_L))$ in $B_{L_\infty(\mathbf{P})}$, using the definition of the weak* topology $\sigma(L_\infty(\mathbf{P}), L_1(\mathbf{P}))$, the

second equality follows from (A.39), and the third inequality follows from the fact that the sequence $(t_{LH}^{**}(\beta_{H,n}))_{n\in\mathbb{N}}$ converges to $t_{LH}^{**}(\tilde{\beta}_H(\beta_L))$ in $[0,1]$ and that F is continuous as $\mathbf{P}$ is nonatomic. Because F is strictly increasing as $\mathbf{P}$ has full support, (A.40) implies that the sequence $(t_L^{**}(\beta_{H,n}))_{n\in\mathbb{N}}$ converges to some limit t_∞. To complete the proof, notice that, for any Borel subset A of $[0,1]$,

$$\begin{aligned}\int_A \pi_L^{**}(\tilde{\beta}_H(\beta_L))(\theta)\,\mathbf{P}(\mathrm{d}\theta) &= \lim_{n\to\infty}\int_A \pi_L^{**}(\beta_{H,n})(\theta)\,\mathbf{P}(\mathrm{d}\theta)\\ &= \lim_{n\to\infty}\mathbf{P}[A\cap(t_{LH}^{**}(\beta_{H,n}),t_L^{**}(\beta_{H,n})]],\end{aligned}\tag{A.41}$$

using again the definition of the weak* topology $\sigma(L_\infty(\mathbf{P}),L_1(\mathbf{P}))$ along with (A.39). Finally, we can substitute $A=(t_{LH}^{**}(\tilde{\beta}_H(\beta_L),t_\infty]$ and $A=(t_\infty,1]$ in (A.41) and use the fact that the sequence $((t_{LH}^{**}(\beta_{H,n}),t_L^{**}(\beta_{H,n})))_{n\in\mathbb{N}}$ converges to $(t_{LH}^{**}(\tilde{\beta}_H(\beta_L),t_\infty)$ to conclude that in fact $t_\infty=t_L^{**}(\tilde{\beta}_H(\beta_L))$ and

$$\pi_L^{**}(\tilde{\beta}_H(\beta_L))(\theta)=1_{\{t_{LH}^{**}(\tilde{\beta}_H(\beta_L))<\theta\le t_L^{**}(\tilde{\beta}_H(\beta_L))\}}$$

up to a $\mathbf{P}$-null set. The result follows. ■

Proposition 4 is then an immediate consequence of Lemma A.6. Hence the result. ■

Proof of Lemma 3. We solve (A.16) for the optimal cutoffs (t_{LH}^{**},t_L^{**})—the existence of which we established in Proposition 4—under the assumption that the individually optimal mechanisms with cutoffs (t_H^*,t_L^*) are not simultaneously implementable. We first claim that we can restrict attention to cutoffs (t_{LH},t_L) such that $t_L\ge t_L^*$. We distinguish two cases. If $t_L^*>t_L^h$, then (36) is satisfied if and only if $t_L\ge t_L^*$. If $t_L^*=t_L^h$, then, for any given t_{LH}, any cutoff $t_L<t_L^h$ would induce an inefficiently high rate of abstinence for type L and would tighten (38) compared to $t_L=t_L^h$; hence an optimal cutoff t_L must satisfy $t_L\ge t_L^h$, which is incentive compatible as $t_L^h=t_L^*$. The claim follows. Replacing expectations in (A.16) by integrals then yields the equivalent problem

$$\max\left\{p_L\beta_L\int_0^{t_L}(t_L^h-\theta)f(\theta)\,\mathrm{d}\theta+p_H\beta_H\int_0^{t_{LH}}(t_H^h-\theta)f(\theta)\,\mathrm{d}\theta\right\},\tag{A.42}$$

subject to the constraints

$$\int_{t_{LH}}^{t_L}(\theta-t_H^a)f(\theta)\,\mathrm{d}\theta\ge 0,\tag{A.43}$$

$$t_L-t_L^*\ge 0,\tag{A.44}$$

$$1-t_L\ge 0.\tag{A.45}$$

The objective in (A.42) is continuous in (t_{LH},t_L) and the feasible set defined by $(t_{LH},t_L)\in$

$[0,1]^2$ and (A.43)–(A.45) is compact. Hence problem (A.42)–(A.45) has a solution (t_{LH}^{**}, t_L^{**}). The proof consists of four steps.

Step 1 We first show that $t_L^{**} > t_H^a > t_{LH}^{**} \geq t_H^h$ in any solution (t_{LH}^{**}, t_L^{**}) to (A.42)–(A.45). That $t_L^{**} > t_H^a$ follows from our preliminary observation that $t_L \geq t_L^h$ along with Assumption 2. As for t_{LH}^{**}, suppose, by way of contradiction, that $t_{LH}^{**} \geq t_H^a$. Because $t_L^{**} > t_H^a$, we have

$$\int_{t_H^a}^{t_L^{**}} (\theta - t_H^a) f(\theta) \, \mathrm{d}\theta > 0.$$

Hence lowering t_{LH}^{**} to a value $t_H^a - \varepsilon$ for some small $\varepsilon > 0$ would preserve (A.43) and strictly increase the objective in (A.42), a contradiction. Thus $t_H^a > t_{LH}^{**}$, as claimed. The proof that $t_{LH}^{**} \geq t_H^h$ is similar, observing that the left-hand side of (A.43) is strictly increasing in $t_{LH} \in [0, t_H^a]$ and the objective in (A.42) is strictly increasing in $t_{LH} \in [0, t_H^h]$.

Step 2 We next verify that constraints (A.43)–(A.45) satisfy the Mangasarian–Fromovitz qualification conditions at (t_{LH}^{**}, t_L^{**}) (Mangasarian (1969, 11.3.5)). Letting g be the mapping defined by the left-hand sides of the binding constraints at (t_{LH}^{**}, t_L^{**}), we must prove that $\nabla g(t_{LH}^{**}, t_L^{**}) z^T > 0$ has a solution $z \in \mathbb{R}^2$, where $\nabla g(t_{LH}^{**}, t_L^{**})$ is the Jacobian matrix of g at (t_{LH}^{**}, t_L^{**}). This is obvious if (A.43) is not binding. If (A.43) is binding, then the first line of $\nabla g(t_{LH}^{**}, t_L^{**})$ is

$$\mathrm{D}g_1(t_{LH}^{**}, t_L^{**}) \equiv \big((t_H^a - t_{LH}^{**}) f(t_{LH}^{**}) \quad (t_L^* - t_H^a) f(t_L^{**}) \big).$$

We shall exploit the fact that f is strictly positive over $(0,1)$. Notice first that, because $t_H^a > t_{LH}^{**} \geq t_H^h$ by Step 1, we always have $(t_H^a - t_{LH}^{**}) f(t_{LH}^{**}) > 0$. If only (A.43) is binding, then $1 > t_L^{**} > t_H^a$ by Step 1, so that $(t_L^* - t_H^a) f(t_L^{**}) > 0$ and

$$\nabla g(t_{LH}^{**}, t_L^{**}) = \mathrm{D}g_1(t_{LH}^{**}, t_L^{**}).$$

We can then take any $z \in \mathbb{R}^2_{++}$. Next, if (A.43) and (A.44) are binding, then $t_L^{**} = t_L^*$, so that $(t_L^* - t_H^a) f(t_L^{**}) > 0$ and

$$\nabla g(t_{LH}^{**}, t_L^{**}) = \begin{pmatrix} (t_H^a - t_{LH}^{**}) f(t_{LH}^{**}) & (t_L^* - t_H^a) f(t_L^{**}) \\ 0 & 1 \end{pmatrix}.$$

We can then take any $z \in \mathbb{R}^2_{++}$. Finally, if (A.43) and (A.45) are binding, then it is optimal to have $t_{LH}^{**} = t_H^*$ by Propositions 1–2, and

$$\nabla g(t_{LH}^{**}, t_L^{**}) = \begin{pmatrix} (t_H^a - t_{LH}^{**}) f(t_{LH}^{**}) & (t_L^* - t_H^a) f(t_L^{**}) \\ 0 & -1 \end{pmatrix}.$$

We can then take $z = (1, \varepsilon)$ for some small enough $\varepsilon < 0$.

Step 3 According to Step 1, constraints (A.43)–(A.45) are qualified at any solution (t_{LH}^{**}, t_L^{**}) to (A.42)–(A.45). Therefore, by the Kuhn–Tucker necessary optimality conditions for nonconvex optimization problems (Mangasarian (1969, 11.3.6)), there exists a vector of Lagrange multipliers $(\zeta^{**}, \nu^{**}, \chi^{**})$ such that we have:

- Positivity:

$$(\zeta^{**}, \nu^{**}, \chi^{**}) \in \mathbb{R}^3_+. \tag{A.46}$$

- First-order conditions:

$$p_L \beta_L (t_L^h - t_L^{**}) f(t_L^{**}) + \zeta^{**}(t_L^{**} - t_H^a) f(t_L^{**}) + \nu^{**} - \chi^{**} = 0, \tag{A.47}$$

$$p_H \beta_H (t_H^h - t_{LH}^{**}) f(t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) f(t_{LH}^{**}) = 0. \tag{A.48}$$

- Complementary slackness:

$$\zeta^{**} \int_{t_{LH}^{**}}^{t_L^{**}} (\theta - t_H^a) f(\theta) \, \mathrm{d}\theta = 0, \tag{A.49}$$

$$\nu^{**}(t_L^{**} - t_L^*) = 0, \tag{A.50}$$

$$\chi^{**}(1 - t_L^{**}) = 0. \tag{A.51}$$

We distinguish three cases.

Case 1 Suppose first that (A.44) is binding, so that $t_L^{**} = t_L^*$ and $\chi^{**} = 0$ by (A.51), and suppose further, by way of contradiction, that $\zeta^{**} = 0$. Then, by (A.48) along with the fact that $f(t_{LH}^{**}) > 0$ as $t_H^a > t_{LH}^{**} \geq t_H^h$ by Step 1 and f is strictly positive over $(0, 1)$, we must have $t_{LH}^{**} = t_H^h$. Therefore, using the assumption that the individually optimal mechanisms with cutoffs t_H^* and t_L^* are not simultaneously implementable, we obtain that

$$\mathbf{E}[\theta \,|\, t_{LH}^{**} < \theta \leq t_L^*] \leq \mathbf{E}[\theta \,|\, t_H^* < \theta \leq t_L^*] < t_H^a.$$

But then (A.43) is violated at (t_{LH}^{**}, t_L^*), a contradiction. Hence, by (A.46), $\zeta^{**} > 0$, so that, by (A.49), (A.43) must be binding at (t_{LH}^{**}, t_L^*). That is, t_{LH}^{**} must satisfy

$$\int_{t_{LH}^{**}}^{t_L^*} (\theta - t_H^a) f(\theta) \, \mathrm{d}\theta = 0. \tag{A.52}$$

Because f is strictly positive over $(0, 1)$, we have $f(t_L^*) > 0$; moreover, as argued above, $f(t_{LH}^{**}) > 0$. Because $\chi^{**} = 0 \leq \nu^{**}$ by (A.46), the first-order conditions (A.47)–(A.48) rewrite as

$$p_L \beta_L (t_L^h - t_L^*) + \zeta^{**}(t_L^* - t_H^a) \leq 0, \tag{A.53}$$

$$p_H\beta_H(t_H^h - t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) = 0. \tag{A.54}$$

Because $\zeta^{**} > 0$ and $t_L^* \geq t_L^h > t_H^a$, (A.53) implies $t_L^* > t_L^h$. Hence the bracketed terms in (A.53) are different from zero. Moreover, because the bracketed terms in (A.54) cannot simultaneously be zero, none of them can be zero. We can thus divide (A.54) by (A.53), which yields

$$\frac{p_H\beta_H}{p_L\beta_L}\,\frac{t_{LH}^{**} - t_H^h}{t_L^* - t_L^h} \leq \frac{t_H^a - t_{LH}^{**}}{t_L^* - t_H^a}. \tag{A.55}$$

Case 2 Suppose next that (A.45) is binding, so that $t_L^{**} = 1$ and $\nu^{**} = 0$ by (A.50). By Propositions 1–2, it is then optimal to have $t_{LH}^{**} = t_H^*$. Because f is strictly positive over $(0, 1)$, we have $f(t_H^*) > 0$. The first-order condition (A.48) then rewrites as

$$p_H\beta_H(t_H^h - t_H^*) - \zeta^{**}(t_H^* - t_H^a) = 0, \tag{A.56}$$

so that $t_H^* > t_H^h$ if and only if $\zeta^{**} > 0$. If $f(1) > 0$, then, because $\chi^{**} \geq 0 = \nu^{**}$ by (A.46), we can also simplify (A.47) to obtain

$$p_L\beta_L(t_L^h - 1) + \zeta^{**}(1 - t_H^a) \geq 0. \tag{A.57}$$

The argument leading to (A.57) is a bit more involved if $f(1) = 0$. In that case, it follows from (A.47) and $\nu^{**} = 0$ that $\chi^{**} = 0$ as well. Hence the relevant part of the Lagrangian, to be maximized with respect to t_L, can be written as

$$\int_{t_H^*}^{t_L} [p_L\beta_L(t_L^h - \theta) + \zeta^{**}(\theta - t_H^a)] f(\theta)\,\mathrm{d}\theta,$$

which, as f is strictly positive over $(0, 1)$, is maximum for $t_L = 1$ only if (A.57) holds. By (A.46) and (A.57), $\zeta^{**} > 0$, so that, by (A.49), (A.43) must be binding at $(t_H^*, 1)$. That is, t_H^* must satisfy

$$\int_{t_H^*}^{1} (\theta - t_H^a) f(\theta)\,\mathrm{d}\theta = 0, \tag{A.58}$$

which generically implies that $t_H^* > t_H^h$, so that the unconstrained-optimal mechanism for type H is not incentive-compatible. The terms $t_H^* - t_H^a$ and $1 - t_H^a$ in (A.56)–(A.57) are by construction different from zero. We can thus divide (A.56) by (A.57), which yields

$$\frac{p_H\beta_H}{p_L\beta_L}\,\frac{t_H^* - t_H^h}{1 - t_L^h} \geq \frac{t_H^a - t_H^*}{1 - t_H^a}. \tag{A.59}$$

Case 3 Suppose finally that (A.44)–(A.45) are not binding, so that $\nu^{**} = \chi^{**} = 0$ by

(A.50)–(A.51). As f is strictly positive over $(0,1)$, we have $f(t_L^{**}) > 0$ and, as argued in Case 1, $f(t_{LH}^{**}) > 0$. The first-order conditions (A.47)–(A.48) then rewrite as

$$p_L\beta_L(t_L^h - t_L^{**}) + \zeta^{**}(t_L^{**} - t_H^a) = 0, \tag{A.60}$$

$$p_H\beta_H(t_H^h - t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) = 0. \tag{A.61}$$

We must have $\zeta^{**} > 0$, and hence, by (A.49), (A.43) must be binding, for, otherwise, the individually unconstrained-optimal mechanisms for types H and L would be simultaneously implementable. That is, (t_{LH}^{**}, t_L^{**}) must satisfy

$$\int_{t_{LH}^{**}}^{t_L^{**}} (\theta - t_H^a) f(\theta)\,\mathrm{d}\theta = 0. \tag{A.62}$$

Because, as a result, the terms on the left- and the right-hand sides in each of (A.60)–(A.61) cannot simultaneously be zero, none of them can be zero. Dividing yields

$$\frac{p_H\beta_H}{p_L\beta_L}\,\frac{t_{LH}^{**} - t_H^h}{t_L^{**} - t_L^h} = \frac{t_H^a - t_{LH}^{**}}{t_L^{**} - t_H^a}. \tag{A.63}$$

Step 4 To complete the proof, we only need to delineate the circumstances under which each of the cases discussed in Step 3 arises. In each case, (A.43) is binding, see (A.52), (A.58), and (A.62). Let accordingly

$$\mathcal{T}_L \equiv \{t_L \geq t_L^* : \text{there exists } t_H \leq t_L \text{ such that } \mathbf{E}[\theta \,|\, t_H < \theta \leq t_L] = t_H^a\}. \tag{A.64}$$

Because $t_L^* > t_H^a$ and $\mathbf{E}[\theta \,|\, t_H^* < \theta \leq t_L^*] < t_H^a$ as the individually optimal mechanisms with cutoffs t_H^* and t_L^* are not simultaneously implementable, $t_L^* \in \mathcal{T}_L$. Because $\mathbf{E}[\theta \,|\, t_H < \theta \leq t_L]$ is strictly increasing in t_H and t_L, $\mathcal{T}_L$ is thus an interval $[t_L^*, \sup \mathcal{T}_L]$, and there exists a unique strictly decreasing function $\hat{t}_{LH} : \mathcal{T}_L \to [0, t_H^a)$ implicitly defined by

$$\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L] = t_H^a. \tag{A.65}$$

By (A.52), (A.58), and (A.62), given t_L^{**}, t_{LH}^{**} is uniquely pinned down by

$$t_{LH}^{**} = \hat{t}_{LH}(t_L^{**}). \tag{A.66}$$

As f is strictly positive over $(0,1)$, a straightforward application of the implicit function theorem implies that $\hat{t}_{LH}$ is differentiable over the interior of $\mathcal{T}_L$, with

$$\hat{t}'_{LH}(t_L) = -\,\frac{f(t_L)}{f(\hat{t}_{LH}(t_L))}\,\frac{t_L - \mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L]}{\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L] - \hat{t}_{LH}(t_L)} < 0. \tag{A.67}$$

While (A.66) holds in each of Cases 1, 2, and 3, these cases differ as to whether (A.55),

(A.70), or (A.63) holds. Defining accordingly

$$\kappa(t_L) \equiv \frac{p_H \beta_H}{p_L \beta_L} \frac{\hat{t}_{LH}(t_L) - t_H^h}{t_L - t_L^h} - \frac{t_H^a - \hat{t}_{LH}(t_L)}{t_L - t_H^a}, \tag{A.68}$$

we have $\kappa(t_L^*) \leq 0$, $\kappa(1) \geq 0$, and $\kappa(t_L^{**}) = 0$ in Cases 1, 2, and 3, respectively. To conclude, we only need to show that these cases are mutually exclusive. For this, we only need to show that κ single-crosses zero, from above. Indeed, if $\kappa(t_L) = 0$, then

$$\begin{aligned}
\kappa'(t_L) &= \frac{p_H \beta_H}{p_L \beta_L} \left[\frac{\hat{t}'_{LH}(t_L)}{t_L - t_L^h} - \frac{\hat{t}_{LH}(t_L) - t_H^h}{(t_L - t_L^h)^2} \right] + \frac{\hat{t}'_{LH}(t_L)}{t_L - t_H^a} + \frac{t_H^a - \hat{t}_{LH}(t_L)}{(t_L - t_H^a)^2} \\
&< -\frac{p_H \beta_H}{p_L \beta_L} \frac{\hat{t}_{LH}(t_L) - t_H^h}{(t_L - t_L^h)^2} + \frac{t_H^a - \hat{t}_{LH}(t_L)}{(t_L - t_H^a)^2} \\
&= \frac{[t_H^a - \hat{t}_{LH}(t_L)](t_H^a - t_L^h)}{(t_L - t_L^h)(t_L - t_H^a)^2} \\
&< 0,
\end{aligned} \tag{A.69}$$

where the first inequality follows from (A.67), the second equality follows from (A.68) along with $\kappa(t_L) = 0$, and the second inequality follows from Assumption 2. Thus Case 1 occurs if and only if $\kappa(t_L^*) \leq 0$, so that $\kappa(t_L) < 0$ for all $t_L > t_L^*$, Case 2 occurs if and only if $\kappa(1) \geq 0$, so that $\kappa(t_L) > 0$ for all $t_L < 1$, and Case 3 occurs if and only if $\kappa(t_L^*) > 0$ and $\kappa(1) < 0$, so that $\kappa(t_L)$ changes sign from positive to negative only at $t_L = t_L^{**}$. The result follows. ■

Proof of Corollary 5. The proof consists of three steps.

Step 1 Consider first the boundary $\underline{p}$, starting with the case $t_L^* > t_L^h$. Define the function $\hat{t}_{LH}$ as in (A.65). By Assumption 2, $t_L^* > t_H^a$, and, by construction, $\hat{t}_{LH}(t_L^*) < t_H^a$. Moreover, because the individually optimal mechanisms with cutoffs t_H^* and t_L^* are not simultaneously implementable, $\hat{t}_{LH}(t_L^*) > t_H^*$ and thus $\hat{t}_{LH}(t_L^*) > t_H^h$. Hence

$$\frac{\beta_H}{\beta_L} \frac{\hat{t}_{LH}(t_L^*) - t_H^h}{t_L^* - t_L^h} > 0 \text{ and } \frac{t_H^a - \hat{t}_{LH}(t_L^*)}{t_L^* - t_H^a} > 0.$$

As $p \mapsto p/(1-p)$ is a strictly increasing continuous mapping between $(0,1)$ and $(0,\infty)$, there exists a unique $\underline{p} \in (0,1)$ such that

$$\frac{\underline{p} \beta_H}{(1-\underline{p})\beta_L} \frac{\hat{t}_{LH}(t_L^*) - t_H^h}{t_L^* - t_L^h} = \frac{t_H^a - \hat{t}_{LH}(t_L^*)}{t_L^* - t_H^a},$$

so that

$$\frac{p_H \beta_H}{p_L \beta_L} \frac{\hat{t}_{LH}(t_L^*) - t_H^h}{t_L^* - t_L^h} \leq \frac{t_H^a - \hat{t}_{LH}(t_L^*)}{t_L^* - t_H^a}$$

if and only if $p_H \in [0, \underline{p}]$. Defining κ as in (A.68), we thus have $\kappa(t_L^*) \leq 0$ for any such p_H. It then follows from Step 4 of the proof of Lemma 3 that $(t_{LH}^{**}, t_L^{**}) = (\hat{t}_{LH}(t_L^*), t_L^*)$. We have thus proven that, if $t_L^* > t_L^h$, there exists $\underline{p} \in (0,1)$ such that, for all $p_H \in (0, \underline{p}]$, type L faces his individually optimal incentive-compatible mechanism. To complete the proof, we only need to check that if $t_L^* = t_L^h$ and type L faces his individually optimal incentive-compatible mechanism, so that $t_L^{**} = t_L^* = t_L^h$, then it must be that $p_H = 0$, in which case we can set $\underline{p} \equiv 0$ by convention. Indeed, from (A.53) in Case 1 of the proof of Lemma 3, if we impose the constraint (A.43), which is relevant only if $p_H > 0$, then $\zeta^{**} > 0$, and $t_L^{**} = t_L^*$ implies $t_L^* > t_L^h$. Thus $t_L^{**} = t_L^* = t_L^h$ implies $p_H = 0$, as desired.

Step 2 Consider next the boundary $\overline{p}$, starting with the case $t_H^* > t_H^h$. Then

$$\frac{\beta_H}{\beta_L} \frac{t_H^* - t_H^h}{1 - t_L^h} > 0 \quad \text{and} \quad \frac{t_H^a - t_H^*}{1 - t_H^a} > 0.$$

As $p \mapsto p/(1-p)$ is a strictly increasing continuous mapping between $(0,1)$ and $(0,\infty)$, there exists a unique $\overline{p} \in (0,1)$ such that

$$\frac{\overline{p}\beta_H}{(1-\overline{p})\beta_L} \frac{t_H^* - t_H^h}{1 - t_L^h} = \frac{t_H^a - t_H^*}{1 - t_H^a},$$

so that

$$\frac{p_H\beta_H}{p_L\beta_L} \frac{t_H^* - t_H^h}{1 - t_L^h} \geq \frac{t_H^a - t_H^*}{1 - t_H^a}$$

if and only if $p_H \in [\overline{p}, 1]$. Defining κ as in (A.68), we thus have $\kappa(1) \geq 0$ for any such p_H. It then follows from Step 4 of the proof of Lemma 3 that $(t_{LH}^{**}, t_L^{**}) = (t_H^*, 1)$. We have thus proven that, if $t_H^* > t_H^h$, there exists $\overline{p} \in (0,1)$ such that, for all $p_H \in [\overline{p}, 1)$, type H faces his individually optimal incentive-compatible mechanism. To complete the proof, we only need to check that if $t_H^* = t_H^h$ and type H faces his individually optimal incentive-compatible mechanism, so that $t_{LH}^{**} = t_H^* = t_H^h$, then it must be that $p_H = 1$, in which case we can set $\overline{p} \equiv 1$ by convention. Indeed, from (A.56) in Case 2 of the proof of Lemma 3, $t_H^* = t_H^h$ implies $\zeta^{**} = 0$. Because $t_{LH}^{**} = t_H^*$ implies $t_L^{**} = 1$, (A.57) implies $p_L = 0$, as desired.

Step 3 According to Steps 1–2,

$$\frac{p_H\beta_H}{p_L\beta_L} \frac{\hat{t}_{LH}(t_L^*) - t_H^h}{t_L^* - t_L^h} > \frac{t_H^a - \hat{t}_{LH}(t_L^*)}{t_L^* - t_H^a} \quad \text{and} \quad \frac{p_H\beta_H}{p_L\beta_L} \frac{t_H^* - t_H^h}{1 - t_L^h} < \frac{t_H^a - t_H^*}{1 - t_H^a}$$

if and only if $p_H \in (\underline{p}, \overline{p})$. Defining κ as in (A.68), we thus have

$$\kappa(p_H, t_L^{**}) = \frac{p_H\beta_H}{(1-p_H)\beta_L} \frac{\hat{t}_{LH}(t_L^{**}) - t_H^h}{t_L^{**} - t_L^h} - \frac{t_H^a - \hat{t}_{LH}(t_L^{**})}{t_L^{**} - t_H^a} = 0 \tag{A.70}$$

for any such p_H, where we make the dependence of κ on p_H explicit. It then follows from Step 4 of the proof of Lemma 3 that (t_{LH}^{**}, t_L^{**}) is the unique solution to (44). Let us accordingly denote by $\hat{t}_L(p_H)$ the unique solution to (A.70). We clearly have $(\partial\kappa/\partial p_H)(p_H, t_L) > 0$ and, from (A.69), $(\partial\kappa/\partial p_H)(p_H, t_L) < 0$ if $\kappa(p_H, t_L) = 0$. A straightforward application of the implicit function theorem then implies that $\hat{t}_L$ is differentiable over $(\underline{p}, \overline{p})$, with $\hat{t}_L' > 0$. Summarizing, because

$$(t_{LH}^{**}, t_L^{**}) = (\hat{t}_{LH}(\hat{t}_L(p_H)), \hat{t}_L(p_H))$$

for all $p_H \in (\underline{p}, \overline{p})$, where $\hat{t}_{LH}$ is strictly decreasing over $\mathcal{T}_L$ by (A.67), the probabilities of consumption $F(\hat{t}_{LH}(\hat{t}_L(p_H)))$ and $F(\hat{t}_L(p_H))$ of type H and type L are strictly decreasing and strictly increasing in $p_H \in (\underline{p}, \overline{p})$, respectively. Hence the result. ■.

References

[1] Aliprantis, C.D., and K.C. Border (2006): *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, Berlin, Heidelberg, New York: Springer.

[2] Bryson, M.C., and M.M. Siddiqui (1969): "Some Criteria for Aging," *Journal of the American Statistical Association*, 64(328), 1472–1483.

[3] Clarke, F. (2013): *Functional Analysis*, *Calculus of Variations and Optimal Control*, London: Springer-Verlag.

[4] Giorgi, G., and S. Komlósi (1992): "Dini Derivatives in Optimization–Part I," *Decisions in Economics and Finance*, 15(1), 3–30.

[5] Mangasarian, O.L. (1969): *Nonlinear Programming*, New York: McGraw-Hill.