Information nudges and self control

Thomas Mariotti, Nikolaus Schweizer, Nora Szech, Jonas von Wangenheim

WORKING PAPERS

N° 914

June 2021

"Information Nudges and Self-Control"

Thomas Mariotti, Nikolaus Schweizer, Nora Szech
and Jonas von Wangenheim

Toulouse
School of
Economics

Information Nudges and Self-Control*

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

June 25, 2021

Abstract

We study the optimal design of information nudges for present-biased consumers who make sequential consumption decisions without exact prior knowledge of their long-term consequences. For any distribution of risks, there exists a consumer-optimal information nudge that is of cutoff type, recommending abstinence if the risk is high enough. Depending on the distribution of risks, more or less consumers have to be sacrificed, as they cannot be credibly warned even though they would like to be. Under a stronger bias for the present, the target group receiving a credible warning to abstain must be tightened, but this need not increase the probability of harmful consumption. If some consumers are more strongly present-biased than others, traffic-light nudges turn out to be optimal and, when subgroups of consumers differ sufficiently, the optimal traffic-light nudge is also subgroup-optimal. We finally compare the consumer-optimal nudge with those a health authority or a lobbyist would favor.

Keywords: Nudges, Information Design, Present-Biased Preferences, Self-Control.

*We thank Sandro Ambuehl, Bruno Biais, Kai Barron, Catherine Casamatta, Francesc Dilme, Laura Doval, Jannis Engel, Daniel Garrett, Bertrand Gobillard, Paul Heidhues, Alessandro Ispano, Yves Le Yaouanq, George Loewenstein, Stefano Lovo, Collin Raymond, Frank Rosar, Sebastian Schweighofer-Kodritsch, Roland Strausz, Jean Tirole, and Takuro Yamashita for very valuable feedback. We also thank seminar audiences at HEC Paris, Toulouse School of Economics, Universität Bonn, and Universität Freiburg, as well as several 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, the 2018 ZEW Workshop on Market Design, the 2019 Bavarian Micro Day at Universität Ulm, the 2019 Berlin IO Day, the 2019 Nordic Conference on Behavioral and Experimental Economics at Kiel Institut für Weltwirtschaft, and the 2019 Paris Workshop on Signaling in Markets, Auctions and Games for many useful discussions. Anke Greif-Winzrieth, Michelle Hörrmann, Nicola Hüholt, and Christine Knopf provided excellent research assistance. This research has benefited from the financial support of the Agence Nationale de la Recherche (Programme d'Investissement d'Avenir ANR-17-EURE-0010), the German Research Foundation (CRC TRR 190), and the research foundation TSE-Partnership.

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

‡Department of Econometrics and Operations Research, 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.

¶Institute for Microeconomics, University of Bonn, Bonn, Germany. Email: jwangenheim@uni-bonn.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 suggest that, while consumption first declined during Prohibition, it increased again after a few years, once the illegal market had adapted; consumption remained stable after Prohibition ended (Miron and Zwiebel (1991)). On top of being illiberal 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)).

However, empirical research also documents that consumers do not always feel properly addressed. In a study with adolescents, McCool, Webb, Cameron, and Hoek (2012, page 1271) report that many participants questioned whether the graphic labels "portrayed an *authentic* representation of the harm caused by smoking. Indeed, the majority perceived such 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 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, and what target groups to address best. 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 may still be the only feasible outcome in intrapersonal equilibrium, with possibly dreadful consequences.

Incomplete Information The decision maker initially has incomplete information about the harmful consequences of consumption. This can be because their likelihood hinges on his individual risk type, which he need not know with precision. Such may for instance be the case 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 assume that decision makers do not know the actual risk they are 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, alcohol, or food warnings. To avoid the negative effects of overstating 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. Building on results

[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 the Bayesian-persuasion literature, in particular Kolotilin (2015), we characterize the optimal information structure from the decision maker's perspective prior to taking any consumption decision. Our own analysis starts with an extensive comparative-statics study of the resulting consumer-optimal information nudge. We then tackle the more challenging but more realistic case where different consumers can exhibit different biases for the present.

There always exists a consumer-optimal information nudge that is of cutoff type. In the corresponding persuasion mechanism, the decision maker learns that the risk he is facing is either high or low, depending on whether it lies above or below a certain cutoff. The intuition is that cutoff mechanisms have good efficiency properties, because consumption only takes place 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 warnings against consumption for high-risk individuals within the target group of the information nudge. 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 nudge is easy, in that it requires pinning down one single parameter. What makes this task challenging is that it requires a precise knowledge of the decision maker's bias for the present.

The optimal cutoff structure outperforms perfect transparency because, by pooling risk types above the cutoff, it enables more consumers to actually find the strength to abstain from consuming once they have learned that the risk they face is relatively high, though they would have engaged in harmful consumption under full information. 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 lower but still significant risk types who end up trapped 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.

We provide two sets of comparative-statics results for the consumer-optimal information nudge. We first prove that a shift of the risk distribution towards higher levels of risk leads to a strictly lower probability of consumption. This reflects two complementary effects. First, such a shift makes abstinence more desirable; second, in line with the above properties of cutoff mechanisms, it makes it easier to sustain abstinence in an incentive-compatible way. We next investigate the impact of a shift in the decision maker's level of self-control, as measured by his bias for the present. Our main focus is on when harmful consumption takes place under the optimal information nudge, and on how the probability of this event is affected by changes in the decision maker's bias for the present.[2]

Surprisingly, the answers to these questions turn out to depend on fine properties of the distribution of risks. We first show that, if the distribution of risks satisfies the monotone-hazard-rate property, harmful consumption will take place under the optimal information nudge if and only if the decision maker's bias for the present is severe enough. However, this condition does not ensure that the probability of *harmful* consumption is monotonic in the decision maker's level of self-control—though a more severe bias for the present always comes with a higher probability that consumption, be it harmless or harmful, takes place. Indeed, it is easy to construct examples in which a lower bias for the present leads to a higher probability of harmful consumption. The basic tradeoff is that abstinence is easier to incentivize for a decision maker with higher self-control—which tends to reduce the probability of harmful consumption—but that such a decision maker is also more sensitive to the hazards of consumption. Which effect dominates on average depends on the relative probability densities of two margins of risk, that above which abstinence can be sustained, and that above which consumption becomes harmful.

While the previous results rely on a fixed and known bias for the present, levels of self-control in fact typically 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, and Kamm (2014)). This leads us to analyze the more realistic scenario in which decision makers may have high or low self-control, a characteristic that is their private information. We focus on the case in which a single information nudge has to be designed for an entire population of

[2]Notice that harmful consumption is a behavioral property of our model that can be identified in the data by asking subjects whether they would rather consume at all dates or abstain at all dates. The unhappy smoker is a case in point.

consumers. This case is empirically relevant for tobacco, alcohol, or food warnings, for which information nudges are often printed on the items consumers can choose from. Both types of decision makers can then be optimally informed via the same information structure, which turns out to take the form of a green-yellow-red traffic-light nudge. While the strongest, red warning label is drastic enough to make decision makers abstain regardless of their level of self-control, the intermediate yellow warning label convinces at least decision makers with high self-control to abstain.

Our results are threefold. First, when the two types of decision makers are sufficiently different in terms of self-control, they exert no externality on each other, so that their individually optimally information nudges can be combined into a single traffic-light nudge without affecting incentives. This traffic-light nudge is monotone in that it has a two-cutoff structure: a green label is issued for low levels of risk, a yellow one for intermediate levels of risk, and a red one for high levels of risk, with the same cutoffs as under the individually optimal nudges. Next, when the two types of decision makers become more alike in terms of self-control, a monotone traffic-light nudge remains optimal, but the corresponding cutoffs have to be modified to preserve incentive-compatibility, which in particular requires that consumers with high self-control abstain when a yellow warning is issued; hence the two types exerts an externality on each other, and it may become optimal, depending on their relative shares in the population, to sacrifice one type to the benefit of the other. This discrimination property may be a reason why traffic-light nudges, which are intuitively perceived as monotone, are one if not the most frequently used nonnumerical information structures, in addition to their potential saliency.[3] Finally, when the two types of decision makers are very similar in terms of self-control, the monotonicity of the optimal traffic-light nudge may be lost and a three-cutoff mechanism may become optimal, whereby a yellow warning is issued for both intermediate and extreme levels of risk. In that case, the intuitive content of traffic-light nudges is more questionable.

While our analysis in most of the paper takes a liberal perspective, focusing on the consumer-optimal information nudge that maximizes the consumer's expected utility at date 0, it is also interesting to derive the information nudges that are optimal from other perspectives. Examples include a health authority aiming at minimizing the probability of consumption, a lobbyist aiming at maximizing the probability of consumption, or a social planner aiming at maximizing a weighted sum of the decision maker's utilities at different dates. The cutoff structure of optimal incentive-compatible persuasion mechanisms carries

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

over to these alternative scenarios; yet, of course, the cutoffs are chosen differently. For instance, while a health authority prefers to make as many consumers as possible shy away from harmful consumption, a lobbyist prefers to lower willpower in as many consumers as possible by convincing them that the risk is not that high, so that many consumers who would wish for an information nudge that helped them abstaining are instead trapped in harmful consumption. A policy maker unaware of consumers' self-control problems may even misinterpret the information structure implemented by a lobbyist as health-concerned, when, in fact, the lobbyist deliberately chose the target group of the warning label to minimize the deterrence effect of the nudge. Finally, from a liberal perspective, it would be ideal to choose the cutoff in the consumer-optimal information nudge so that consumption is recommended if and only if it involves no harm. Yet, as we have seen, 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, and the consumer-optimal information nudge coincides with the one a health authority would favor.

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 abstain if they were instead exposed to the coarser consumer-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), Kamenica and Gentzkow (2011), and especially, in the continuous-state case, Gentzkow and Kamenica (2016). Because our mechanism designer has linear preferences about which risk types should better abstain, our baseline model is most closely related to the setting of linear, type-dependent sender preferences in Kolotilin (2015), which we follow to derive the consumer-optimal information nudge. Our analysis starts with the comparative statics of this nudge, notably with respect to the decision maker's bias for the present. But our main methodological contribution to the Bayesian-persuasion literature lies in the analysis of optimal traffic-light nudges under present-bias heterogeneity. Similar to Guo and Shmaya (2019), we find that the optimal information nudge can be implemented by an interval structure of warnings. The reason for the optimality of interval structure is, however, different. In Guo and Shmaya (2019), it stems from monotonicity of likelihood-utility ratio over receiver types, an assumption that is violated in our model. Intuitively, this is because, in our model, disagreement on the right action between consumers with low and high self-control is the strongest for intermediate risk types. As a result, when consumers with low and high self-control become more alike in our model, a nonmonotone traffic-light nudge can be optimal.

What sets our paper apart from most of the Bayesian-persuasion 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. 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 (2020) 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 based on a noisy signal that depends on both the agent's type and effort; this provides 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 corresponding signal can vary drastically according to the political goal. Policy makers unaware of or underestimating consumers' self-control problems risk implementing an information nudge that completely misses its goal. It may even maximize consumption when minimizing consumption is intended.

The paper is organized as follows. Section 2 describes the model. Section 3 characterizes optimal information disclosure. Section 4 gathers our comparative statics results. Section 5 studies the case of a mixed population in which some consumers suffer from more severe self-control problems than others. Section 6 considers alternative objective functions. Section 7 concludes. Proofs not given in the text can be found in Appendices A–C.

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 possibly harmful in the long term. The novelty of the model is that the decision maker's information about the inherent riskiness of consumption is optimized by a mechanism designer (she).

2.1 Actions and Payoffs

The decision maker lives at dates $\tau = 0, 1, 2, 3$. 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

[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.

discount function with parameters β and δ. That is, his vNM utility functions at dates 0 and 1 are given by

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

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

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 the riskiness θ are represented by a distribution $\mathbf{P}$ with cumulative distribution function F over $[0, 1]$. We assume that the support of $\mathbf{P}$ is an interval $\Theta \equiv [\underline{\theta}, \overline{\theta}]$ and that $\mathbf{P}$ admits a continuous density f that is strictly positive over the interior of Θ. Thus F is strictly increasing over Θ and, for each $\gamma \in [0, 1]$, the γ-quantile $F^{-1}(\gamma) \in \Theta$ is well-defined.[5]

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.

We throughout assume that the mechanism designer and the decision maker have common prior beliefs about θ. This is an important assumption, for, otherwise, the optimal mechanism may take a different form. Moreover, in most of our analysis, we take a liberal perspective;

[5]The assumption that $\mathbf{P}$ admits a density is only made to simplify the exposition and can easily be relaxed. When F is discontinuous, the optimal persuasion mechanism may involve mixing at an atom of $\mathbf{P}$.

that is, we assume that the mechanism designer is benevolent, so that her interests are aligned with those of the decision maker at date 0. This is especially relevant in the case of alcohol or food consumption, in which it is plausible that even a health authority internalizes some of the enjoyable aspects of consumption. Alternative objective functions for the mechanism designer—such as, for instance, that of a health authority or a lobbyist aiming at minimizing or maximizing tobacco use, respectively, or that of a social planner taking into account the decision maker's utilities at both dates 0 and 1—are considered in Section 6.

2.3 Applications

Our model applies to situations in which a mechanism designer can determine how much information she wants to reveal regarding the riskiness θ. She can pool information by issuing a coarse signal. Yet she needs to stick to the truth: that is, she cannot fool Bayesian decision makers by systematically lying to them. Depending on the application, the riskiness may be a characteristic of the product the decision maker can consume, 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.[6] 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 riskiness 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-label item or as an unhealthy, red-label 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: for instance, a doctor or a medical agency may have superior information about a patient's riskiness, and optimize the way it is communicated to him in order to influence his behavior.[7] In the latter case, the riskiness is an individual characteristic of the patient. The doctor can choose to disclose it to her patient perfectly, but she can also only tell him that he belongs to a group of smaller or larger riskiness.

A key observation in that respect is that, even if the decision maker has some private

[6]For more discussion of both theory and applications of such collective testing problems, see Aprahamian, Bish, and Bish (2019).

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

information, he may still lack the ability to translate it into his individual riskiness. This essentially amounts 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 the decision maker has some knowledge of survival probabilities, but only at the aggregate population level. Then, although he may have 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.[8] The mechanism designer has access to the relevant computation model and can offer him personalized information. Again, she may decide to issue a coarse signal pooling decision makers of different riskiness.

From now on, and bearing in mind the above two interpretations of the model, we shall uniformly refer to θ as the decision maker's *type*.

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 in θ of the date-0 and date-1 selves' utilities, 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 $\hat{\boldsymbol{\theta}} \neq t^a$ about θ, the intrapersonal game has a unique 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 $\hat{\boldsymbol{\theta}} > t^a$.*

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 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

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

[8] 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.

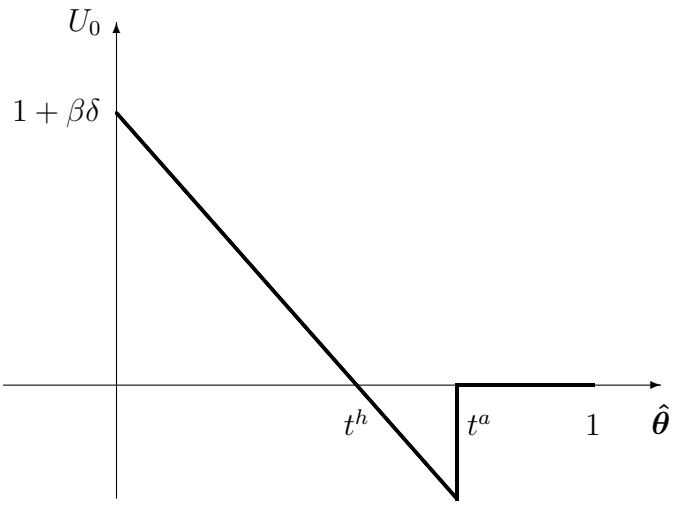

Figure 1: The decision maker's date-0 equilibrium payoff correspondence.

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 say that *harmful consumption* then 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, it is convex in $\hat{\boldsymbol{\theta}}$, reflecting that the value of information for a time-consistent decision maker is always nonnegative (Blackwell (1953)).

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 abstain, and reciprocally. Because the date-0 self can do nothing to influence the behavior of the date-1 self, and reciprocally, there exist a continuum of subgame-perfect equilibria in which both selves abstain with arbitrary probabilities in $[0, 1]$. We will follow the convention in the information-design literature by assuming that, in case of indifference, the decision maker chooses the mechanism designer's preferred pair of actions; for instance, he abstains if the mechanism designer is benevolent. Figure 1 illustrates the decision maker's date-0 equilibrium payoff correspondence.

2.5 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 his perspective at date 0. 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

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

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 $\theta > t^a$ 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.

3 Optimal Information Disclosure

The above argument shows that, because full transparency can destroy beneficial beliefs that help overcome temptation, 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 literature initiated by Brocas and Carrillo (2007), Rayo and Segal (2010), and Kamenica and Gentzkow (2011). We assume throughout this section that the mechanism designer is benevolent.

3.1 Persuasion Mechanisms

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 every type $\theta \in \Theta$ and for every element ω of some sample space Ω, a recommendation $x(\theta, \omega)$ to abstain (0) or to consume (1) at dates 0 and 1.[9] As in Aumann (1964), we can take Ω to be $[0, 1]$, endowed with

[9] If the decision maker's mean posterior belief is different from t^a, then no other recommendation would be followed by him, because his equilibrium behavior is uniquely determined by Lemma 1. If the decision maker's mean posterior belief is equal to t^a, then the intrapersonal game has multiple subgame-perfect equilibria, but the benevolent mechanism designer prefers that he abstain at each date.

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

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

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 mechanisms.

Remark Issuing messages at date 0 only involves no loss of generality. Indeed, if the mechanism designer could commit to different messages at dates 0 and 1, then certainly the date-1 self would take both messages into account when forming his consumption decision. Because less information at date 0 can only hurt the date-0 self—and, hence, the mechanism designer—it is always optimal for the mechanism designer to provide the date-0 self with any information that she provides the date-1 self with.

3.2 Incentive-Compatibility and Optimality

In this section, we formulate the relevant incentive constraints and the mechanism designer's optimization problem, and characterize the optimal information nudge. To make the problem meaningful, we assume that the support Θ of $\mathbf{P}$ is sufficiently spread out.

Assumption 1 *$F(t^h) > 0$ and $F(t^a) < 1$.*

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

A 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, there are information states in which the decision maker would be strictly better off abstaining but cannot help consuming. This reflects that abstention is not a simple default option in our setup: the decision maker abstains only if he perceives the potential consequences from consumption to be drastic enough.

Consider a mechanism π under which both 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 $\mathbf{E}[\theta \,|\, x(\theta,\omega)=1] \leq t^a$, that is,[10]

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

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,

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

A mechanism π is *incentive-compatible* (IC) 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 $\frac{1+\delta}{t^a}$, be formulated as

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

The objective function in (8) as well as the constraints (6)–(7) are all affine in π. Due to this simple structure, deriving the optimal IC persuasion mechanism lies within scope of earlier results in the Bayesian-persuasion literature, in particular those of Kolotilin (2015). This quickly leads us to Proposition 1, which offers such a characterization in our setup. Appendix A provides a self-contained derivation of this result and its extension to more general objective functions presented in Section 6.

It turns out that we can restrict attention to the class of *cutoff mechanisms* π_t, $t \in \Theta$. The cutoff mechanism π_t recommends to consume if θ is below the cutoff value t,

$$\pi_t(\theta) \equiv 1_{\{\theta<t\}},$$

and to abstain otherwise. For each $\gamma \in [0,1]$, we denote by $t_\gamma \equiv F^{-1}(\gamma)$ the γ-quantile of θ; observe that the associated cutoff mechanism π_{t_γ} recommends consumption with probability γ, $\mathbf{E}[\pi_{t_\gamma}(\theta)] = \gamma$. We now identify two crucial properties of cutoff mechanisms.

Lemma 2 *The following holds*:

(i) Among all mechanisms π such that $\mathbf{E}[\pi(\theta)] = \gamma$, the cutoff mechanism π_{t_γ} minimizes $\mathbf{E}[\theta\pi(\theta)]$.

(ii) If a mechanism π with $\mathbf{E}[\pi(\theta)] = \gamma$ is IC, then π_{t_γ} is IC as well.

[10] More generally, the left-hand side of constraint (6) is not well-defined if $\pi = 0$ **P**-almost surely, and similarly the left-hand side of constraint (7) is not well-defined if $\pi = 1$ **P**-almost surely. We adopt the convention that the undefined constraint is then emptily satisfied.

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. Moreover, they have good incentive properties because they recommend abstaining when the news about θ is the most alarming.

The two parts of Lemma 2 together imply that designing the optimal IC persuasion mechanism boils down to finding the optimal cutoff t. There is a case distinction. The objective in (8) is maximized by the unconstrained-optimal mechanism π_{t^h} that recommends consuming whenever the net benefit $t^h - \theta$ from consuming is positive. Whenever π_{t^h} is IC, it solves (8). Otherwise, the constraint (7) becomes binding.

Proposition 1 *If*

$$\mathbf{E}[\theta \,|\, \theta \geq t^h] \geq t^a, \tag{9}$$

then the optimal IC persuasion mechanism is π_{t^h}. *Otherwise, the optimal IC persuasion mechanism is* π_{t^c}, *where* $t^c \in (t^h, t^a)$ *is uniquely defined by*

$$\mathbf{E}[\theta \,|\, \theta \geq t^c] = t^a. \tag{10}$$

Combining the two cases, the optimal IC persuasion mechanism is

$$\pi_{t^*}(\theta) \equiv 1_{\{\theta < t^*\}}$$

where $t^* \equiv \max\{t^h, t^c\}$ and t^c is the cutoff value that satisfies the binding IC constraint (10).[11] For cutoff mechanisms, (10) is equivalent to (7) being binding. Thus harmful consumption takes place under the optimal IC persuasion mechanism if and only if (9) does not hold. The key insight of (10) is that, following the recommendation to abstain, the decision maker is on the verge of falling into the harmful-consumption interval (t^h, t^a). This is because 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. An optimal balance between the credibility and the efficiency of the mechanism is thereby striked: any higher value of the critical cutoff would undermine the credibility of the mechanism, whereas any lower value would render the recommendation to abstain inefficiently alarming. Notice that the optimal information nudge only warns the high-risk types. Potentially, a sizable mass of somewhat lower-risk types would prefer a warning as well. Yet the optimal nudge has to sacrifice the latter in order to convince at least the high-risk types to abstain. An example of such selective

[11] By convention, t^c is set equal to $\underline{\theta}$ if there exists no solution to (10), that is, if $\mathbf{E}[\theta] > t^a$.

nudging are alcohol warnings that target pregnant women instead of the whole population of consumers who should better drink less.

There are multiple ways of implementing the optimal IC persuasion mechanism: for example, consumption for types $\theta < t^c$ can indifferently be triggered by fully disclosing these types, or by sending the message that $\theta < t^c$. 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.

3.3 The Benefits of Optimal Information Design: An Example

To develop an intuition for the potential benefits of optimal information design in our setting, suppose that θ is uniformly distributed over $[\underline{\theta}, 1]$ for some $\underline{\theta}$, which, in line with Assumption 1, we allow to vary in $[0, t^h]$. For simplicity, we fix the values of all the other parameters, $\beta = \frac{1}{2}$, $\delta = 1$, $C = 3$, so that the welfare-optimal cutoff for consumption is $t^h = \frac{1}{2}$, while the behavioral cutoff is $t^a = \frac{2}{3}$. Given this parametrization, the optimal information nudge recommends to consume when $\theta < t^h$, and, hence, involves no harmful consumption. Indeed, the corresponding recommendation to abstain reveals that $\theta \geq t^h = \frac{1}{2}$, leading to a mean posterior belief $\hat{\boldsymbol{\theta}} = \frac{3}{4} > \frac{2}{3} = t^a$, which makes this recommendation IC.

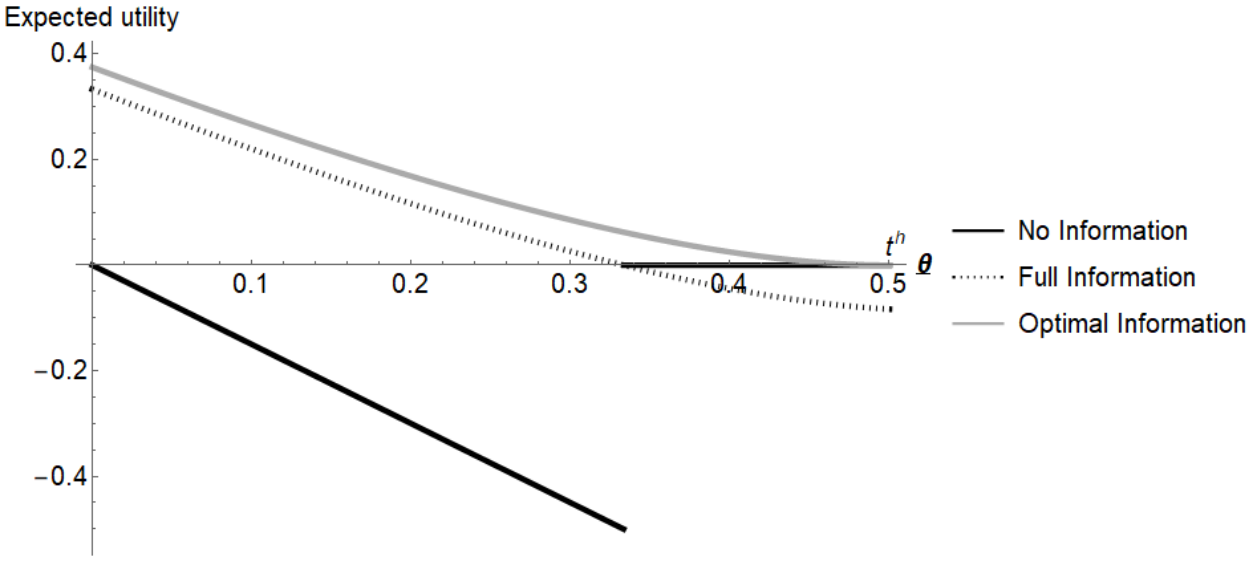

Figure 2: Expected utilities in different information scenarios for varying support $[\underline{\theta}, 1]$ of θ.

Figure 2 compares the date-0 expected utilities from no information, full information, and from the optimal information nudge for different values of $\underline{\theta} \in [0, t^h]$.[12] With no information,

[12] Direct calculations show that these expected utilities are $-\frac{3}{2}\underline{\theta}$ for $\underline{\theta} \in [0, \frac{1}{3})$ and 0 otherwise, $-\frac{3}{2}\underline{\theta} + \frac{1}{3(1-\underline{\theta})}$, and $-\frac{3}{2}\underline{\theta} + \frac{3}{8(1-\underline{\theta})}$, respectively.

the decision maker consumes if $\mathbf{E}[\theta] < t^a$, which holds for $\underline{\theta} < \frac{1}{3}$. Consumption is most harmful for $\underline{\theta}$ just below that threshold, while, for $\underline{\theta} \geq \frac{1}{3}$, the decision maker receives his abstention utility of zero. With full information, the picture is flipped. The gains from consumption outweigh the losses for $\underline{\theta} < 1/3$, while, for $\underline{\theta} > \frac{1}{3}$, harmful consumption dominates and expected utility is negative. The optimal information nudge completely avoids harmful consumption in this example. Moreover, we see that the gains from careful information design are greatest in the critical cases around the harmful-consumption trap. For $\underline{\theta}$ close to zero, the expected utility from the full-information mechanism is not much smaller than from the optimal one because consumption is mostly welfare-enhancing. For $\underline{\theta}$ approaching t^h, abstention is guaranteed when receiving no information is an option. The greatest gains from optimal design arise in between these two extremes, for instance, around $\underline{\theta} = \frac{1}{3}$, where the welfare ordering of full information and no information is reversed.

3.4 Sampling versus Information Design

It is instructive to compare our results to those obtained by Carrillo and Mariotti (2000), who suppose that the decision maker can sample costless information about θ by throwing i.i.d. biased coins with success probability θ before making his consumption decisions. We assume that $\mathbf{P}$ has full support over $[0, 1]$ and that (9) does not hold, so that the optimal IC persuasion mechanism is π_{t^c} and involves harmful consumption.

In the sampling model, 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 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 decision to consume or lead him to rationally abstain. By contrast, in the information-design model, 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 $t^c < t^a$; 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 IC 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 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 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, 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 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 would be better off avoiding the optimal information nudge. Its only shortcoming may be that the target group is smaller than some types may wish for. However, a tightening of the target group is necessary to preserve credibility and efficiently mitigate self-control problems.

3.5 Pricing Decisions

We conclude this section with a remark on responsive monopoly pricing. So far, we have abstained from describing pricing decisions. Yet our model can easily be interpreted as a model of perfect competition in which firms' marginal costs—and, hence, competitive equilibrium prices— are normalized to zero. In the following paragraph, we consider a monopolist who produces the good at zero cost. We argue that the described mechanism is equally implementable when pricing decisions are made endogenously as an optimal response to the information nudge.[13]

Consider first the case where the incentive constraint (7) binds, so that the cutoff mechanism is π_{t^c}, with t^c defined by (10). Then, for any type $\theta \geq t^c$ with a recommendation to abstain, the willingness-to-pay is

$$\mathbf{E}[1 - \beta\delta^2\theta C \,|\, \theta \geq t^c] = 1 - \beta\delta^2 t^a C = 0,$$

so that the monopolist cannot sell profitably to these types. Hence the monopolist optimally serves only types $\theta < t^c$ by charging their expected utility

$$p^m \equiv 1 - \beta\delta^2 \mathbf{E}[\theta \,|\, \theta < t^c] C > 0.$$

Similarly, if the incentive constraint (7) does not bind, then the willingness-to-pay of types

[13]The derivation of the consumer-*optimal* information signal with responsive monopoly pricing is beyond the focus of this paper. In line with Roesler and Szentes (2017), the optimal signal would involve the construction of unit-elastic demand above some cutoff. Such a signal makes the monopolist indifferent between prices on an entire interval.

with a recommendation to abstain is even below zero, and the monopolist sells to all types $\theta < t^h$ by charging their expected utitity

$$p^m \equiv 1 - \beta\delta^2 \mathbf{E}[\theta \,|\, \theta < t^h]C > 0.$$

In particular, the unconstrained optimum is implementable in a monopoly whenever it is implementable under perfect competition.

4 Comparative Statics

For comparative statics, the more interesting scenario arises when (9) does not hold, so that the optimal IC mechanism $\pi_{t^*} = \pi_{t^c}$ involves harmful consumption and $t^* = t^c > t^h$ is the unique solution to (10). As a result, one also has $t^c < t^a$: thus 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 $\pi^*_{t^c}$. That $t^c > t^h$ holds reflects the fact that consumption must take place for types for which consumption is slightly harmful to preserve the credibility of the mechanism. In this way, the optimal mechanism can recommend abstention for types for whom consumption is even more harmful, but would nevertheless take place if these types were disclosed.

4.1 Changes in the Distribution of Risks

In case (9) does not hold, the characterization (10) of the cutoff t^c leads to straightforward comparative statics in terms of the distribution $\mathbf{P}$; for simplicity, we shall assume that all distributions of risks have full support over $[0, 1]$. Suppose for instance that $\overline{\mathbf{P}}$ dominates $\underline{\mathbf{P}}$ in the hazard-rate order, that is,

$$\frac{1 - \overline{F}}{1 - \underline{F}} \text{ is nondecreasing over } [0, 1).$$

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 $\underline{\mathbf{P}}[\cdot \mid \theta > t]$ (Shaked and Shanthikumar (2007, Section 1.B.1)). This, in turn, implies that $\overline{\mathbf{E}}[\theta \,|\, \theta > t] \geq \underline{\mathbf{E}}[\theta \,|\, \theta > t]$ for any such t. It then follows from (10) that the cutoff t^c is smaller under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$, $\overline{t}^c \leq \underline{t}^c$.

As a result, if the optimal IC persuasion mechanism when θ is distributed according to $\underline{\mathbf{P}}$ involves no consumption for some type, then neither does the optimal IC 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. Such types are more likely under $\overline{\mathbf{P}}$ than under $\underline{\mathbf{P}}$ by first-order stochastic dominance. Finally, note that cases where (9) holds can be discussed in a similar way as the value of t^a does not depend on the distribution of risks and $\overline{t}^c \leq \underline{t}^c$ implies $\overline{t}^* = \max\{t^h, \overline{t}^c\} \leq \max\{t^h, \underline{t}^c\} = \underline{t}^*$. In fact, in line with Kolotilin (2015), we can derive the following stronger result.

Corollary 1 *If the distribution* $\overline{\mathbf{P}}$ *dominates the distribution* $\underline{\mathbf{P}}$ *in the increasing convex order, that is,*

$$\mathbf{E}^{\overline{\mathbf{P}}}[h(\theta)] \geq \mathbf{E}^{\underline{\mathbf{P}}}[h(\theta)] \text{ for all nondecreasing and convex } h, \tag{11}$$

then the probability of consuming is 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, because the optimal abstinence cutoff under $\underline{\mathbf{P}}$ is a fortiori IC under $\overline{\mathbf{P}}$.

4.2 Changes in the 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, which is inversely related to β. We start with the basic observation that the cutoff $t^* = \max\{t^h, t^c\}$ for θ above which abstinence is recommended is strictly decreasing in β. Indeed, if (9) holds, then $t^* = t^h \geq t^c$, and this directly follows from (3)–(4); if (9) does not hold, then $t^* = t^c > t^h$, and this directly follows from (3) and (10). Thus, if the optimal IC persuasion mechanism for a time-inconsistency parameter $\underline{\beta}$ involves abstinence for a given value of θ, then so does the optimal IC persuasion mechanism for any time-inconsistency parameter $\overline{\beta} > \underline{\beta}$. That is, a more severe bias for the present induces a higher probability that consumption takes place, which corresponds to a tightening of the target group receiving a credible warning to abstain. Notice from the above reasoning that this intuitive property is satisfied regardless of whether the optimal IC persuasion mechanism involves harmful consumption.

We now turn to the more subtle question of how a change in β affects the probability of harmful consumption. We start with a closer examination of the condition (9) under which

the unconstrained-optimal mechanism π_{t^h} is IC. By (3)–(4), this condition amounts to

$$\mathbf{E}\left[\theta \,|\, \theta > \frac{1+\beta\delta}{(1+\delta)\beta\delta^2 C}\right] \geq \frac{1}{\beta\delta^2 C}. \tag{12}$$

Because a time-consistent decision maker never engages into harmful consumption, a natural guess is that the optimal IC persuasion mechanism involves no harmful consumption and, hence, coincides with π_{t^h} 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 (12) to hold is that β be close enough to 1. By assuming some additional regularity on $\mathbf{P}$, we can turn this sufficient condition into an equivalence. Specifically, we assume that $\mathbf{P}$ satisfies the strict monotone-hazard-rate property (MHRP):

$$r(t) \equiv \frac{f(t)}{1-F(t)} \text{ is strictly increasing in } t \in [0,1).$$

The following result then holds.

Corollary 2 *If the distribution* $\mathbf{P}$ *satisfies MHRP, then* π_{t^h} *is IC if and only if* $\beta \geq \beta^u$, *where* β^u *is the unique value of* $\beta \in (\frac{1}{\delta^2 C}, 1)$ *that achieves equality in (12).*

Thus, for a fixed distribution $\mathbf{P}$ satisfying MHRP, if the optimal IC persuasion mechanism for a decision maker with time-inconsistency parameter $\underline{\beta}$ involves no harmful consumption, then neither does the optimal IC 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.

Now, consider a value of $\underline{\beta} \in (\frac{1}{\delta^2 C}, \beta^u)$ of β such that, therefore, harmful consumption takes place under the optimal IC persuasion mechanism. Does a small increase in self-control, that is, a small increase in β to some value $\overline{\beta} \in (\underline{\beta}, \beta^u)$, necessarily involve less harmful consumption? There are two opposite effects at play here. On the one hand, from the reasoning at the beginning of this section, there are values of θ such that the decision maker would be trapped in harmful consumption under $\underline{\beta}$ but abstains 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 the harmful consequences of consumption if he has more self-control. The first effect tends to reduce the optimal harmful-consumption interval (t^h, t^c); the second, to increase it. Hence, any statement about how harmful consumption varies with β under the optimal IC 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 MHRP, harmful consumption is more likely to take place, the more severe the decision maker's bias for the present.

Corollary 3 *If the distribution* $\mathbf{P}$ *satisfies MHRP and its density* f *is such that*

$$\text{For all } t > t',\ f(t) > \frac{1}{1+\delta}\, f(t'), \tag{13}$$

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

Condition (13) is satisfied, for instance, if $\mathbf{P}$ is the uniform distribution. More generally, it requires that the density f does not decrease too fast over $[0,1]$, so that the margin of risk above which abstinence can be sustained, t^c, remains in a probabilistic sense more important than that above which consumption becomes harmful, t^h. However, condition (13) is not satisfied, for instance, if $\mathbf{P}$ is a Beta(a,b) distribution with $a, b > 1$, which satisfies MHRP (Bagnoli and Bergstrom (2005)), but not condition (13) as then $f(1) = 0$. The following result shows that Corollary 3 does not extend to this case. Specifically, 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.

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

$$f(1) < \frac{1}{2(1+\delta)}\, f\left(\frac{1+\frac{1}{\delta C}}{1+\delta}\right), \tag{14}$$

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

The somewhat convoluted condition (14) can be intuitively interpreted as follows. If initially $\beta \approx \frac{1}{\delta^2 C}$, then nearly all types consume under the optimal IC persuasion mechanism, that is, $t^c \approx 1$. If β increases by $\mathrm{d}\beta$, then the cutoff t^c above which abstinence can be sustained 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 becomes harmful,

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

decreases by an amount

$$\mathrm{d}t^h \approx \frac{\mathrm{d}t^c}{1+\delta}\, \frac{\mathrm{d}}{\mathrm{d}t} \left\{\mathbf{E}[\theta \,|\, \theta > t]\right\}\Big|_{t=1^-} \geq \frac{\mathrm{d}t^c}{2(1+\delta)}$$

under the weak conditions we impose on f.[14] The mass of new types thus trapped in harmful consumption is bounded below by $\frac{1}{2(1+\delta)} f(t^h)\,\mathrm{d}t^c$, which exceeds the mass $f(1)\,\mathrm{d}t^c$ of neutralized types if condition (14) is satisfied. As a result, the probability of the harmful-consumption interval (t^h, t^c) locally increases in β. Notice that, because f is assumed to be strictly positive over $(0,1)$, condition (14) is satisfied as soon as $f(1) = 0$.

5 Traffic-Light Nudges

Our analysis so far has relied on the assumption that there is a single decision maker with known bias for the present or, if there is a population of decision makers, that they have identical characteristics. Yet, in practice, 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 other factors such as the level of glucose available, so that a hungry individual may display comparatively little self-control (Gailliot and Baumeister (2007)). This makes it important to assess which of the insights from our basic model carry over to a more realistic scenario in which decision makers' self-control levels are not known to the mechanism designer.

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 $0 < \beta_L < \beta_H \leq 1$.[15] The vNM utility functions at dates 0 and 1 for type $i = L, H$ are given by

$$U_{i,0}(x_0, x_1, \theta) \equiv x_0(1 - \beta_i\delta^2\theta C) + x_1\beta_i\delta(1 - \delta^2\theta C), \tag{15}$$

$$U_{i,1}(x_0, x_1, \theta) \equiv -x_0\beta_i\delta C + x_1(1 - \beta_i\delta^2\theta C). \tag{16}$$

The riskiness θ is assumed to be known to the mechanism designer and the same for each decision maker, regardless of his level of self-control. Whether a specific decision maker is of

[14] The intuition for the factor $\frac{1}{2}$ 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 $\mathrm{d}[\frac{1}{2}(t+1)] = \frac{1}{2}\,\mathrm{d}t$.

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

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, or *joint mechanism*, simultaneously targeting types L and H. As a result, both types L and H are exposed to the same information, which is empirically the case for tobacco, alcohol, or food warnings. For simplicity, we assume that the density f of $\mathbf{P}$ is strictly positive over $(0,1)$ and, whenever needed, that $\mathbf{P}$ satisfies MHRP.

It is clear from (15)–(16) that, for any mean posterior belief $\hat{\boldsymbol{\theta}}$, type L consumes whenever type H does. Therefore, we can focus on measurable direct joint 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

$$\pi_j(\theta) \equiv \boldsymbol{\lambda}[\{\omega \in \Omega : x(\theta, \omega) = j\}]. \tag{17}$$

As in Section 3.1, we can identify x with $\pi \equiv (\pi_0, \pi_L, \pi_{LH})$. For each type $i = L, H$, 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 for the individually optimally optimal mechanisms.

5.1 The No-Externality Case

We first determine when the two types exert no externality on each other, in the sense that there is no relevant strategic interaction between their individually optimal mechanisms. The latter can then be straightforwardly combined into a joint mechanism without affecting incentives. For each type $i = L, H$, the optimal IC persuasion mechanism characterized in Proposition 1 recommends abstinence if and only 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 joint mechanism that merges the two individually optimal mechanisms,[16]

$$(\pi_0^*, \pi_L^*, \pi_{LH}^*)(\theta) \equiv (1_{\{\theta > t_L^*\}}, 1_{\{t_H^* < \theta \leq t_L^*\}}, 1_{\{\theta \leq t_H^*\}}) \tag{18}$$

is IC. By inspection, this is the case if and only if, upon receiving recommendation L, type H is willing to abstain, that is,

$$\mathbf{E}[\theta \,|\, t_H^* < \theta \leq t_L^*] \geq t_H^a. \tag{19}$$

For β_H close enough to 1, we have $t_H^* = t_H^h \approx t_H^a$, and the incentive constraint (19) 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

[16] That is, the *meet* of the two individually optimal partitions of $[0,1]$.

constraint (19) 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 2 *If the distribution* $\mathbf{P}$ *satisfies MHRP, then, for each* $\beta_L > \frac{1}{\delta^2 C}$, *there exists a threshold* $\beta_H^{ne}(\beta_L) \in (\beta_L, 1)$ *such that the joint mechanism (18) is IC if and only if* $\beta_H \geq \beta_H^{ne}(\beta_L)$. *The threshold* $\beta_H^{ne}(\beta_L)$ *is strictly greater than* β^u *and is strictly increasing in* β_L.

Owing to its two-cutoff structure, the joint mechanism (18) has a natural interpretation as a *monotone* traffic-light nudge, whereby the green-yellow-red labels are used to signal low-intermediate-high riskiness. This makes this nudge especially simple to understand, adding to its potential salience.

5.2 The Externality Case

We now analyze the case where (19) does not hold. Then the two individually optimal mechanisms are not simultaneously implementable, so that the two types $i = L, H$ exert an externality on each other: at least one of them is bound to suffer from the existence of the other. In line with (6)–(7), the joint mechanism (π_0, π_L, π_{LH}) is IC if and only if

$$\frac{\mathbf{E}[\theta\pi_0(\theta)]}{\mathbf{E}[\pi_0(\theta)]} \geq t_L^a, \tag{20}$$

$$\frac{\mathbf{E}[\theta\pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} \leq t_L^a, \tag{21}$$

$$\frac{\mathbf{E}[\theta\pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} \geq t_H^a, \tag{22}$$

$$\frac{\mathbf{E}[\theta\pi_{LH}(\theta)]}{\mathbf{E}[\pi_{LH}(\theta)]} \leq t_H^a. \tag{23}$$

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[17]

$$\max\left\{\sum_{i=L,H} p_i\beta_i\{t_i^h\mathbf{E}[\Pi_i(\theta)] - \mathbf{E}[\theta\Pi_i(\theta)]\} : \pi \text{ is IC}\right\}. \tag{24}$$

For simplicity, we will first 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.

[17]Notice that the population shares p_L and p_H in (24) can also be interpreted as Pareto weights in the mechanism designer's social-welfare function. This is because we study a pure information-design problem, with no aggregate resource constraint.

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

Assumption 2 intuitively states that, 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), this is equivalent to

$$\beta_H > \beta_H^{no}(\beta_L) \equiv \frac{(1+\delta)\beta_L}{1+\beta_L\delta} \in (0,1),$$

so that β_H is large enough relative to β_L. This lower bound is nevertheless consistent with $\beta_H < \beta_H^{ne}(\beta_L)$, in which case, according to Proposition 2, 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 $\beta_H = \beta_H^{no}(\beta_L)$, we have $t_L^h = t_H^a > t_H^*$, and constraint (19) 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 $\beta_H = \beta_H^{no}(\beta_L)$, the joint mechanism (18) is not IC and the threshold $\beta_H^{ne}(\beta_L)$ in Proposition 2 satisfies $\beta_H^{ne}(\beta_L) > \beta_H^{no}(\beta_L)$, as required.

Under Assumption 2, designing an IC joint mechanism is straightforward; for instance, the mechanism designer may sacrifice type H by letting him consume when $\theta \leq t_H^a$, while offering type L his individually optimal information nudge by setting

$$(\pi_0, \pi_L, \pi_{LH})(\theta) \equiv (1_{\{\theta > t_L^*\}}, 1_{\{t_H^a < \theta \leq t_L^*\}}, 1_{\{\theta \leq t_H^a\}}). \tag{25}$$

Notice that the joint mechanism (25) is again a monotone traffic-light nudge, and that we have used Assumption 2, which ensures that $t_H^a < t_L^h \leq t_L^*$. However, as the expectation of θ conditional on the yellow warning, that is, recommendation L being issued, is strictly higher than t_H^a, this joint mechanism can be improved by lowering the cutoff below which consumption for type H is recommended, while keeping the same cutoff t_L^* below which consumption for type L is recommended. We will see in Corollary 5 that such a modification can be optimal provided the share of type L in the population is high enough. Meanwhile, the upshot of this discussion is that there are gains from pooling intermediate values of θ into a yellow warning label. Indeed, the central result of this section more generally states that, under Assumption 2, a two-cutoff joint mechanism is optimal.

Proposition 3 *Under Assumption 2, there exist two cutoffs* $0 < t_{LH}^{**} \leq t_L^{**} \leq 1$ *such that*

$$(\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})(\theta) \equiv (1_{\{\theta > t_L^{**}\}}, 1_{\{t_{LH}^{**} < \theta \leq t_L^{**}\}}, 1_{\{\theta \leq t_{LH}^{**}\}}) \tag{26}$$

is an optimal IC joint mechanism.

When $t^{**}_{LH} < t^{**}_{L}$, the optimal IC joint mechanism can be implemented by a three-label monotone traffic-light nudge; as we will see in Lemma 3 below, this inequality is always satisfied 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 warning label; all consumers will find their riskiness 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 warning label; consumers with high self-control will find their riskiness high enough to abstain, while 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 label; all consumers will find their riskiness low enough to consume. Such a monotone traffic-light nudge can thus optimally reach consumers with low self-control without sacrificing those with high self-control; it also has an easy-to-grasp connotation.[18] Koenigstorfer, Groeppel-Klein, and Kamm (2014) confirm this prediction in an empirical study, comparing consumers with high and low levels of self-control.

Proposition 3 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 because the red label 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 because this relaxes the key incentive constraint (22). In Appendix C, we prove that, under Assumption 2, the first effect dominates, which gives rise to a monotone traffic-light nudge.

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 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

[18] A red label may be an especially salient warning. 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.

information nudges in order to support healthy eating habits and fight obesity. Proposition 3 suggests that heterogeneity in levels of self-control across consumers is a possible rationale for monotone traffic-light nudges that use green-yellow-red labels as an ordered signal of a potentially harmful good's riskiness.

Our next result explicitly characterizes the optimal cutoffs (t_{LH}^{**}, t_L^{**}) that correspond to the green-yellow and yellow-red boundaries, respectively.

Lemma 3 *Suppose that (19) 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*

$$\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L^*) < \theta \leq t_L^*] = t_H^a. \tag{27}$$

Then the optimal cutoffs (t_{LH}^{**}, t_L^{**}) in (26) satisfy $t_{LH}^{**} < t_L^{**}$ and are given by

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

$$\frac{p_H \beta_H}{p_L \beta_L} \, \frac{\hat{t}_{LH}(t_L^*) - t_H^h}{t_L^* - t_L^h} \leq \frac{t_H^a - \hat{t}_{LH}(t_L^*)}{t_L^* - t_H^a}, \tag{28}$$

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

$$\frac{p_H \beta_H}{p_L \beta_L} \, \frac{t_H^* - t_H^h}{1 - t_L^h} \geq \frac{t_H^a - t_H^*}{1 - t_H^a}, \tag{29}$$

3. *the unique solution to*

$$\mathbf{E}[\theta \,|\, t_{LH}^{**} < \theta \leq t_L^{**}] = t_H^a \;\; \textit{and} \;\; \frac{p_H \beta_H}{p_L \beta_L} \, \frac{t_{LH}^{**} - t_H^h}{t_L^{**} - t_L^h} = \frac{t_H^a - t_{LH}^{**}}{t_L^{**} - t_H^a} \tag{30}$$

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 into 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 (22) from below, which is proportional to $t_H^a - t_{LH}^{**}$. Similarly, pooling marginally more risks into red rather than into yellow by decreasing t_L^{**} comes at a benefit proportional to $p_L \beta_L (t_L^{**} - t_L^h)$ due to higher abstinence of type L. Yet there is also the marginal cost of tightening type H's incentive constraint (22) from above, which is proportional to $t_L^{**} - t_H^a$. 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^*, and is thus privileged to the detriment of type H. 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, who is thus privileged to the detriment of type L.

The cutoff characterization conditions (28)–(29) affords us straightforward comparative statics with respect to the population share of type H, which determines which of Cases 1–3 in Lemma 3 arises.

Corollary 5 *Suppose that (19) 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 IC joint mechanism implements the individually optimal cutoff t_L^* for type L and the cutoff for type H is determined by (27),*

2. *for $p_H \in [\overline{p}, 1]$, the optimal IC joint 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 IC joint mechanism implements the interior solution to (30). 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 IC in the sense of Proposition 1, and similarly $\overline{p} = 1$ if and only if the individually unconstrained-optimal mechanism for type L is IC in the sense of Proposition 1.

We conclude this section with a short discussion of what can happen when Assumption 2 does not hold. In that case, pooling the highest-risk types into yellow rather than red may prove so efficient at relaxing incentive constraint (22) that it becomes optimal to screen type L and type H when they have high riskiness. In that case, a nonmonotone traffic-light nudge may be optimal, but we should stress that such a nudge loses much of the intuitive appeal of those we have encountered so far, as the yellow warning label now pools intermediate and extreme values of θ. Alternatively, if the two types are very similar, a pooling outcome can emerge, in which both types face the individually optimal information nudge for type L.

Proposition 4 *In general, there exist three cutoffs* $0 < t_{LH}^{**} \leq \underline{t}_L^{**} \leq \bar{t}_L^{**} \leq 1$ *such that*

$$(\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})(\theta) \equiv (1_{\{\underline{t}_L^{**} < \theta \leq \bar{t}_L^{**}\}}, 1_{\{t_{LH}^{**} < \theta \leq \underline{t}_L^{**}\}} + 1_{\{\theta > \bar{t}_L^{**}\}}, 1_{\{\theta \leq t_{LH}^{**}\}}) \tag{31}$$

is an optimal IC joint mechanism.

5.3 A Remark on the Continuous Case

When there is more heterogeneity in the decision makers' degree of self-control in the population, incentivizing all types of decision makers with a single information nudge becomes increasingly difficult. Intuitively, this is because a nudge that is just strong enough to convince a certain β-type to abstain will be just too weak to induce a slightly lower β-type to abstain. To speak to this issue, we study in this section how the structure of optimal information nudges changes when the degree of self-control β in the population is continuously distributed, with a strictly positive density over the interval $(\frac{1}{\delta^2 C}, 1)$.

Because there is a one-to-one correspondence between β and the cutoff $t^a = \frac{1}{\beta\delta^2 C}$, it is convenient, for the purpose of this section, to think of a decision maker's private type as being described by $t^a \in [t_0, 1] \equiv [\frac{1}{\delta^2 C}, 1]$ rather than by β; a high value of t^a corresponds to a low degree of self-control β. We denote by H the cumulative distribution function of t^a induced by the distribution of β and by h the corresponding density function over the interval $[t_0, 1]$, which we assume to be continuously differentiable over $(t_0, 1)$. In this new interpretation of our model, the decision maker thus draws a cutoff t^a as his private type; the mechanism designer then observes θ but not t^a.

The following result then follows from adapting the arguments in Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017, Theorems 1–2).

Proposition 5 *If the density* h *of the distribution of cutoffs* t^a *is log-concave with* $h'(t_0) > 0$*, then there exists a cutoff* $\tilde{t} > t_0$ *such that an optimal IC persuasion mechanism prescribes full disclosure of* θ *for* $\theta \leq \tilde{t}$ *and issues a red warning label for* $\theta > \tilde{t}$*, which is an IC recommendation to abstain for decision makers with* $t^a \leq \mathbf{E}[\theta \mid \theta > \tilde{t}]$*. If, in addition,* $h(1) = 0$ *and* $h'(1) < 0$*, then we have an interior solution,* $\tilde{t} < 1$*.*

Because $t^a = \frac{1}{\beta\delta^2 C}$, the log-concavity of h is equivalent to assuming that the distribution of $\frac{1}{\beta}$ is log-concave. If we think of the discount factor β as being generated by a discount rate R, that is, $\beta \equiv \frac{1}{1+R}$, then the log-concavity of h is thus equivalent to the log-concavity of the density of R. Together with the assumption that $h'(t_0) > 0$ and $h'(1) < 0$, Proposition 5 thus requires that the distribution of R be well-behaved and unimodal.

When information is continuously distributed both on the designer's and on the decision maker's sides, we do not expect a simple three-color traffic light to remain optimal—and, indeed, this is not what we find. Instead, there is still a clear red warning label, while the yellow and green labels are replaced by a more precise, continuous signal. However, it seems plausible, in light of the analysis of Sections 5.1–5.2, that a traffic-light structure, possibly with more than three labels, remains optimal as long as the distribution of β, as perceived by the mechanism designer, is discrete. Whether the relevant information nudge has a traffic-light structure, as in Propositions 2–3, or a more continuous structure, as in Proposition 5, thus ultimately hinges on the precision of the probabilistic information the mechanism designer has about the decision makers' degrees of self-control.

6 More General Objective Functions

So far, we have focused on benevolent persuasion mechanisms that maximize the decision maker's date-0 utility. We now return to the setting of Section 3 but take a look at more general objective functions for the mechanism designer. 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 might want to use an information nudge that deters as many people as possible from consuming as it is not interested in its short-term enjoyable aspects.

Motivated by these considerations, we analyze in this section more flexible objective functions for the mechanism designer, with the only restriction that the designer's utility from consumption is continuous and nonincreasing in the consumer's risk type. Specifically, we assume that the designer's utility is of the form

$$V(x_0, x_1, \theta) \equiv x_0 v_0(\theta) + x_1 v_1(\theta) \tag{32}$$

for some continuous and nonincreasing period utility functions v_0 and v_1. This class of objective functions includes the above mentioned cases of a lobbyist with $v_0 \equiv v_1 \equiv 1$ and of a health authority with $v_0 \equiv v_1 \equiv -1$, assuming that such mechanism designers only care about the probability of consumption irrespective of θ. It also includes the case of a mechanism designer who does not internalize the decision maker's bias for the present, and thus maximizes the expected utility of a decision maker with $\beta = 1$, or that of a mechanism designer who maximizes a weighted sum of the expected utilities of the date-0 and date-1 selves or takes into account the expected utilities of overlapping generations of consumers currently at different life stages.

As the decision problem of the decision maker is the same as in Section 3, his consumption decision at each date $t = 0, 1$ is again pinned down by his mean posterior belief $\hat{\boldsymbol{\theta}}$. Hence, a persuasion mechanism is IC if and only if it satisfies (6)–(7) and the designer's realized utility is given by 0 if the decision maker abstains and by

$$v(\theta) \equiv V(1, 1, \theta) = v_0(\theta) + v_1(\theta)$$

if the decision maker consumes.

In Appendix A, we show that Lemma 2 still holds in this more general environment, so that we can again with no loss of generality restrict attention to cutoff mechanisms π_t, $t \in \Theta$. We further show that the set of IC cutoffs is an interval $\mathcal{I} \equiv [t^c, t^d]$, where

$$t^c \equiv \inf \{t \in \Theta : \mathbf{E}[\theta \,|\, \theta \geq t] \geq t^a\}, \tag{33}$$

$$t^d \equiv \sup \{t \in \Theta : \mathbf{E}[\theta \,|\, \theta < t] < t^a\}. \tag{34}$$

The definition of t^c in (33) extends the one given in Section 3, defining t^c as the smallest cutoff such that a recommendation to abstain convinces all types above it. Conversely, t^d is the largest cutoff such that a recommendation to consume convinces all types below it. The cutoffs t^c and t^d are thus the extremal values at which the two IC constraints (6) and (7) are still satisfied. The interval $\mathcal{I}$ is always nonempty as $t^c < t^a < t^d$. Moreover, one of the two IC constraints is always trivially satisfied, $t^c = \underline{\theta}$ or $t^d = \overline{\theta}$, depending on whether $\mathbf{E}[\theta]$ is above or below t^a.

Thus, we can find IC mechanisms that maximize (32) by solving the problem

$$\max \{\mathbf{E}[v(\theta)\pi_t(\theta)] : t \in \mathcal{I}\}.$$

Because v is nonincreasing, solving this problem is immediate. Three cases can arise.

Proposition 6 *The following holds*:

(i) If there exists $t^* \in \mathcal{I}$ such that $v(t^*) = 0$, then π_{t^*} is an optimal IC mechanism.

(ii) If $v(t) < 0$ for all $t \in \mathcal{I}$, then π_{t^c} is an optimal IC mechanism.

(iii) If $v(t) > 0$ for all $t \in \mathcal{I}$, then π_{t^d} is an optimal IC mechanism.

We can interpret Proposition 6 as follows. If the designer-optimal cutoff t^* and the consumer-optimal cutoff t^h are close enough—that is, their interests are sufficiently aligned—the mechanism designer's optimal cutoff is IC. If the mechanism designer favors a sufficiently

lower consumption rate than the consumers—as in the case of a health authority—then her preferred abstention warning may violate incentive-compatibility in (7). Hence she must tighten the target group that receives a consumption warning. Conversely, if the mechanism designer favors a sufficiently higher consumption rate than the consumers—as in the case of a lobbyist—then her preferred abstention warning may violate incentive-compatibility in (6). Hence she will sacrifice some high-risk types in order to achieve IC consumption recommendations for the largest possible population of low-risk types. There may thus be types 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.[19] 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 at especially high risk of stroke or heart disease, but for everybody.[20] Thus the release of a warning to a high-risk group can at the same time function as a justification to continue harmful consumption for lower-risk groups. Policy makers need to be aware of this problem, which arises because of present-biased preferences.

7 Concluding Remarks

In this paper, we have studied the optimal design of credible information nudges targeted at a population 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 severe bias need a much more drastic signal in order to avoid

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

[20] 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.

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, in addition to the strong, red warning, a specifically milder, yellow warning. Thus heterogeneity in self-control is a rationale for the traffic-light nudges we observe in practice.

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 the effects of self-control when it comes to the design and evaluation of powerful information nudges to limit harmful consumption.

Appendix A: Proofs for Sections 3 and 6

This appendix provides the technical details for the claims of Sections 3 and 6. To simultaneously prove Proposition 1 and 6, we study a maximization problem that is slightly more general than (8), namely,

$$\max\{\mathbf{E}[v(\theta)\pi(\theta)] : \pi \text{ is IC}\}, \tag{A.1}$$

for some continuous nonincreasing function v; in (8), $v(\theta) = t^h - \theta$. The IC constraints are given by (6)–(7) as in Section 3.

We first generalize Lemma 2 by showing that we can with no loss of generality restrict attention to cutoff mechanisms π_t, $t \in \Theta$. To prove this claim, notice first that, for each $\gamma \in [0,1]$, among all mechanisms π with recommendation probability $\mathbf{E}[\pi(\theta)] = \gamma$, the cutoff mechanism π_{t_γ} with cutoff $t_\gamma \equiv F^{-1}(\gamma)$ concentrates as much mass as possible on small values of θ; hence, as v is nonincreasing, it maximizes $\mathbf{E}[v(\theta)\pi(\theta)]$ in this class of mechanisms. Moreover, inspecting the IC conditions (6)–(7), we find that they depend on the mechanism π only through $\mathbf{E}[\pi(\theta)]$ and $\mathbf{E}[\theta\pi(\theta)]$. In particular, holding $\gamma = \mathbf{E}[\pi(\theta)]$ fixed, both constraints only become easier to satisfy when $\mathbf{E}[\theta\pi(\theta)]$ is made smaller, which is again achieved by the cutoff mechanism π_{t_γ}. Thus, if a mechanism π with $\mathbf{E}[\pi(\theta)] = \gamma$ is IC, then π_{t_γ} is IC as well. The claim follows.

We next characterize under which conditions a cutoff mechanism is IC. To this end, notice that for π_t the IC constraints (6)–(7) can be written as

$$m(t) \equiv \mathbf{E}[\theta \,|\, \theta < t] \leq t^a, \tag{A.2}$$

$$M(t) \equiv \mathbf{E}[\theta \,|\, \theta \geq t] \geq t^a. \tag{A.3}$$

Under our assumptions on $\mathbf{P}$, both m and M are continuous, strictly increasing functions of $t \in \Theta$, which satisfy $m(\underline{\theta}) = \underline{\theta}$, $m(\overline{\theta}) = \mathbf{E}[\theta]$, $M(\underline{\theta}) = \mathbf{E}[\theta]$, and $M(\overline{\theta}) = \overline{\theta}$. We distinguish two cases. If $t^a > \mathbf{E}[\theta]$, then (A.2) is automatically satisfied, and (A.3) is satisfied if and only if $t \geq t^c$, where $t^c \equiv M^{-1}(t^a) \in (\underline{\theta}, \overline{\theta})$. Alternatively, if $t^a \leq \mathbf{E}[\theta]$, then (A.3) is automatically satisfied, and (A.2) is satisfied if and only if $t \leq t^d$, where $t^d \equiv m^{-1}(t^a) \in (\underline{\theta}, \overline{\theta}]$. In either case, we denote by $\mathcal{I} \equiv [t^c, t^d]$ the set of IC cutoffs; thus $t^d = \overline{\theta}$ if $t^a > \mathbf{E}[\theta]$ and $t^c = \underline{\theta}$ if $t^a \leq \mathbf{E}[\theta]$. Using the definitions (A.2)–(A.3) of the functions m and M and the definitions of t^c and t^d, it is straightforward to check that $t^a \in (t^c, t^d)$, as expected.

In light of the above discussion, (A.1) boils down to maximizing

$$\mathbf{E}[v(\theta)\pi_t(\theta)] = \int_{\underline{\theta}}^{t} v(\theta) f(\theta)\, \mathrm{d}\theta$$

with respect to $t \in \mathcal{I}$. Solving this problem is immediate given that v is nonincreasing, which implies that the objective function is quasiconcave. If $v > 0$ over $\mathcal{I}$, then $t = t^d$ is the unique solution. If $v < 0$ over $\mathcal{I}$, then $t = t^c$ is the unique solution. Finally, if v vanishes over $\mathcal{I}$, then any $t \in \mathcal{I} \cap v^{-1}(0)$ is a solution. This concludes the proof of Proposition 6.

It remains to complete the proof of Proposition 1. In that case, $v(\theta) = t^h - \theta$ is strictly decreasing and linear in θ, with $v(t^h) = 0$ for $t^h < t^a$. Because the set $\mathcal{I} = [t^c, t^d]$ of IC cutoffs is such that $t^c < t^a < t^d$, we can conclude that $t^h < t^d$, so that the cutoff t^h can never be too high for being IC. Accordingly, when $t^h \geq t^c$, which is exactly (9), we have $t^h \in \mathcal{I}$ and the optimal IC mechanism is π_{t^h}. By contrast, when $t^h < t^c$, which is exactly the negation of (9), we have $t^c > \underline{\theta}$ and the optimal IC mechanism is π_{t^c} because in that case $v < 0$ over $\mathcal{I}$; finally, (10) is exactly $M(t^c) = t^a$. This concludes the proof of Proposition 1. An important observation is that the constraint (6) is slack at the optimum.

Appendix B: Proofs for Section 4

Proof of Corollary 1. We focus with no loss of generality on the case where (9) does not hold under both $\overline{\mathbf{P}}$ and $\underline{\mathbf{P}}$. We need to show that that $\underline{F}(\underline{t}^c) \geq \overline{F}(\bar{t}^c)$. For this we show that the unique $\hat{t}$ defined by $\underline{F}(\underline{t}^c) = \overline{F}(\hat{t})$ satisfies $\hat{t} \geq \bar{t}^c$. We have

$$\begin{aligned}
\int_{\hat{t}}^{1} (\theta - t^a)\, \overline{\mathbf{P}}(\mathrm{d}\theta) &= \int_{\overline{F}^{-1}(\underline{F}(\underline{t}^c))}^{1} (\theta - t^a)\, \overline{\mathbf{P}}(\mathrm{d}\theta) \\
&= \int_{\underline{F}(\underline{t}^c)}^{1} [\overline{F}^{-1}(p) - t^a]\, \mathrm{d}p \\
&\geq \int_{\underline{F}(\underline{t}^c)}^{1} [\underline{F}^{-1}(p) - t^a]\, \mathrm{d}p \\
&= \int_{\underline{t}^c}^{1} (\theta - t^a)\, \underline{\mathbf{P}}(\mathrm{d}\theta) \\
&= 0 \\
&= \int_{\bar{t}^c}^{1} (\theta - t^a)\, \overline{\mathbf{P}}(\mathrm{d}\theta),
\end{aligned}$$

where the inequality follows from Shaked and Shanthikumar (2007, Section 4.A.1) and the last two equalities follow from (10). If $\hat{t} \geq t^a$, then a fortiori $\hat{t} > \bar{t}^c$. Otherwise, $\max\{\bar{t}^c, \hat{t}\} < t^a$ implies that $\theta - t^a < 0$ for θ between $\bar{t}^c$ and $\hat{t}$, so that $\hat{t} \geq \bar{t}^c$ from the above inequality. Hence the result. ■

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 $\frac{1}{\delta^2 C} < b \leq 1$. Corollary 2 corresponds to the special case $b = 1$. A first observation is that the MHR property is preserved by left-truncation.

Lemma B.1 *If the distribution* $\mathbf{P}$ *satisfies MHRP, then, for each* $b \in (0, 1)$*, the distribution* $\mathbf{P}_b$ *satisfies MHRP as well.*

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

$$r_b(t) \equiv \frac{f_b(t)}{1 - F_b(t)} \propto \frac{f(t)}{F(b) - F(t)} = r(t)\, \frac{1 - F(t)}{F(b) - F(t)},$$

so that $r_b(t)$ is the product of two strictly positive and strictly increasing functions of t. The result follows. ■

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

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

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 $\phi_b(\frac{1}{b\delta^2 C}) < 0$, 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 $\xi \equiv \frac{1}{\beta\delta^2 C}$ and changing variables accordingly, it is equivalent to the claim that

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

is strictly decreasing in $\xi \in (\frac{1}{\delta^2 C}, b)$. A classical result from reliability theory (see, for instance, Bryson and Siddiqui (1969)) states that, for a distribution that satisfies MHRP, the mean residual life is strictly decreasing in the age. By Lemma B.1, such is the case of $\mathbf{P}_b$, and thus

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}_b[\theta \,|\, \theta > t]\right\} < 1$$

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

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

for all $\xi \in (\frac{1}{\delta^2 C}, b)$ as $\frac{1}{1+\delta} < 1$. Hence the result. ■

Proof of Corollary 3. According to (3)–(4) and (10), we can rewrite the probability of harmful consumption as

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

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

$$H(t) \equiv F(t) - F\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + \frac{1}{\delta C}}{1+\delta}\right) \tag{B.2}$$

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 $t \in (t^u, 1)$,

$$t > \frac{\mathbf{E}[\theta \,|\, \theta > t] + \frac{1}{\delta C}}{1+\delta} \tag{B.3}$$

because, as β^u is the unique value of $\beta \in (\frac{1}{\delta^2 C}, 1)$ that achieves equality in (12), (B.3) becomes an equality at $t = t^u$ and because, as $\mathbf{P}$ satisfies MHRP, the mapping $t \mapsto (1+\delta)t - \mathbf{E}[\theta \,|\, \theta > t]$ is strictly increasing over $[0, 1)$. Then, for each $t \in (t^u, 1)$,

$$\begin{aligned} H'(t) &= f(t) - \frac{1}{1+\delta} f\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + \frac{1}{\delta C}}{1+\delta}\right) \frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} \\ &\geq f(t) - \frac{1}{1+\delta} f\left(\frac{\mathbf{E}[\theta \,|\, \theta > t] + \frac{1}{\delta C}}{1+\delta}\right) \\ &> 0, \end{aligned} \tag{B.4}$$

where the first inequality again follows from MHRP, and the second inequality follows from (13) and (B.3). Hence the result. ■

Proof of Corollary 4. Defining H as in (B.2), we have

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

in a strict right-neighborhood of $\beta = \frac{1}{\delta^2 C}$ if and only if $H' < 0$ in a strict left-neighborhood of $t = 1$ or, equivalently,

$$f(1) - \frac{1}{1+\delta} f\left(\frac{1+\frac{1}{\delta C}}{1+\delta}\right) \liminf_{t\to 1^-} \frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} < 0, \tag{B.5}$$

according to (B.4). We need to show that (14) implies (B.5) if $f(1) > 0$ or, if $f(1) = 0$, if f is nonincreasing in a left-neighborhood of $t = 1$.[21] That is, we need to show that, under these assumptions,

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

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 $\varepsilon > 0$,

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

for all n. Then, because

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\{\mathbf{E}[\theta \,|\, \theta > t]\right\} = \frac{f(t)}{1-F(t)}\left\{\mathbf{E}[\theta \,|\, \theta > t] - t\right\},$$

we have

$$f(t_n)\left\{\int_{t_n}^1 \theta f(\theta)\,\mathrm{d}\theta - t_n[1-F(t_n)]\right\} - \frac{1-\varepsilon}{2}\left[1-F(t_n)\right]^2 < 0 \tag{B.6}$$

for all n. Consider then the function

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

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 (B.6), 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 $\mathbf{E}[\theta \,|\, \theta > t] - t$ is strictly positive for all $t \in [0,1)$, the same conclusion obtains if f is nonincreasing, so that its right upper Dini derivative D^+f is nonpositive in a strict left-neighborhood of $t = 1$. Hence the result. ■

[21] Notice that, in the latter case, condition (14) is automatically satisfied.

Appendix C: Proofs for Section 5

Proof of Proposition 2. 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 (10) as

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

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

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

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 $t_L^* < t_L^a = t^a(\beta_L)$, $t^*(1) = t^a(1) = t^h(1) < t_L^h \leq t_L^*$, 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 IC and, therefore, $t^*(\beta_H) = t^c(\beta_H) > t^h(\beta_H)$. In this case, from (C.1)–(C.2), 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), and thus the desired cutoff cannot belong to that interval.

Case 2 If $\beta_H \geq \max\{\beta_L, \beta^u\}$, then the unconstrained-optimal mechanism for type H is IC 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 (B.1) with $b = t_L^*$. As shown in the proof of Corollary 2, because $\mathbf{P}$ satisfies MHRP, $\phi_{t_L^*}$ is strictly increasing and vanishes at a single point $\beta_{t_L^*}^u$, which defines the desired threshold $\beta_H^{ne}(\beta_L)$. That $\beta_H^{ne}(\beta_L) > \beta^u$ was shown in Case 1. That $\beta_H^{ne}(\beta_L)$ is strictly increasing in β_L follows from the fact that $t_L^* = t^*(\beta_L)$ and, thus, $\phi_{t_L^*}$ are strictly decreasing in β_L. Hence the result. ■

Proof of Proposition 3. A useful preliminary observation is that, because the mechanism designer always prefers a higher abstinence rate than the decision maker, we can, in analogy with the proof of Proposition 1, neglect constraints (21) and (23) in our quest for an optimal IC joint mechanism. That is, the following result holds.

Lemma C.1 *Any solution to the relaxed problem*

$$\max \left\{ \sum_{i=L,H} p_i \beta_i \{ t_i^h \mathbf{E}[\Pi_i(\theta)] - \mathbf{E}[\theta \Pi_i(\theta)] \} : \pi \text{ satisfies (20) and (22)} \right\} \tag{C.3}$$

is a solution to problem (24).

Proof. We show that any solution to (C.3) satisfies (21) and (23), and thus is a solution to (24). We accordingly distinguish two cases.

Case 1 Suppose, by way of contradiction, that a solution (π_0, π_L, π_{LH}) to (C.3) violates (21). 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 that type L abstain in this case, and a fortiori that type H abstain as $t_H^a < t_L^a$. Therefore, the joint mechanism $(\pi_0 + \pi_L, 0, \pi_{LH})$ would satisfy (20) and (22) and improve upon the solution to (C.3), a contradiction.

Case 2 Suppose, by way of contradiction, that a solution (π_0, π_L, π_{LH}) to (C.3) violates (23). 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 that type H abstain in this case. Therefore, the joint mechanism $(\pi_0, \pi_L + \pi_{LH}, 0)$ would satisfy (20) and (22) and improve upon the solution to (C.3), once again a contradiction. The result follows. ■

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

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

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 (20) and (22), as they issue recommendations to abstain to higher-risk types than any other joint mechanism with the same probabilities of consumption recommendations that also satisfies these constraints.

Lemma C.2 *For any joint mechanism $\pi = (\pi_0, \pi_L, \pi_{LH})$ that satisfies (20) and (22), there exists a joint mechanism $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_L, \tilde{\pi}_{LH})$ that also satisfies (20), (22), and such that*

$$\mathbf{E}[\tilde{\pi}_j(\theta)] = \mathbf{E}[\pi_j(\theta)], \quad j = 0, L, LH, \tag{C.4}$$

$$\tilde{\pi}_{LH}(\theta) = 1_{\{\theta < t_{\gamma_{LH}}\}} \tag{C.5}$$

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 (C.3) than π, and strictly so if π does not satisfy (C.5) on a $\mathbf{P}$-nonnull set.

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

persuasion mechanism

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

and let $\tilde{\pi} \equiv (\tilde{\pi}_0, \tilde{\pi}_L, \tilde{\pi}_{LH})$ be the corresponding joint mechanism. The direct joint persuasion mechanism $\tilde{x}$ is constructed such that recommendation probabilities are the same as under x, but consumption is recommended to both types if and only if $\theta \leq t_{\gamma_{LH}}$. Hence (C.4)–(C.5) hold by construction. Moreover, $\tilde{\pi}$ satisfies the incentive constraints (20) and (22), 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 (C.5) on a $\mathbf{P}$-nonnull set. The result follows. ■

Lemma C.2 implies that any solution $\pi^{**} = (\pi_0^{**}, \pi_L^{**}, \pi_{LH}^{**})$ to (C.3) is such that, for some cutoff t_{LH}^{**}, we have

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

up to a $\mathbf{P}$-null set. For any such joint 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 (C.3) reduces to finding a measurable function $\pi_L^{**} : [0,1] \to [0,1]$ that vanishes over $[0, t_{LH}^{**}]$ and that solves

$$\max\{t_L^h \mathbf{E}[\pi_L(\theta)] - \mathbf{E}[\theta \pi_L(\theta)] : \pi \text{ satisfies (20) and (22)}\}. \tag{C.6}$$

As in Section 3.2, the left-hand side of constraint (20) is not well-defined if $\pi_0 = 0$ $\mathbf{P}$-almost surely over $(t_{LH}^{**}, 1)$, and similarly the left-hand side of constraint (22) is not well-defined if $\pi_0 = 0$ $\mathbf{P}$-almost surely over $(t_{LH}^{**}, 1)$. To circumvent this problem, we again adopt the convention that the undefined constraint is emptily satisfied, which allows us to linearize the constraints (20) and (22). We start with an existence result.

Lemma C.3 *Problems (C.6), (C.3), and (24) have a solution.*

Proof. Our convention on the constraints (20) and (22) allows us to rewrite (C.6) as

$$\begin{aligned} \max\{t_L^h \mathbf{E}[\pi_L(\theta)] - \mathbf{E}[\theta \pi_L(\theta)] : \mathbf{E}[\theta[1-\pi_L(\theta)]] \geq t_L^a \mathbf{E}[1-\pi_L(\theta)] \\ \text{and } \mathbf{E}[\theta \pi_L(\theta)] \geq t_H^a \mathbf{E}[\pi_L(\theta)]\}, \end{aligned} \tag{C.7}$$

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 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 (C.7); 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 (C.7) is continuous in π_L by duality, (C.7) and, hence, (C.6) have a solution. To complete the proof, observe that, by Lemma C.1, we only need to show that (C.3) has a solution. Treating t^{**}_{LH} as a parameter, Berge's maximum theorem (Aliprantis and Border (2006, Theorem 17.31)) implies that the solutions to (C.6) 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 C.2, (C.3) 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 (C.6).

Lemma C.4 *Under Assumption 2, problem (C.6) has a solution of the form (26).*

Proof. We distinguish two cases.

Case 1 If constraint (22) is slack at the optimum, then (C.6) reduces to finding an optimal mechanism for type L alone, as described in Section 3. Proposition 1 yields 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 joint mechanism (18), which is IC if and only if the no-externality condition (19) holds.

Case 2 If constraint (22) is binding at the optimum, that is, according to Case 1, if the no-externality condition (19) does not hold, then

$$\frac{\mathbf{E}[\theta \pi_L(\theta)]}{\mathbf{E}[\pi_L(\theta)]} = t^a_H. \tag{C.8}$$

Plugging (C.8) into the objective of (C.6), the problem becomes[22]

$$\max\{(t^h_L - t^a_H)\mathbf{E}[\pi_L(\theta)] : \pi \text{ satisfies (20) and (C.8)}\}. \tag{C.9}$$

Our convention on the constraints (20) and (22) allows us to replace expectations in (C.9) by integrals, yielding the equivalent problem

$$\max\left\{(t^h_L - t^a_H)\int_{t^{**}_{LH}}^1 \pi_L(\theta) f(\theta)\,\mathrm{d}\theta : \int_{t^{**}_{LH}}^1 \theta[1-\pi_L(\theta)] f(\theta)\,\mathrm{d}\theta \geq t^a_L \int_{t^{**}_{LH}}^1 [1-\pi_L(\theta)] f(\theta)\,\mathrm{d}\theta \right.$$
$$\left. \text{and } \int_{t^{**}_{LH}}^1 \theta \pi_L(\theta) f(\theta)\,\mathrm{d}\theta = t^a_H \int_{t^{**}_{LH}}^1 \pi_L(\theta) f(\theta)\,\mathrm{d}\theta \right\},$$

where the maximum is taken over the set S defined in the proof of Lemma C.3. 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 by construction is a solution to (C.9) and (C.6), there exists a vector of Lagrange multipliers $(\eta^{**}, \lambda^{**}, \mu^{**})$ such that the following properties are satisfied:

[22] We keep the multiplicative constant $t^h_L - t^a_H$, which is strictly positive under Assumption 2, in order to make Lemma C.4 relevant when this assumption does not hold, as in Proposition 4.

- Nontriviality:

$$(\eta^{**}, \lambda^{**}, \mu^{**}) \neq (0, 0, 0). \tag{C.10}$$

- Positivity:

$$\eta^{**} \in \{0, 1\} \quad \text{and} \quad \lambda^{**} \in \mathbb{R}_+. \tag{C.11}$$

- Lagrangian maximization:

$$\pi_L^{**} \in \arg\max \left\{ \int_{t_{LH}^{**}}^1 h^{**}(\theta) \pi_L(\theta) f(\theta) \, \mathrm{d}\theta : \pi_L \in S \right\}, \tag{C.12}$$

  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:

$$\lambda^{**} \left\{ \int_{t_{LH}^{**}}^1 \theta [1 - \pi_L^{**}(\theta)] f(\theta) \, \mathrm{d}\theta - t_L^a \int_{t_{LH}^{**}}^1 [1 - \pi_L^{**}(\theta)] f(\theta) \, \mathrm{d}\theta \right\} = 0. \tag{C.13}$$

- Equality constraint:

$$\int_{t_{LH}^{**}}^1 \theta \pi_L^{**}(\theta) f(\theta) \, \mathrm{d}\theta = t_H^a \int_{t_{LH}^{**}}^1 \pi_L^{**}(\theta) f(\theta) \, \mathrm{d}\theta. \tag{C.14}$$

We distinguish four subcases.

Subcase 2.1 If $h^{**}(\theta) > 0$ for all $\theta \in (t_{LH}^{**}, 1)$, then the objective in (C.12) is uniquely (up to a $\mathbf{P}$-null set) maximized over S by

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

which corresponds to a cutoff $t_L^{**} = 1$ in (26). Notice that (C.13) is automatically satisfied and that (C.14) 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 mechanism.

Subcase 2.2 If $h^{**}(\theta) < 0$ for all $\theta \in (t_{LH}^{**}, 1)$, then the objective in (C.12) is uniquely (up to a $\mathbf{P}$-null set) maximized over S by

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

which corresponds to a cutoff $t_L^{**} = t_{LH}^{**}$ in (26). Notice that (C.14) is automatically satisfied, and that (C.13) becomes

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

Hence we must have $t_{LH}^{**} = t_L^c$ if $\lambda^{**} > 0$.

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

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

We claim that $\lambda^{**} + \mu^{**} > 0$. Indeed, if $\lambda^{**} + \mu^{**} < 0$, then the objective in (C.12) is uniquely (up to a $\mathbf{P}$-null set) maximized over S by

$$\pi_L^{**}(\theta) = 1_{\{\theta \geq t_L^{**}\}}, \tag{C.15}$$

so that

$$\pi_0^{**}(\theta) = 1_{\{t_{LH}^{**} < \theta < t_L^{**}\}}. \tag{C.16}$$

Now, given (C.16), (20) requires

$$\mathbf{E}[\theta \,|\, t_{LH}^{**} < \theta < t_L^{**}] \geq t_L^a. \tag{C.17}$$

However, we know from Lemma C.1 that any solution to (C.3) and, hence, to (C.6) and (C.9), is also a solution to (24). In particular, given (C.15), (21) requires

$$\mathbf{E}[\theta \,|\, \theta \geq t_L^{**}] < t_L^a. \tag{C.18}$$

Because (C.17)–(C.18) contradict each other, we obtain $\lambda^{**}+\mu^{**} > 0$, as claimed, and the objective in (C.12) is uniquely (up to a $\mathbf{P}$-null set) maximized over S by

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

once again in line with (26).

Subcase 2.4 Suppose finally that h^{**} is identically zero over $(t_{LH}^{**}, 1)$—so that, 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 (C.11); otherwise, by (C.11) again, $\eta^{**} = \lambda^{**} = \mu^{**} = 0$, which violates (C.10). Applying (C.11) yet again, we obtain $t_H^a \geq t_L^h$, with equality if and only if $\lambda^{**} = 0$. Hence this subcase cannot arise under Assumption 2. The result follows. ■

Proposition 3 is then an immediate consequence of Lemma C.4. Hence the result. ■

Proof of Lemma 3. We solve (C.3) for the optimal cutoffs (t_{LH}^{**}, t_L^{**})—the existence of which we established in Proposition 3—under the assumption that the individually optimal mechanisms with cutoffs t_H^* and t_L^* are not simultaneously implementable, that is, (19) does not hold. We first claim that we can restrict attention to cutoffs (t_{LH}, t_L) such that $t_L \geq t_L^*$. To prove this claim, we distinguish two cases. If $t_L^* > t_L^h$, then (20) is satisfied if and only if $t_L \geq t_L^*$. 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 (22) compared to $t_L = t_L^h$; hence an optimal cutoff t_L must satisfy $t_L \geq t_L^h$, which is IC as $t_L^h = t_L^*$. The claim follows. Replacing expectations in (C.3) by integrals then yields the equivalent problem

$$\max \left\{ p_L \beta_L \int_0^{t_L} (t_L^h - \theta) f(\theta)\, \mathrm{d}\theta + p_H \beta_H \int_0^{t_{LH}} (t_H^h - \theta) f(\theta)\, \mathrm{d}\theta \right\}, \tag{C.19}$$

subject to the constraints

$$\int_{t_{LH}}^{t_L} (\theta - t_H^a) f(\theta)\,\mathrm{d}\theta \geq 0, \tag{C.20}$$

$$t_L - t_L^* \geq 0, \tag{C.21}$$

$$1 - t_L \geq 0. \tag{C.22}$$

The objective in (C.19) is continuous in (t_{LH}, t_L) and the feasible set defined by $(t_{LH}, t_L) \in [0,1]^2$ and (C.20)–(C.22) is nonempty and compact. Hence problem (C.19)–(C.22) 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 (C.19)–(C.22). 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 (C.20) and strictly increase the objective in (C.19), 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 (C.20) is strictly increasing in $t_{LH} \in [0, t_H^a]$ and the objective in (C.19) is strictly increasing in $t_{LH} \in [0, t_H^h]$.

Step 2 We next verify that the constraints (C.20)–(C.22) 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 (C.20) is not binding. If (C.20) is binding, then the first line of $\nabla g(t_{LH}^{**}, t_L^{**})$ is

$$\mathrm{D}g_1(t_{LH}^{**}, t_L^{**}) \equiv \begin{pmatrix} (t_H^a - t_{LH}^{**}) f(t_{LH}^{**}) & (t_L^* - t_H^a) f(t_L^{**}) \end{pmatrix}.$$

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 (C.20) 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 (C.20) and (C.21) 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 (C.20) and (C.22) are binding, then it is optimal to have $t_{LH}^{**} = t_H^*$ by Proposition 1, 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 2, constraints (C.20)–(C.22) are qualified at any solution (t_{LH}^{**}, t_L^{**}) to (C.19)–(C.22). 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 the following properties are satisfied:

- Positivity:

$$(\zeta^{**}, \nu^{**}, \chi^{**}) \in \mathbb{R}^3_+. \tag{C.23}$$

- First-order conditions:

$$p_L \beta_L (t_L^h - t_L^{**}) f(t_L^{**}) + \zeta^{**}(t_L^{**} - t_H^a) f(t_L^{**}) + \nu^{**} - \chi^{**} = 0, \tag{C.24}$$
$$p_H \beta_H (t_H^h - t_{LH}^{**}) f(t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) f(t_{LH}^{**}) = 0. \tag{C.25}$$

- Complementary slackness:

$$\zeta^{**} \int_{t_{LH}^{**}}^{t_L^{**}} (\theta - t_H^a) f(\theta)\, \mathrm{d}\theta = 0, \tag{C.26}$$
$$\nu^{**}(t_L^{**} - t_L^*) = 0, \tag{C.27}$$
$$\chi^{**}(1 - t_L^{**}) = 0. \tag{C.28}$$

We distinguish three cases.

Case 1 Suppose first that (C.21) is binding, so that $t_L^{**} = t_L^*$ and $\chi^{**} = 0$ by (C.28), and suppose further, by way of contradiction, that $\zeta^{**} = 0$. Then, by (C.25) 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 $t_{LH}^{**} = t_H^h \leq t_H^*$. 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 (C.20) is violated at (t_{LH}^{**}, t_L^*), a contradiction. Hence, by (C.23), $\zeta^{**} > 0$, so that, by (C.26), (C.20) must be binding at (t_{LH}^{**}, t_L^*). That is, t_{LH}^{**} must satisfy

$$\int_{t_{LH}^{**}}^{t_L^*} (\theta - t_H^a) f(\theta)\, \mathrm{d}\theta = 0. \tag{C.29}$$

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 (C.23), the first-order conditions (C.24)–(C.25) rewrite as

$$p_L \beta_L (t_L^h - t_L^*) + \zeta^{**}(t_L^* - t_H^a) \leq 0, \tag{C.30}$$
$$p_H \beta_H (t_H^h - t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) = 0. \tag{C.31}$$

Because $\zeta^{**} > 0$ and $t_L^* \geq t_L^h > t_H^a$, (C.30) implies $t_L^* > t_L^h$. Hence the bracketed terms in (C.30) are different from zero. Moreover, because the bracketed terms in (C.31) cannot simultaneously be zero, none of them can be zero. Because $t_H^h \leq t_{LH}^{**} < t_H^a$ by Step 1, we can thus divide (C.31) by (C.30), which yields

$$\frac{p_H \beta_H}{p_L \beta_L} \frac{t_{LH}^{**} - t_H^h}{t_L^* - t_L^h} \leq \frac{t_H^a - t_{LH}^{**}}{t_L^* - t_H^a}. \tag{C.32}$$

Case 2 Suppose next that (C.22) is binding, so that $t_L^{**} = 1$ and $\nu^{**} = 0$ by (C.27). By Proposition 1, 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 (C.25) then rewrites as

$$p_H \beta_H (t_H^h - t_H^*) - \zeta^{**}(t_H^* - t_H^a) = 0, \tag{C.33}$$

so that $t_H^* > t_H^h$ if and only if $\zeta^{**} > 0$. If $f(1) > 0$, then, because $\chi^{**} \geq 0 = \nu^{**}$ by (C.23), we can also simplify (C.24) to obtain

$$p_L \beta_L (t_L^h - 1) + \zeta^{**}(1 - t_H^a) \geq 0. \tag{C.34}$$

The argument leading to (C.34) is a bit more involved if $f(1) = 0$. In that case, it follows from (C.24) 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 (C.34) holds. By (C.23) and (C.34), $\zeta^{**} > 0$, so that, by (C.26), (C.20) must be binding at $(t_H^*, 1)$. That is, t_H^* must satisfy

$$\int_{t_H^*}^{1} (\theta - t_H^a) f(\theta) \, \mathrm{d}\theta = 0, \tag{C.35}$$

which generically implies that $t_H^* > t_H^h$, so that the unconstrained-optimal mechanism for type H is not IC. The terms $t_H^* - t_H^a$ and $1 - t_H^a$ in (C.33)–(C.34) are by construction different from zero. We can thus divide (C.33) by (C.34), which yields

$$\frac{p_H \beta_H}{p_L \beta_L} \frac{t_H^* - t_H^h}{1 - t_L^h} \geq \frac{t_H^a - t_H^*}{1 - t_H^a}. \tag{C.36}$$

Case 3 Suppose finally that (C.21)–(C.22) are not binding, so that $\nu^{**} = \chi^{**} = 0$ by (C.27)–(C.28). 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 (C.24)–(C.25) then rewrite as

$$p_L \beta_L (t_L^h - t_L^{**}) + \zeta^{**}(t_L^{**} - t_H^a) = 0, \tag{C.37}$$

$$p_H \beta_H (t_H^h - t_{LH}^{**}) - \zeta^{**}(t_{LH}^{**} - t_H^a) = 0. \tag{C.38}$$

We must have $\zeta^{**} > 0$ and, hence, by (C.26), (C.20) must be binding, for, otherwise, by (C.37)–(C.38), we would have $t_{LH}^{**} = t_H^h$ and $t_L^{**} = t_L^h$, so that the individually unconstrained-optimal mechanisms for types H and L would be simultaneously implementable, a contradiction. That is, (t_{LH}^{**}, t_L^{**}) must satisfy

$$\int_{t_{LH}^{**}}^{t_L^{**}} (\theta - t_H^a) f(\theta) \, \mathrm{d}\theta = 0. \tag{C.39}$$

Because $t_L^{**} > t_{LH}^{**}$ by Step 1, it follows that the bracketed terms on the left-hand sides of (C.37)–(C.38) cannot be zero. Dividing yields

$$\frac{p_H \beta_H}{p_L \beta_L} \frac{t_{LH}^{**} - t_H^h}{t_L^{**} - t_L^h} = \frac{t_H^a - t_{LH}^{**}}{t_L^{**} - t_H^a}. \tag{C.40}$$

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, (C.20) is binding, see (C.29), (C.35), and (C.39). Let accordingly

$$\mathcal{T}_L \equiv \{t_L \geq t_L^* : \text{there exists } t_H \leq t_L \text{ such that } \mathbf{E}[\theta \,|\, t_H < \theta \leq t_L] = t_H^a\}. \tag{C.41}$$

Because $t_L^* > t_H^a$ and $\mathbf{E}[\theta \mid t_H^* < \theta \leq t_L^*] < t_H^a$ 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

$$\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L] = t_H^a \tag{C.42}$$

for all $t_L \in \mathcal{T}_L$. By (C.29), (C.35), and (C.39), given t_L^{**}, t_{LH}^{**} is uniquely pinned down by

$$t_{LH}^{**} = \hat{t}_{LH}(t_L^{**}). \tag{C.43}$$

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

$$\hat{t}'_{LH}(t_L) = -\frac{f(t_L)}{f(\hat{t}_{LH}(t_L))} \, \frac{t_L - \mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L]}{\mathbf{E}[\theta \,|\, \hat{t}_{LH}(t_L) < \theta \leq t_L] - \hat{t}_{LH}(t_L)} < 0. \tag{C.44}$$

While (C.43) holds in each of Cases 1, 2, and 3, these cases differ as to whether (C.32), (C.36), or (C.40) holds. Defining accordingly

$$\kappa(t_L) \equiv \frac{p_H \beta_H}{p_L \beta_L} \, \frac{\hat{t}_{LH}(t_L) - t_H^h}{t_L - t_L^h} - \frac{t_H^a - \hat{t}_{LH}(t_L)}{t_L - t_H^a}, \tag{C.45}$$

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

$$\begin{aligned}
\kappa'(t_L) &= \frac{p_H \beta_H}{p_L \beta_L} \left[\frac{\hat{t}'_{LH}(t_L)}{t_L - t_L^h} - \frac{\hat{t}_{LH}(t_L) - t_H^h}{(t_L - t_L^h)^2} \right] + \frac{\hat{t}'_{LH}(t_L)}{t_L - t_H^a} + \frac{t_H^a - \hat{t}_{LH}(t_L)}{(t_L - t_H^a)^2} \\
&< -\frac{p_H \beta_H}{p_L \beta_L} \, \frac{\hat{t}_{LH}(t_L) - t_H^h}{(t_L - t_L^h)^2} + \frac{t_H^a - \hat{t}_{LH}(t_L)}{(t_L - t_H^a)^2} \\
&= \frac{[t_H^a - \hat{t}_{LH}(t_L)](t_H^a - t_L^h)}{(t_L - t_L^h)(t_L - t_H^a)^2} \\
&< 0,
\end{aligned} \tag{C.46}$$

where the first inequality follows from (C.44), the second equality follows from (C.45) 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 (C.42). 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 $p \mapsto \frac{p}{1-p}$ is a strictly increasing continuous mapping from $(0,1)$ to $(0,\infty)$, 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 (C.45), 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 mechanism. To complete the proof, we only need to check that if $t_L^* = t_L^h$ and type L faces his individually optimal 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 (C.30) in Case 1 of the proof of Lemma 3, if we impose the constraint (C.20), 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 $p \mapsto \frac{p}{1-p}$ is a strictly increasing continuous mapping from $(0,1)$ to $(0,\infty)$, 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 (C.45), 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 mechanism. To complete the proof, we only need to check that if $t_H^* = t_H^h$ and type H faces his individually optimal 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 (C.33) in Case 2 of the proof of Lemma 3, $t_H^* = t_H^h$ implies $\zeta^{**} = 0$. Because $t_{LH}^{**} = t_H^*$ implies $t_L^{**} = 1$, (C.34) 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 (C.45), we thus have

$$\kappa(p_H, t_L^{**}) = \frac{p_H \beta_H}{(1-p_H)\beta_L} \frac{\hat{t}_{LH}(t_L^{**}) - t_H^h}{t_L^{**} - t_L^h} - \frac{t_H^a - \hat{t}_{LH}(t_L^{**})}{t_L^{**} - t_H^a} = 0 \tag{C.47}$$

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 (30). Let us accordingly denote by $\hat{t}_L(p_H)$ the unique solution to (C.47). We clearly have $\mathrm{D}_{p_H}\kappa(p_H, t_L) > 0$ and, from (C.46), $\mathrm{D}_{t_L}\kappa(p_H, t_L) < 0$ 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, for each $p_H \in (\underline{p}, \overline{p})$,

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

where $\hat{t}_{LH}$ is strictly decreasing over the interval $\mathcal{T}_L$ by (C.44), the probabilities $F(\hat{t}_{LH}(\hat{t}_L(p_H)))$ and $F(\hat{t}_L(p_H))$ that type H and type L consume, respectively, are strictly decreasing and strictly increasing in $p_H \in (\underline{p}, \overline{p})$, respectively. Hence the result. ■

Proof of Proposition 4. By Proposition 3, if Assumption 2 holds, then there exists an optimal IC joint mechanism of the form (31) with $\overline{t}_L^{**} = 1$. Suppose then that Assumption 2 does not hold. The result is immediate if we are in Case 1 or Subcases 2.1–2.3 of Lemma C.4; note, incidentally, that we can be in Subcase 2.2, which, according to Lemma 3, cannot arise under Assumption 2. There remains to consider Subcase 2.4 of Lemma C.4, in which the affine function h^{**} is identically zero over $(t_{LH}^{**}, 1)$.

Case 1 We first assume that $t_H^a > t_L^h$. Then, by arguments already invoked, $\lambda^{**} > 0$ and, by (C.13), any solution π_L^{**} to (C.6) must satisfy (C.14) and

$$\int_{t_{LH}^{**}}^1 \theta[1 - \pi_L^{**}(\theta)] f(\theta)\, \mathrm{d}\theta = t_L^a \int_{t_{LH}^{**}}^1 [1 - \pi_L^{**}(\theta)] f(\theta)\, \mathrm{d}\theta. \tag{C.48}$$

Notice that, because Lemma C.3 guarantees that a solution π_L^{**} to (C.6) exists, there exists a solution to (C.14) and (C.48). Conversely, because h^{**} is identically zero over $(t_{LH}^{**}, 1)$, any solution to (C.14) and (C.48) is a solution to the maximization condition (C.12) and, hence, to (C.6) as this is a convex problem and $\eta^{**} > 0$ (Clarke (2013, Exercise 9.7)). Let us then fix a solution π_L^{**} to (C.14) and (C.48). 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 of Lemma C.4. That is, we focus on the case where both constraints (20) and (22) in (C.3) are well-defined and binding. In particular, we must have

$$t_{LH}^{**} < t_H^a < \mathbf{E}[\theta \,|\, \theta > t_{LH}^{**}] < t_L^a. \tag{C.49}$$

Summing (C.14) and (C.48) and rearranging, we obtain that any solution to (C.14) and (C.48) satisfies

$$\int_{t_{LH}^{**}}^1 [1 - \pi_L^{**}(\theta)] f(\theta)\, \mathrm{d}\theta = \rho \equiv \frac{\mathbf{E}[\theta \,|\, \theta > t_{LH}^{**}] - t_H^a}{t_L^a - t_H^a} [1 - F(t_{LH}^{**})] < 1 - F(t_{LH}^{**}). \tag{C.50}$$

We claim that, in line with (31), there exists a solution to (C.14) and (C.48) of the form

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

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

$$\int_{\underline{t}}^{\bar{t}} \theta f(\theta)\,\mathrm{d}\theta = t_L^a[F(\bar{t}) - F(\underline{t})] \tag{C.51}$$

$$\int_{t_{LH}^{**}}^{\underline{t}} \theta f(\theta)\,\mathrm{d}\theta + \int_{\bar{t}}^{1} \theta f(\theta)\,\mathrm{d}\theta = t_H^a[F(\underline{t}) - F(t_{LH}^{**}) + 1 - F(\bar{t})], \tag{C.52}$$

has a unique solution. As above, summing (C.51)–(C.52) yields

$$F(\bar{t}) - F(\underline{t}) = \rho, \tag{C.53}$$

and, hence, (C.51) 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 (C.53) and that $\mathbf{P}$ has full support over $[0,1]$, and the second inequality follows from (C.49). 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

$$\frac{\int_{F^{-1}(1-\rho)}^{1} \theta f(\theta)\,\mathrm{d}\theta}{\rho} \geq t_L^a. \tag{C.54}$$

But we know that there exists a solution to (C.14) and (C.48), 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 (C.50), 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 (C.54). The claim follows. In case (C.54) holds as an equality, we have $\bar{t}_L^{**} = 1$, and π_L^{**} has the same form as in Subcase 2.3 of Lemma C.4.

Case 2 The proof for the limiting case $t_H^a = t_L^h$ or, equivalently, $\beta_H = \beta_H^{no}(\beta_L)$, relies on a simple continuity argument. From the proof of Lemma C.3, for each $\beta_H \geq \beta_H^{no}(\beta_L)$, any solution to (C.3) 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 $\beta_H^{no}(\beta_L)$. 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 (C.3) 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 $(t^{**}_{LH}(\beta^{no}_H(\beta_L)), \pi^{**}_L(\beta^{no}_H(\beta_L)))$ to (C.3) for $\beta^{no}_H(\beta_L)$. We can with no loss of generality assume that this sequence converges. For each $n \in \mathbb{N}$, we have $\beta_{H,n} > \beta^{no}_H(\beta_L)$ and, hence,

$$\pi^{**}_L(\beta_{H,n})(\theta) = 1_{\{t^{**}_{LH}(\beta_{H,n}) < \theta \leq t^{**}_L(\beta_{H,n})\}} \tag{C.55}$$

by Subcases 2.1–2.3 of the proof of Lemma C.4. Therefore,

$$\begin{aligned}\int \pi^{**}_L(\beta^{no}_H(\beta_L))(\theta)\,\mathbf{P}(\mathrm{d}\theta) &= \lim_{n\to\infty} \int \pi^{**}_L(\beta_{H,n})(\theta)\,\mathbf{P}(\mathrm{d}\theta) \\ &= \lim_{n\to\infty} F(t^{**}_L(\beta_{H,n})) - F(t^{**}_{LH}(\beta_{H,n})) \\ &= \lim_{n\to\infty} F(t^{**}_L(\beta_{H,n})) - F(t^{**}_{LH}(\beta^{no}_H(\beta_L))),\end{aligned} \tag{C.56}$$

where the first equality follows from the fact that the sequence $(\pi^{**}_L(\beta_{H,n}))_{n\in\mathbb{N}}$ converges in $B_{L_\infty(\mathbf{P})}$ to $\pi^{**}_L(\beta^{no}_H(\beta_L))$, using the definition of the weak* topology $\sigma(L_\infty(\mathbf{P}), L_1(\mathbf{P}))$, the second equality follows from (C.55), and the third inequality follows from the fact that the sequence $(t^{**}_{LH}(\beta_{H,n}))_{n\in\mathbb{N}}$ converges to $t^{**}_{LH}(\beta^{no}_H(\beta_L))$ in $[0,1]$ and that F is continuous as $\mathbf{P}$ is nonatomic. Because F is strictly increasing as $\mathbf{P}$ has full support, (C.56) 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 $[0,1]$,

$$\begin{aligned}\int_A \pi^{**}_L(\beta^{no}_H(\beta_L))(\theta)\,\mathbf{P}(\mathrm{d}\theta) &= \lim_{n\to\infty} \int_A \pi^{**}_L(\beta_{H,n})(\theta)\,\mathbf{P}(\mathrm{d}\theta) \\ &= \lim_{n\to\infty} \mathbf{P}\left[A \cap (t^{**}_{LH}(\beta_{H,n}), t^{**}_L(\beta_{H,n})]\right],\end{aligned} \tag{C.57}$$

using again the definition of the weak* topology $\sigma(L_\infty(\mathbf{P}), L_1(\mathbf{P}))$ along with (C.55). Finally, we can substitute $A = (t^{**}_{LH}(\beta^{no}_H(\beta_L), t_\infty]$ and $A = (t_\infty, 1]$ in (C.57) and use the fact that the sequence $((t^{**}_{LH}(\beta_{H,n}), t^{**}_L(\beta_{H,n})))_{n\in\mathbb{N}}$ converges to $(t^{**}_{LH}(\beta^{no}_H(\beta_L), t_\infty)$ to conclude that in fact $t_\infty = t^{**}_L(\beta^{no}_H(\beta_L))$ and

$$\pi^{**}_L(\beta^{no}_H(\beta_L))(\theta) = 1_{\{t^{**}_{LH}(\beta^{no}_H(\beta_L)) < \theta \leq t^{**}_L(\beta^{no}_H(\beta_L))\}}$$

up to a $\mathbf{P}$-null set. Hence the result. ■

Proof of Proposition 5. The proof consists in transforming our model into one to which the results of Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017) can be adapted, and then to apply those results to characterize the optimal mechanism. First, because the date-0 and date-1 selves share the same β and the same information about θ, the optimal consumption decisions x_0 and x_1 must coincide in our model, $x_0 = x_1 = x$. Second, we may identify the *receiver* as each self of the decision maker, with decision utility

$$u \equiv x(1 - \beta\delta^2 C\theta) = x\left(\frac{t^a - \theta}{t^a}\right),$$

and the *sender* as the mechanism designer, with intertemporal utility

$$v \equiv x[1 + \beta\delta - (1+\delta)\beta\delta^2 C\theta] = -x(1-\beta)\delta + (1+\delta)u.$$

Because a high action is good for high types in Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017)

and is denoted by a, we let $a \equiv 1 - x$ stand for abstention, and rewrite these utilities as

$$u \equiv a\left(\frac{\theta - t^a}{t^a}\right) + \frac{t^a - \theta}{t^a}$$

and

$$v \equiv a(1-\beta)\delta + (1+\delta)u - (1-\beta)\delta.$$

Any summand that does not depend on a does not enter any of the relevant optimization problems, and thus may be dropped without loss of generality. Similarly, the multiplicative factor $\frac{1}{t^a}$ does not affect the receiver's optimization problem, and can be omitted from its objective function u. However, as the sender does not know t^a, we cannot omit the multiplicative factor $\frac{1}{t^a}$ from her objective function v without implicitly changing the weight she puts on each private type.[23] We instead rescale v with a multiplicative factor $1 + \delta$. With a slight abuse of notation, we continue to denote the resulting transformed utilities by u and v, and obtain

$$u(a, \theta, t^a) \equiv a(\theta - t^a) \quad \text{and} \quad v(a, \theta, t^a) \equiv a\rho(t^a) + \frac{u(a, \theta, t^a)}{t^a}, \tag{C.58}$$

where

$$\rho(t^a) \equiv \frac{\delta}{1+\delta}(1-\beta) = \frac{\delta}{1+\delta}\left(1 - \frac{t_0}{t^a}\right).$$

These final expressions depend on β only through t^a. Now, the only differences with the persuasion problem studied in Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017) are that t^a is distributed on $[t_0, 1]$ rather than on $[0, 1]$, and that there is an additional factor $\rho(t^a)$ in the first summand of $v(a, \theta, t^a)$. The following result is a modification of their Lemma 2, accounting for these changes.

Lemma C.5 *For every incentive-compatible mechanism x, let $U^x(t^a)$ and $V^x(t^a)$ be the receiver's and the sender's expected interim utilities induced by x, respectively. Then the sender's expected utility is given by*

$$\int_{t_0}^1 V^x(t^a)\,\mathrm{d}H(t^a) = \int_{t_0}^1 U^x(t^a) J(t^a)\,\mathrm{d}t^a, \tag{C.59}$$

where $J(t^a) \equiv (\rho h)'(t^a) + \frac{h(t^a)}{t^a}$.

Proof. Define, for each $t^a \in [t_0, 1]$,

$$\tilde{v}(a, \theta, t^a) \equiv \frac{v(a, \theta, t^a)}{\rho(t^a)} = a + \frac{u(a, \theta, t^a)}{\rho(t^a) t^a}, \quad \tilde{h}(t^a) \equiv \frac{\rho(t^a) h(t^a)}{Z}, \quad \text{and} \quad \tilde{\rho}(t^a) \equiv \frac{1}{\rho(t^a) t^a},$$

where

$$Z \equiv \int_{t_0}^1 h(t^a)\rho(t^a) dt^a$$

is a normalizing constant that makes $\tilde{h}$ a density over $[t_0, 1]$. The reparameterized model written in terms of u, $\tilde{v}$, $\tilde{h}$ and $\tilde{\rho}$ is such that we can apply Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017, Lemma 2) to obtain

$$\int_{t_0}^1 V^x(t^a)\,\mathrm{d}H(t^a) = Z\int_{t_0}^1 \tilde{V}^x(t^a)\tilde{h}(t^a)\,\mathrm{d}t^a = Z\tilde{h}(t_0)\mathbf{E}[\theta] + \int_{t_0}^1 U^x(t^a) Z\tilde{J}(t^a) dt^a, \tag{C.60}$$

[23] Indeed, Lemma C.5 below accounts for this problem by appropriately rescaling the density h.

where $\tilde{J}(t^a) \equiv \tilde{h}'(t^a) + \tilde{\rho}(t^a)\tilde{h}(t^a)$ for all $t^a \in [t_0, 1]$. The first term on the right-hand side of (C.60) is zero as ρ and, hence, $\tilde{h}$ vanish at $t_0 = \frac{1}{\delta^2 C}$, while the second term can be rewritten using

$$Z\tilde{I}(t^a) = Z\left[\tilde{h}'(t^a) + \frac{\tilde{h}(t^a)}{\rho(t^a)t^a}\right] = (h\rho)'(t^a) + \frac{h(t^a)}{t^a},$$

which implies (C.59). The result follows. ■

To derive an optimal mechanism, it is key to understand the sign pattern of the function J defined in Lemma C.5. The following result provides sufficient conditions for J to be either always nonnegative or to cross zero at most once, from positive to negative.

Lemma C.6 *Suppose that h is log-concave with $h'(t_0) > 0$. Then, either J is non-negative over $[t_0, 1]$, or there exists $t_1 \in (t_0, 1)$ such that $J(t_1) = 0$ and $J(t) \gtrless 0$ if $t \lessgtr t_1$. The latter case obtains if $h(1) = 0$ and $h'(1) < 0$.*

Proof. A direct computation yields

$$J(t^a) = h(t^a)\left[\frac{1}{t^a} + \frac{1}{(1+\delta)\delta C(t^a)^2} + \frac{\delta}{1+\delta}\left(1 - \frac{t_0}{t^a}\right)\frac{h'(t^a)}{h(t^a)}\right]. \tag{C.61}$$

Clearly, if $h'(t_0) > 0$, then $J > 0$ in an open right-neighborhood of t_0. Similarly, if $h(1) = 0$ and $h'(1) < 0$, then $J(1) < 0$. In the latter case, the existence of a $t_1 \in (t_0, 1)$ such that $J(t_1) = 0$ follows from the intermediate value theorem. To conclude the proof, it thus remains to show that there can be at most one $t_1 \in (t_0, 1)$ such that $J(t_1) = 0$. We need to ensure that the term in square brackets in (C.61) crosses zero at most once and from above. Rearranging the condition that this term equal zero yields

$$-\frac{h'(t^a)}{h(t^a)} = \frac{(1+\delta)t^a + \frac{1}{\delta C}}{\delta t^a(t^a - t_0)}. \tag{C.62}$$

Because h is log-concave, the left hand-side of (C.62) is nondecreasing. Thus, because the right-hand side of (C.62) is strictly decreasing, we obtain that (C.62) can have at most one solution over $(t_0, 1)$. The result follows. ■

By Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017, Theorem 1), the receiver's utility profile U^x is implementable by a mechanism x if and only if U^x is a convex function that lies between his utility profiles from full information and no information. With no information, the receiver abstains if and only if $\mathbf{E}[\theta] > t^a$, in which case he obtains (abstention) utility $\mathbf{E}[\theta] - t^a$. His utility profile from no information is thus given by $\underline{U}(t^a) \equiv \max\{\mathbf{E}[\theta] - t^a, 0\}$. With full information, the receiver abstains if and only if $\theta > t^a$, in which case he obtains (abstention) utility $\theta - t^a$. His utility profile from full information is thus given by $\overline{U}(t^a) \equiv [1 - F(t^a)]\mathbf{E}[\theta - t^a \,|\, \theta > t^a]$. Clearly, $\underline{U}$ and $\overline{U}$ are nonincreasing convex functions that satisfy $\underline{U} \leq \overline{U}$, $\underline{U}(t_0) = \overline{U}(t_0)$, and $\underline{U}(1) = \overline{U}(1)$. Designing the optimal mechanism boils down to finding a convex function U^x that maximizes the right-hand side of (C.59) subject to the constraint $\underline{U} \leq U^x \leq \overline{U}$.

To complete the proof, we simply follow the logic in Kolotilin, Mylovanov, Zapechelnyuk, and Li (2017, Section 4.2), in particular the discussion below their Theorem 2 and their Example 1, which treats the case where J switches signs at most once, from positive to negative. They argue that the optimal utility profile U^x must be piecewise linear wherever it does not coincide with the upper bound $\overline{U}$. If J is everywhere nonnegative, the optimal U^x satisfies $U^x = \overline{U}$, so that a

full-information signal is optimal. If J switches sign once, from positive to negative, they show that there exists $\tilde{t} \in (t_0, 1]$ such that the optimal U^x satisfies $U^x(t^a) = \overline{U}(t^a)$ for $t^a \leq \tilde{t}$. For $t^a > \tilde{t}$, U_1^x continues as a tangent, that is, it is piecewise linear with slope $\overline{U}'(\tilde{t})$ until it hits $\underline{U}$, after which it coincides with $\underline{U}$. This choice of U^x corresponds to a mechanism that fully discloses each $\theta \leq \tilde{t}$, whereas all $\theta > \tilde{t}$ are pooled into a red warning label. As a result, decision makers with high self-control, $t^a \leq \tilde{t}$, consume whenever it is optimal for them to do so. Decision makers with intermediate self-control, $t^a \in (\tilde{t}, \mathbf{E}[\theta \,|\, \theta > \tilde{t}]]$, abstain from consuming whenever they observe a red warning, which leaves them to abstain more often than under full information. Finally, decision makers with $t^a > \mathbf{E}[\theta \mid \theta > \tilde{t}]$ have so little self-control that they always consume and obtain (abstention) utility 0. Hence the result. ■

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

[4] Aprahamian, H., D.R. Bish, and E.K. Bish (2019): "Optimal Risk-Based Group Testing," *Management Science*, 65(9), 4365–4384.

[5] 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.

[6] 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.

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

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

[9] Benkert, J.-M., and N. Netzer (2018): "Informational Requirements of Nudging," *Journal of Political Economy*, 126(6), 2323–2355.

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

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

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

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

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

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

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

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

[18] 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.

[19] 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.

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

[21] 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.

[22] Gentzkow, M., and E. Kamenica (2016): "A Rothschild-Stiglitz Approach to Bayesian Persuasion," *American Economic Review*, 106(5), 597–601.

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

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

[25] Guo, Y., and E. Shmaya (2019): "The Interval Structure of Optimal Disclosure," *Econometrica*, 87(2): 653–675.

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

[27] Habibi, A. (2020): "Motivation and Information Design," *Journal of Economic Behavior and Organization*, 169, 1-18.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] Kolotilin, A. (2015): "Experimental Design to Persuade," *Games and Economic Behavior*, 90, 215–226.

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

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

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

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

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

[42] 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.

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

[44] 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.

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

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

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

[48] 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.

[49] 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.

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

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

[52] Roesler, A.K., and B. Szentes (2017): "Buyer-Optimal Learning and Monopoly Pricing." *American Economic Review*, 107(7), 2072–2080.

[53] Schweizer, N., and N. Szech (2018): "Optimal Revelation of Life-Changing Information," *Management Science*, 64(11), 4967–5460.

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

[55] 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.

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

[57] 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.

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

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

[60] 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.

[61] 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.