games

Article

Call to Action: Intrinsic Motives and Material Interests

Vasileios Kotsidis

Faculty of Economics, University of Cambridge; Austin Robinson Building, Sidgwick Ave,
Cambridge CB3 9DD, UK; vk340@cam.ac.uk

Received: 30 August 2018; Accepted: 6 November 2018; Published: 14 November 2018

Abstract: We provide a game-theoretic account of endogenous intrinsic motivation within a principal–agent framework. We explore the incentives of an altruistic principal who, by exerting costly effort, can intrinsically motivate a present-biased agent to exhibit a direct preference for more far-sighted behaviour. We characterize the conditions under which this happens. We show that allowing for endogenous intrinsic motivation generates interesting interplays between exogenous economic incentives and endogenous motivation, including the possibility of crowding out. Our model can be applied in a wide variety of contexts, including public policy, self-control, and cultural transmission.

Keywords: endogenous preferences; present bias

JEL Classification: D01; D02; D64; D91; H30

1. Introduction

Standard economic theory postulates that preferences are given and immutable. Hobbes prompts us to think of humans as if they were mushrooms, attaining full development prior to engaging in any form of interaction with each other [1]. His position has been widely adhered to by traditional economic approaches. In the view of [2], tastes tend to be relatively stable and qualitatively similar across people. As such, they are prone to being considered as constant in the analysis of economic behaviour. This view of preferences can lead to important insights into the causal mechanisms underlying behaviour.

However, it is clear that the conception of stable, universal preferences is only a first-order approximation. Economists and other social scientists agree, to some extent, that people's preferences are moulded by a number of factors, including the influence of our parents (e.g., [3]), teachers (e.g., [4]), leaders (e.g., [5]), and others. Bowles remarks that thinking of preferences as fixed does result in the simplification of the task facing economists, but also compromises economic analysis in terms of explanatory power, relevance, and ethical consistency [6]. Indeed, to the extent that preferences are, even partly, affected by the environment where the individuals live and interact, the implications for economic theory and the design of public policy can be quite significant.

This paper therefore takes an alternative approach, and focuses on the incentives of an altruistic principal who, by exerting costly effort, can intrinsically motivate an agent to exhibit a direct preference for a given behaviour. For instance, a parent may wish to instil in her child a preference for hard work. A teacher may wish to instil in his pupil a preference for studying hard. A doctor may wish to instil in her patient a preference for a healthy lifestyle. An individual himself may want to cultivate a preference for a principled stance, if only as a commitment device. Such instances are indicative of a more general mechanism of preference formation, which this paper aims to model.

Our approach, owing to its particular incentive structure, departs from the standard principal–agent model, which has been extensively developed and deployed to account for agency problems.[1] Specifically, we assume that the (altruistic) principal cares, to some extent, about the material welfare of the agent.[2] We show that this feature of our design gives rise to interesting links between the underlying economic environment and the principal's optimal action.

A crucial ingredient of the setup is that the agent is present-biased.[3] This means that he is characterised by an excessive tendency to adopt behaviours that generate short-term rewards and involve long-term costs. In other words, he assigns a very high weight on present outcomes, to the detriment of his future welfare. Present bias is an increasingly acknowledged notion in the economics literature. Following the seminal contributions of Ainslie in the domain of temptation and self-control (see [12,13]), many experimental studies have documented the phenomenon in economics (Refs. [14,15] are two recent examples). This led to a growing literature of formal accounts that have established the phenomenon as a feature of people's preferences (see, e.g., [16–19]; see [20,21] for overviews and discussion).[4]

In our framework, the element of present bias is something that the individuals cannot address directly. For this reason, if the agent is left to his own devices, he will be prone to making myopic choices. To counteract this tendency, the altruistic principal has an incentive to imbue the agent with an intrinsic preference.[5] By intrinsic, we mean a preference that is defined directly over actions rather than the outcomes these actions imply. In other words, an intrinsic preference expresses a direct tendency to behave in a particular way, rather than a taste for the resulting payoff of that behaviour.

We firstly explore the conditions under which the principal in our scenario will optimally endow the agent with such a preference, in order to mitigate the distortionary effects of the agent's present bias. We then explore how the principal's incentive varies with the parameters of our model. We show that allowing for endogenous intrinsic preferences generates interplays between extrinsic economic incentives and intrinsic motivation that can exert significant effects on behaviour. In particular, a standard analysis would predict that the short-sighted behaviour will be less common in environments where it is objectively less attractive (e.g., because it generates, on average, higher long-term costs). This intuition may fail once endogenous preferences are factored in. The reason is that, in these environments, the principal's incentive to instil a direct preference against the short-sighted behaviour is weaker.

The scope of our analysis is quite broad. There are numerous principal–agent relationships of the type we describe here.[6] Parents may wish for their children to attain financial independence. Health-insurance providers may want to encourage their members to abstain from unhealthy habits. An individual may yearn for a life of affluence. We explore some alternative readings in Section 3 and discuss the implications of our analysis using specific examples.

The remainder of the paper proceeds as follows. Section 2 contains the setup and analysis of our model. In Section 3, we discuss some of its potential applications. Section 4 provides conclusions.

2. Model

2.1. Principal–Agent Setup

Consider a two-player sequential game, $\mathcal{G}$, spanning across three periods, denoted by $t \in 0, 1, 2$. The first player, the principal (P), moves at $t = 0$. The second player, the agent (A), observes the

[1] For a detailed exposition of the theory, see [7]. Ref. [8] provides a thorough critical overview of its applications in contracts, while Ref. [9] discusses its potential and limits in the domain of executive compensation.

[2] In this capacity, our model is conceptually close to that of [10], who harnesses the notion of *imperfect empathy*, proposed by [11].

[3] In our framework both individuals exhibit present-biased preferences, but it is the present bias of the agent that is incentivising the principal.

[4] Present bias also has a theoretical rationale as a feature of humans' preferences in an evolutionary sense. If the information reception and processing mechanisms of humans are imperfect (as in the context of [22]), then their uncertainty about the environment may induce them to place a lot of weight on present consumption.

[5] It is interesting to note that the principal has such an incentive even though she exhibits present bias herself.

[6] We are grateful to an anonymous referee for pointing out several such instances.

principal's move and subsequently makes his own, at $t = 1$.[7] The agent must select an action, $\gamma \in \{B, F\}$ (smoke/do not smoke, be extravagant/be thrifty, break/follow the law, etc.). Each of these two actions yields a consumption payoff for the agent. The consumption payoff of action F is normalised to zero.[8] Selecting action B generates an immediate consumption benefit, $b_1 \in \mathbb{R}^{++}$, as well as a delayed cost, $b_2 \in \mathbb{R}^{++}$.[9]

The agent decides with the aim to maximise his utility, which is given by the present discounted value of his consumption payoff over periods 1 and 2, as well as a hedonic component, which is manipulated by the principal (more on this later). There is no hedonic component associated with action B. By choosing F, on the other hand, the agent experiences a (net) degree of intrinsic gratification, denoted by $n \in \mathbb{R}^{+}$. We will refer to n as the level of 'warm glow' player A is conditioned to experience.[10]

Definition 1. *Warm glow: The degree of direct preference, n, for action γ is the level of intrinsic (non-material) utility player A receives upon choosing γ. This is additional to the material payoff resulting from action γ, but relevant only to the 'conditioned' agent, i.e., player A.*

Recall that n is associated with action F. This entails, again, no loss of generality: In principle, agent A can be conditioned to experience some degree of warm glow for each of the two actions he performs, with n describing their difference.[11] The problem of the agent, then, becomes obvious: He has to judge the relative appeal of each option, given their material costs and benefits, as well as the hedonic component n he has been conditioned to experience.

As stated before, the principal moves first, at $t = 0$. Her objective is to maximise the joint welfare of herself and the agent. To do so, she chooses a value for $n \in \mathbb{R}^{+}$. However, indoctrination comes at a cost. This is captured by a function $C : \mathbb{R}^{+} \rightarrow \mathbb{R}^{+}$, which associates each action available to the principal with a material loss she has to incur to take that action. We postulate that no such loss occurs by default, i.e., $C(0) = 0$. We also assume that this loss is increasing linearly in the degree of the principal's interference and, in particular, that $C^{'}(n) = \frac{dC(n)}{dn} = c > 0$. The linearity assumption here is only imposed for simplicity. Our results would be no different in a qualitative sense under an exponentially increasing cost function.[12]

An important element of our framework is that the preferences of both players are present-biased. Specifically, let $\delta \in (0, 1]$ denote the standard, compounding discount factor common to both the principal and the agent. Furthermore, let α denote the present-bias term of the principal and β represent that of the agent, where $\alpha, \beta \in (0, 1]$. Parameter α (β) measures the additional weight by which the principal (agent) discounts *all* future consequences relative to present ones.[13]

Notice that, in this framework, the case where the principal and the agent are different temporal versions of the same individual (the *intertemporal self*) involves $\alpha = \beta$. Notice, further, that we are agnostic as to which of the two players is more biased toward their present. In particular, from P's

[7] The two players may very well be different temporal versions of the same agent (i.e., the same person at different time-periods), but in general this need not be the case.

[8] This is without loss of generality. Given any $\pi_{\bar{t}}^{A}(F)$ and $\pi_{\bar{t}}^{A}(B)$ in some $\bar{t} \in \{1,2\}$, where $\pi_{\bar{t}}^{A}(.)$ is the material-payoff function of agent A in period $\bar{t}$, subtracting $\pi_{\bar{t}}^{A}(F)$ from both will not alter A's decision.

[9] The same relationship could have been achieved by restricting both b_1 and b_2 to be negative. Indeed, the important element is that they are of the same sign. Consider this alternative case, where a present loss is weighted against a future benefit. It is straightforward to show that this scenario is a reflection of ours. Owing to the symmetric structure of the analysis, our results are invariant across the two.

[10] See also [23–28].

[11] This interpretation is relevant to situations where the action space for A contains more than two distinct elements, which we will not examine here. Indeed, due to the associated cost, when two actions are available to agent A, attaching some value of n to one of them only is optimal.

[12] Indeed, $C(n)$ is assumed weakly convex for our proofs in Appendixes A and B.

[13] We say that the players exhibit quasi-hyperbolic, time-inconsistent preferences.

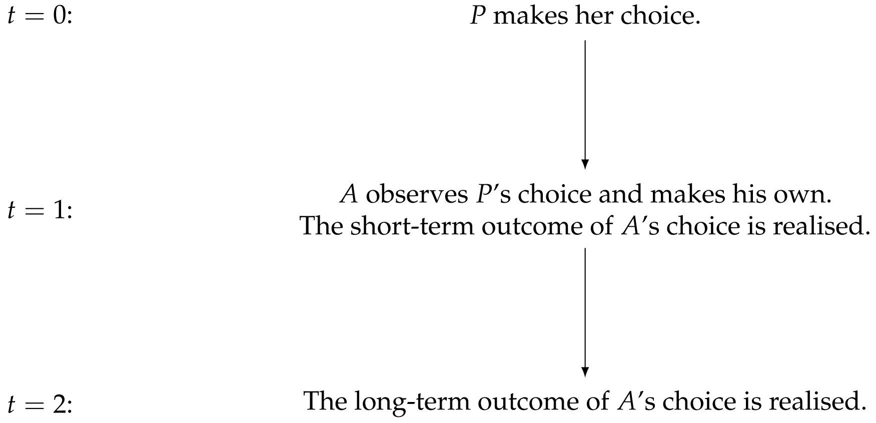
point of view at $t = 0$ all consequences of A's actions materialise in the future and are therefore uniformly discounted by α. Then, P's utility function, evaluated at $t = 0$, can be stated as:

$$U^P = \begin{cases} \alpha\delta(b_1 - \delta b_2) - C(n), & \text{if } \gamma = B, \\ -C(n), & \text{if } \gamma = F. \end{cases} \tag{1}$$

On the other hand, in making his decision, A has to consider present payoffs in addition to future ones. His utility function, evaluated at $t = 1$, can be stated as:

$$U^A = \begin{cases} b_1 - \beta\delta b_2, & \text{if } \gamma = B, \\ n, & \text{if } \gamma = F. \end{cases} \tag{2}$$

To summarise, in the period $t = 0$, the principal selects $n \in \mathbb{R}^+$ so as to maximise the joint utility of herself and the agent, evaluated according to her preferences at that time. The agent observes the principal's move and subsequently makes his own choice, at $t = 1$, between actions F and B. The agent's choice yields both a short- and a long-term outcome. The short-term outcome is realised immediately upon his choice, i.e., at $t = 1$. The long-term outcome is realised in the following period, i.e., at $t = 2$. A timeline of the events is provided in Figure 1.



$t = 0$: P makes her choice.

$t = 1$: A observes P's choice and makes his own.
The short-term outcome of A's choice is realised.

$t = 2$: The long-term outcome of A's choice is realised.

Figure 1. Timeline of events

Throughout our analysis, we apply backward induction to solve the game. The principal knows that, at time $t = 1$, the agent will choose, between actions B and F, the one that maximises his utility, as expressed in Equation (2). She therefore knows the outcome of every sub-game that starts with the agent making a choice. Then, at $t = 0$, the principal will choose the value of n that maximises her utility, as expressed in Equation (1) (or (4), see Section 2.3 below), given the agent's profile of choices. The natural solution concept for this problem is thus the Subgame-Perfect Nash Equilibrium (SPNE).

The fact that the players' preferences are present-biased can create a conflict of interest. This is because the principal does not internalise fully the agent's preferences, but instead applies imperfect empathy. That is, she evaluates the agent's material payoff through the lens of her own preferences at $t = 0$.[14] As a result, from the point of view of the principal, the agent appears impulsive. His present bias may lead him to choose action B, whereas he would be better off, in P's own assessment, were he to opt for F instead. To correct for this bias, given her inability to eliminate it directly, she can opt instead to imbue A with a direct preference for one of the actions.

[14] The general principle of imperfect empathy is quite standard in the principal–agent literature (see [11]).

The inability of the principal to address the present-bias problem of the agent directly is, indeed, critical for our account. At first glance, this might seem arbitrary. Why should the principal not simply invest in eliminating this feature from the agent's preferences? One argument is that our model would still apply in a situation where the principal could indeed influence β, but only to some extent or at too high a cost. A stronger argument can be made about the nature of each source of motivation. In our discussion in Section 1, we have described present bias as an innate characteristic, an impulse similar to the drive for profit. Such an impulse may emerge as an evolutionarily optimal feature of preferences under certain conditions. By contrast, we have described the principal's intervention as a form of indoctrination. That is, the principal is able to interfere with the agent's preferences to some extent, but, by *instilling a preference for a particular action*, rather than *eliminating an impulse*. It is this distinction that renders our account relevant. As a final point, such constraints are common in this literature (see e.g., [22] on the constraints in information processing).

Equations (1) and (2) highlight the potential for discrepancy between the choice favoured by the principal and the one the agent prefers. To see this, consider the following example, where $n = 0$. Here, P would prefer A to choose B iff:

$$U^P(0, B) \geq U^P(0, F) \Rightarrow b_1 - \delta b_2 \geq 0 \Rightarrow b_2 \leq \frac{b_1}{\delta}.$$

On the other hand, A will opt for B iff:

$$U^A(0, B) \geq U^A(0, F) \Rightarrow b_1 - \beta\delta b_2 \geq 0 \Rightarrow b_2 \leq \frac{b_1}{\beta\delta}.$$

Thus, the agent would switch from B to F at a higher threshold value for b_2. From the point of view of the principal that would be sub-optimal. In simple terms, P would like A to opt for action F so long as the period-1 gain from switching to B falls short of the period-2 cost discounted by δ. However, A would need a period-2 cost of at least $\frac{b_1}{\beta\delta}$ in order to be induced to switch to F. That would render him impulsive in P's opinion: due to his present-biased preferences, he would assign an unreasonably high weight on his period-1 utility. This situation, where the principal does not interfere with the agent's preferences at all ($n = 0$), is illustrated in Figure 2.

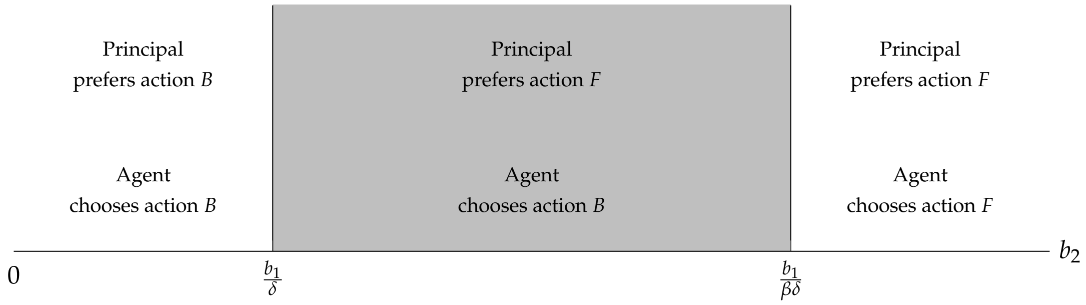


Figure 2. $n = 0$, no preference for a particular action.

Suppose now that the principal chooses instead to instil a direct preference for action F at $t = 0$, so that $n > 0$. That will induce the agent to lower his threshold for switching from B to F. Recall that n is entirely inconsequential and, thus, of no value to the principal. That is, P chooses to condition A to receive some warm glow from action F that is additional to its material consequences.[15] This does not

[15] Notice that in our characterisation the degree of intrinsic utility assigned to an action is *dependent* on its material consequences. The level of n is chosen by the principal in order to account for the agent's present bias and not because she believes that it has any real value. One way to think about this instrumentalist approach would be to consider that P can assign n to virtually any action available to A, so long as, given the underlying parameter values, she prefers it more than he does. Note, however, that for A this additional value *is* meaningful, in the sense that his utility increases by n whenever he performs the associated action.

affect the material consequences implied by the choices available to the agent or his present bias, but it does affect his utility. In this way, it counteracts his impulse and brings his preferences closer to those of the principal. The resulting situation looks like the one depicted in Figure 3.

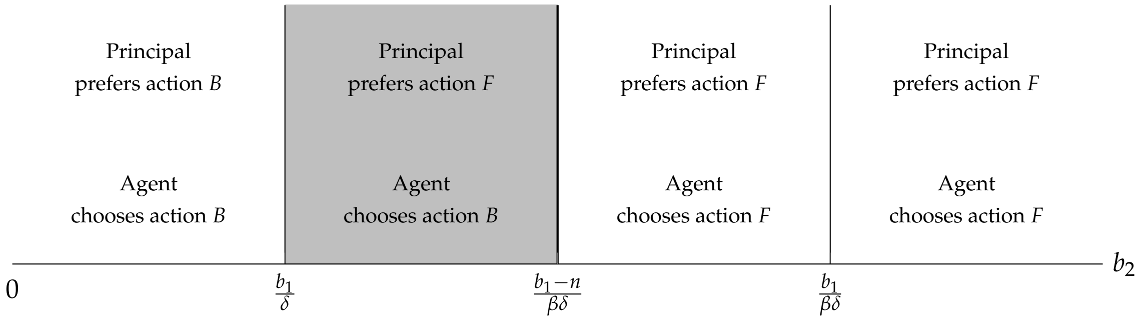


Figure 3. $n > 0$ assigned on action F.

Notice that so far the magnitudes of b_1 and b_2 are both deterministic, that is, there is no uncertainty associated with any of them. We start from this case, in Section 2.2, because it is useful as a basis for comparison. In Section 2.3, we consider a more realistic scenario, by allowing for uncertainty over b_2.

2.2. Baseline

Some important remarks are in order. To start with, notice that the principal would have no incentive to set $n > (1-\beta)b_1$, as that would not only be more costly for her, but also counter-productive. Indeed, such a value for n would induce the agent to choose action F even in instances where the principal would want him to opt for B. In addition, the principal would have no incentive to instil a preference for action B instead.[16] Doing so would also be counter-productive, as it would increase the discrepancy between the two players' preferences.

Lastly, it can be easily shown that the sequences of actions in Figures 2 and 3 would be reversed if it was the case that $b_1, b_2 < 0$. That is, if action B led to a present cost and a future benefit, then both players would favour F for $|b_2| \leq |\frac{b_1}{\delta}|$ and both would choose B for $|b_2| \geq |\frac{b_1}{\beta\delta}|$. For $|b_2| \in (|\frac{b_1}{\delta}|, |\frac{b_1}{\beta\delta}|)$, they would disagree, with the principal favouring B and the agent choosing F. Then, the former would find it optimal to assign $n > 0$ to action B. Taking these observations into account, we can form the following proposition.

Proposition 1. *In any equilibrium of game $\mathcal{G}$, $n \in [0, (1-\beta)b_1)$.*

Proof. Formally, this can be proved by contradiction. Consider first the case where $b_1, b_2 > 0$ and, thus, P assigns n to action F.

i. Suppose $n < 0$: Then, $\forall b_2 \in [\frac{b_1}{\beta\delta}, \frac{b_1-n}{\beta\delta})$ it would be true that $\frac{b_1-n}{\beta\delta} - b_2 > 0$. Thus, A would choose action B and P would have been better off setting $n = 0$.
ii. Suppose $n > (1-\beta)b_1$: Then, $\forall b_2 \in (\frac{b_1-n}{\beta\delta}, \frac{b_1}{\delta}]$ it would be true that $\frac{b_1-n}{\beta\delta} - b_2 < 0$. Thus, A would choose action F, even though P would prefer action B. Therefore, P would have been better off setting $n = (1-\beta)b_1$.

[16] In this paper, we focus on positive values for n in an effort to determine the action that will be *chosen*, as opposed to that which will be *avoided*. The two are equivalent in our framework, where the agent faces a binary-choice problem. However, in a situation with three or more available actions, assigning a negative n to an action (aversion towards a certain type of behaviour) does not generally ensure that the desired action will be chosen. A comparison between the cost of discouraging certain types of behaviour and that of encouraging others is an interesting research project itself. We leave this for the future and focus instead on positive education (encouragement of a particular behaviour).

iii. Suppose $n = (1-\beta)b_1$: For $b_2 \in [\frac{b_1}{\delta}, \frac{b_1}{\beta\delta})$ A would choose action F, in line with P's preferences. If $b_2 = \frac{b_1}{\delta}$, P would be indifferent between actions F and B, as they would both result in $U^A = 0$. Setting $n = (1-\beta)b_1$ would render A indifferent between the two actions at a positive cost to P. Thus, P would be better off setting n slightly below $(1-\beta)b_1$, so as to avoid the unnecessary expenditure in the case where $b_2 = \frac{b_1}{\delta}$.

An equivalent argument holds in the case where $b_1, b_2 < 0$ and P attaches n on action B. □

Proposition 1 describes the upper and lower bound for n. In simple terms, it determines the values of n for which it makes sense for the principal to consider.

Consider now the situation outlined in Section 2.1 from the principal's perspective at $t = 0$. The principal knows that in period 1 the agent will choose based on:

$$n \gtreqless b_1 - \beta\delta b_2 \Rightarrow b_2 \gtreqless \frac{b_1 - n}{\beta\delta}$$

If the future cost from action B is such, that the preferences of the agent are at odds with those of the principal, then the latter may find it optimal to invest in n. In other words, if $\frac{b_1}{\delta} < b_2 < \frac{b_1}{\beta\delta}$, then P may optimally assign $n > 0$ on action F, so as to induce A to choose it at $t = 1$. This depends on the cost of doing so. To simplify the analysis, suppose that, when the agent's preferences render him indifferent between the two options, he always chooses action F. Then, the various different cases are summarised in the following proposition.

Proposition 2. *Given game $\mathcal{G}$ with $b_1, b_2 > 0$, P assigns n^* to action F such, that:*

i. *if $b_2 > \frac{b_1}{\beta\delta}$, then $n^* = 0$ and A will choose action F.*

ii. *if $b_2 < \frac{b_1}{\delta}$, then $n^* = 0$ and A will choose action B.*

iii. *if $\frac{b_1}{\delta} < b_2 < \frac{b_1}{\beta\delta}$, then* $n^* = \begin{cases} b_1 - \beta\delta b_2 & \textit{if } \frac{C(b_1 - \beta\delta b_2)}{\delta} < \delta b_2 - b_1 \\ & \textit{and A will choose action F.} \\ 0 & \textit{if } \frac{C(b_1 - \beta\delta b_2)}{\delta} > \delta b_2 - b_1 \\ & \textit{and A will choose action B.} \end{cases}$

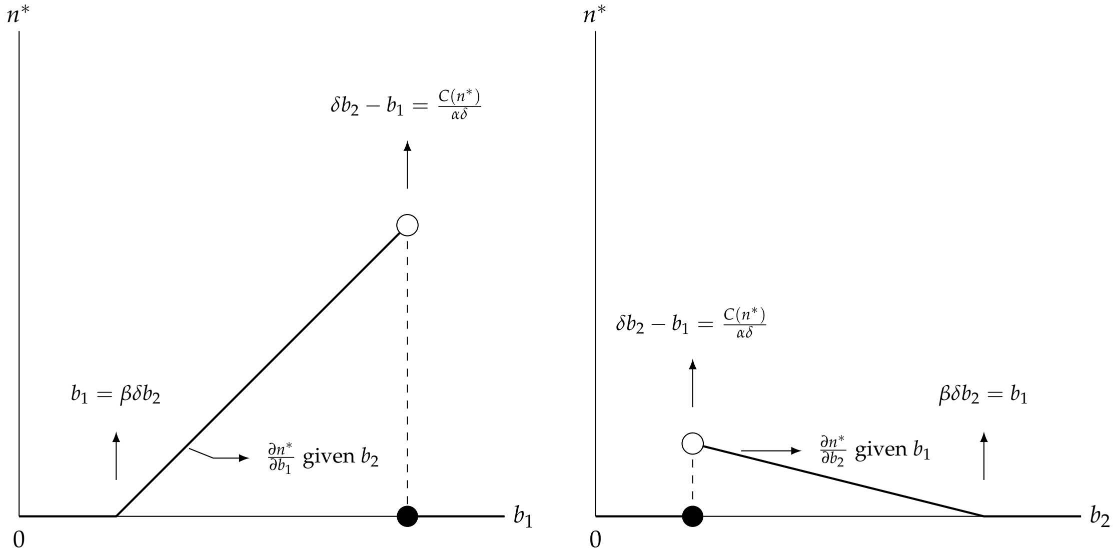
Proof. The proof of this proposition is straightforward. Trying to maximise their joint welfare, the principal compares the material gain that results from n^* with its cost. When both players agree on which action the agent should take, there is no need for investment in n. When they do not, if $n^* > 0$, then it is precisely such that it makes the agent indifferent between F and B (given the assumption stated above, that in such cases the agent opts for F). Any higher or lower value would incur an additional cost to the principal with no added benefit. Thus. the principal has to compare what she gets by setting $n^* = b_1 - \beta\delta b_2$ with the cost, $C(b_1 - \beta\delta b_2)$, of doing so. If the benefit surpasses the cost, then n^* is set equal to $b_1 - \beta\delta b_2$; otherwise, it is set equal to 0. □

The instrumental view of indoctrination championed in our paper gives rise to a rich structure of variations. Recall that the level of n the principal optimally attaches onto an action is dependent on the material consequences implied by that action relative to those implied by the alternative. In our simple scenario, the value of n attached on action F varies with the net benefit/cost of action B. The latter is expressed as a comparison between b_1 and b_2, evaluated at $t = 0$. The following corollaries summarize how changes in these two parameters affect n^*.

Corollary 1. *Consider game $\mathcal{G}$ with $b_1, b_2 > 0$ and n^* assigned on action F. The relationship between n^* and b_1 is non-monotonic. That is, $\exists\ \bar{b}_1 : n^*_{\hat{b}_1} = 0\ \forall\ \hat{b}_1 \geq \bar{b}_1,\ n^*_{\tilde{b}_1} < n^*_{\breve{b}_1}\ \forall\ \tilde{b}_1 < \breve{b}_1 < \bar{b}_1$. In particular, an increase in b_1 will encourage P to increase the level of n^* at a one-to-one rate, so long as b_1 remains lower than $\delta b_2 - \frac{C(n^*)}{\alpha\delta}$. If b_1 becomes equal to or greater than $\delta b_2 - \frac{C(n^*)}{\alpha\delta}$, the value of n^* will drop to zero.*

Corollary 2. *Consider game $\mathcal{G}$ with $b_1, b_2 > 0$ and n^* assigned on action F. The relationship between n^* and b_2 is non-monotonic. That is, $\exists\ \bar{b}_2 : n^*_{\hat{b}_2} = 0\ \forall\ \hat{b}_2 \leq \bar{b}_2,\ n^*_{\tilde{b}_2} > n^*_{\breve{b}_2}\ \forall\ \bar{b}_2 < \tilde{b}_2 < \breve{b}_2$. In particular, an increase in b_2 past $\frac{1}{\delta}\big(b_1 + \frac{C(n^*)}{\alpha\delta}\big)$ will encourage P to decrease n^* at a rate lower than one-to-one (equal to $\beta\delta$), unless n^* is already equal to zero. For b_2 values lower than or equal to $\frac{1}{\delta}\big(b_1 + \frac{C(n^*)}{\alpha\delta}\big)$, n^* will be equal to zero.*

An increase in b_1 increases the appeal of action B to the agent. Therefore, if the principal still thinks that opting for it is non-optimal, she will need to invest in a higher level of n to avert it. As b_1 increases, there comes a point where such an investment is sub-optimal from the principal's point of view: what the agent gains by choosing F is not enough to justify the cost of n necessary to induce him to do so. From that point onward, the only sensible option for the principal is to set $n = 0$. A fall in b_2 also increases the appeal of action B to the agent. Therefore, again, the principal needs to increase n to ensure that the agent will opt for F. As b_2 keeps dwindling, there comes a point where the net material benefit of B does not cover the cost of her investment (as evaluated by the principal). From that point onward, further reductions in b_2 will not change n from zero. Figure 4a,b illustrate these two cases.


n^*
$\delta b_2 - b_1 = \frac{C(n^*)}{\alpha\delta}$
$b_1 = \beta\delta b_2$
$\frac{\partial n^*}{\partial b_1}$ given b_2
0
b_1
n^*
$\delta b_2 - b_1 = \frac{C(n^*)}{\alpha\delta}$
$\beta\delta b_2 = b_1$
$\frac{\partial n^*}{\partial b_2}$ given b_1
0
b_2

(**a**) (**b**)

Figure 4. (**a**) relationship between n^* and b_1 given b_2 and $C(n)$: so long as there is a conflict of preferences between P and A and the cost of indoctrination is sufficiently low, n increases in b_1; (**b**) relationship between n^* and b_2 given b_1 and $C(n)$: given that there is conflict of preferences between P and A and the cost of indoctrination is sufficiently low, n decreases in b_2.

We can describe the variations in n^*, the optimal level of n, as responding to variations in the principal's total utility. Recall that her utility depends on hers and the agent's joint material payoff. This, in turn, is determined by her decision on n and the agent's choice between actions F and B. Based on our previous analysis, the optimal value for n will be either equal to zero or such that will render the agent exactly indifferent between F and B. This is true for any pair of values, b_1 and b_2, preference parameters, δ, α, and β, and linear cost function, $C(n)$. We can, thus, describe n^* as a function of the difference in P's utility between the following two combinations of choices:

$$dU^P \equiv U^P(\bar{n}, F) - U^P(0, B) = \delta b_2 - b_1 - \frac{C(\bar{n})}{\delta}, \quad \bar{n} > 0. \tag{3}$$

Figure 5 illustrates how changes in dU^P affect n^*. It is worth noting that n^* attains its highest levels in our framework for dU^P values close to zero. This is true when the net cost from action B, as evaluated by the principal at $t = 0$, is only marginally higher than the cost of the level of n necessary to avert it. In other words, a relatively high n is necessary when action B is sub-optimal, but only just so.

We now turn to examine the case where the principal does not know b_2 ex ante, only that it follows a certain distribution, $\mathcal{F}(b_2)$.

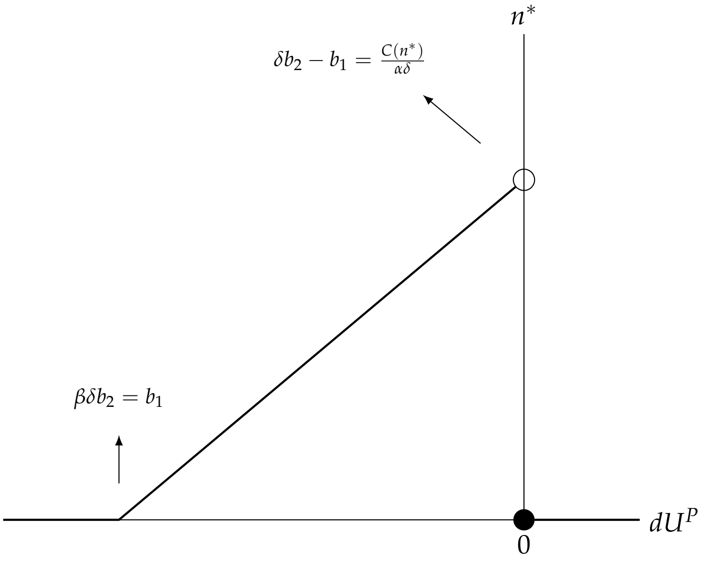
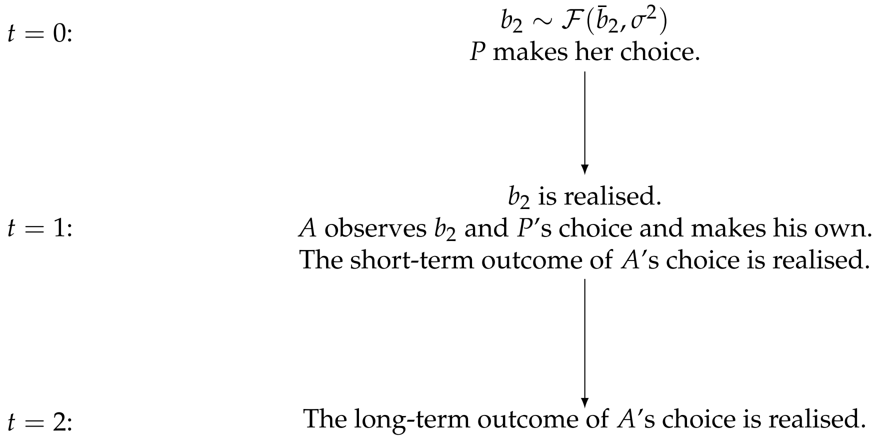


Figure 5. Relationship between n^* and dU^P: Indoctrination is at its highest when its net benefit is only marginal.

2.3. Probabilistic Future Cost

In this sub-section, we allow for some information asymmetry to arise over the value of b_2, the future consequence of action B. Specifically, the principal is now unaware of the actual value of b_2 when she makes her decision. She only knows that it follows a specific distribution, with a positive mean and a certain variance. The agent, on the other hand, knows its exact value when he makes his choice. Suppose that b_2 is normally distributed in $\mathbb{R}^+$ and let $\mathcal{F}(\bar{b}_2, \sigma^2)$ be the cumulative distribution function, with the corresponding probability-density function represented by $f(b_2)$. Then, the timeline of the events is akin to that in Figure 6.



$t = 0$: $b_2 \sim \mathcal{F}(\bar{b}_2, \sigma^2)$
P makes her choice.

$t = 1$: b_2 is realised.
A observes b_2 and P's choice and makes his own.
The short-term outcome of A's choice is realised.

$t = 2$: The long-term outcome of A's choice is realised.

Figure 6. Timeline of events—b_2 uncertain at $t = 0$.

This new structure enhances the generality of our results. To see this, note that our framework accommodates cases where b_2 is ex ante definite as instances where $\sigma^2 = 0$. In addition, in many applications, it is intuitively plausible. Indeed, the principal can be fairly certain about the degree of gratification the agent can expect instantaneously upon making a decision. However, future consequences related to that decision are inherently compromised by environmental volatility-changes in exogenous factors that the principal may not even be aware of, let alone able to influence. In this sense, the agent has an informational advantage simply by being closer to these future consequences. The material payoff of the agent will thus now feature in the principal's utility in expected terms:

$$U^P = \begin{cases} \int_0^\infty (b_1 - \delta b_2) f(b_2) db_2 - \frac{C(n)}{\alpha\delta}, & \text{if } \gamma = B, \\ 0 - \frac{C(n)}{\alpha\delta}, & \text{if } \gamma = F. \end{cases} \tag{4}$$

The agent's utility, on the other hand, is still represented by Equation (2). Taking Equations (2) and (4) into account, the principal's problem can be stated as follows:

$$\max_n U^P = \pi^A - \frac{C(n)}{\alpha\delta} = \int_0^{\frac{b_1 - n}{\beta\delta}} (b_1 - \delta b_2) f(b_2) db_2 - \frac{C(n)}{\alpha\delta}. \tag{5}$$

Here, $\pi^A = \pi_1^A + \pi_2^A$ is the agent's total material payoff across periods 1 and 2. The particular functional form of the distribution of b_2 may imply more than one local maxima for Equation (5). To maintain simplicity, we impose two technical assumptions, which jointly ensure that the maximum is unique.

Assumption 1. *Given game $\mathcal{G}$, let $f(.)$ denote the density function according to which b_2 is distributed. Then, $f(.)$ is quasi-concave in b_2.*

Assumption 2. *In any game $\mathcal{G}$, $\frac{\beta^2}{\alpha} C'(0) < [(1-\beta)b_1] f(\frac{b_1}{\beta\delta})$.*

Assumption 1 implies that the marginal gain from n will not increase again once it has started decreasing. Given that $C(.)$ is increasing in n, a unique maximum point is implied. Assumption 2 precludes the possibility of a minimum occurring before a maximum as n increases. This would be possible if, for example, for n sufficiently small, the cost of increasing it surpassed its additional benefit. Assumptions 1 and 2 together ensure that P's problem attains a unique optimum solution, which confers the maximum return to n.

Assumptions 1 and 2 are rather restrictive, but their purpose is to maintain the analysis simple. Note that the set of values for b_2 that are relevant to P's problem is bounded by $\frac{b_1}{\delta}$ and $\frac{b_1}{\beta\delta}$. Thus, a solution would be attainable even with a different functional form for $f(.)$. The additional complication would be a comparison across all local maxima to determine the global one(s). Moreover, the same would be true even in the presence of local minima. We simply chose to sidestep these additional complexities, in order to refrain from further obscuring our analysis.

Bearing the above in mind, we can now proceed to characterise the solution to P's problem in the face of uncertainty. Proposition 3 presents this result.

Proposition 3. *Consider game $\mathcal{G}$ with $f(b_2)$ and $C(n)$ in line with Assumptions 1 and 2. Then, the optimal n satisfies:*

$$n^* = (1-\beta)b_1 - \frac{\beta^2}{\alpha f(\frac{b_1 - n^*}{\beta\delta})} C'(n^*). \tag{6}$$

The proof can be found in Appendix A. The result is, by construction, consistent with the analytical perspective of methodological individualism. That is, n will be assigned a positive value only if it is instrumental to the achievement of P's goal, and only to the extent that it has a higher rate of return compared to its cost. We thus see that the instrumental character of indoctrination does not change when uncertainty is introduced. The solution to P's problem is qualitatively similar to the one in our baseline version.

What about the agent's decision? In our baseline scenario, the value of n^* would be such that he would always be exactly indifferent between actions B and F, and would eventually choose F, in line with the principal's preference.[17] In this new scenario, however, it is possible that the agent's choice will not reflect the principal's preference, even given her investment in n. The reason is that the actual realisation of b_2 may be so low that he may find it profitable to choose action B even with the additional intrinsic utility n associated with action F. Figure 7 illustrates such a scenario.

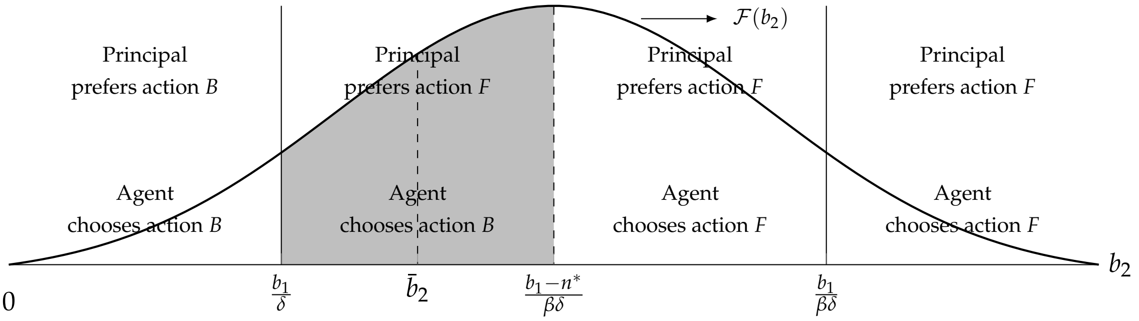

$\mathcal{F}(b_2)$
Principal prefers action B
Principal prefers action F
Principal prefers action F
Principal prefers action F
Agent chooses action B
Agent chooses action B
Agent chooses action F
Agent chooses action F
b_2
0
$\frac{b_1}{\delta}$
$\bar{b}_2$
$\frac{b_1-n^*}{\beta\delta}$
$\frac{b_1}{\beta\delta}$

Figure 7. *Misalignment of preferences:* Here, P has optimally assigned n^* on action F knowing that b_2 is drawn from $\mathcal{F}(b_2)$, but the realised value, $\bar{b}_2$, induces A to opt for action B. The shaded area is the cumulative probability of all such b_2 values.

Given that the possibility is now open for the agent's choice to be different from what the principal would want, we can also assess how the probability of this scenario varies with b_1 and the distribution of b_2. To do so, we need to formally distinguish between cases where the choice of A agrees with P's preference and cases where the two differ.

Definition 2. *Compliance: The degree of conformity following P's choice of $\hat{n}^*$ is the cumulative probability that A's choice will agree with P's preference given $\hat{n}^*$.*

Using Definitions 1 and 2, we now turn to examine how n^* and compliance are affected by changes in b_1 and $\mathcal{F}(b_2)$.

Corollary 3. *Consider game $\mathcal{G}$ satisfying Assumptions 1 and 2. An increase in the value of b_1, from $\bar{b}_1$ to $\hat{b}_1$ may lead to a higher n^*, so long as Assumption 2 remains satisfied. However, compliance may be lower as a result of the increase in b_1.*

Proof. See Appendix B. □

Corollary 3 points out that changes in b_1 may affect n^* and compliance in different ways. Suppose that b_1 increases. This implies that both players will be more inclined to opt for action B than before. However, the discrepancy between their preferences widens. To see this, notice that the agent's switching threshold changes by a greater margin than the principal's one does. Therefore, the range of b_2 values for which their preferences are conflicting is now larger. As a result, if the principal still prefers action F, then the previous level of n^* is no longer optimal. In particular, the increase in b_1 induces her to increase n, in order to account for the additional appeal of action B relative to action F.

[17] The same would be true in expected terms, if the cost of action B was uncertain for both players. So long as P and A had the same distribution of b_2 in mind, A's choice would be anticipated by P: They would both form the same expectation about b_2. Thus, even if the *actual* value of b_2 eventually proved to be different than what they had expected, their choices would coincide.

It is important to bear in mind that, in adjusting n^* to account for the change, the principal is interested in its *marginal* benefit, not what she gets out of it *on average*. It may well be the case that on average the agent will choose action B, contrary to the principal's preference. However, it may still make sense for her to invest in n, so long as what she gets from doing so (in expected terms) is more than what she spends on it. Figure 8 illustrates this situation, given a linear cost function and a normal distribution for b_2. In this scenario, an increase in b_1 results in a higher n^* and a lower degree of compliance.

The positive relation between b_1 and n^* also implies that a decrease in b_1 will likely be followed by a reduction in the level of n^*. Intuitively, the change renders option B less appealing and, therefore, encourages the principal to reduce the level of indoctrination, so as to lower its cost. We thus observe a trade-off between the exogenous incentive to opt for the option that the principal favours and the endogenous intrinsic preference she instils herself.

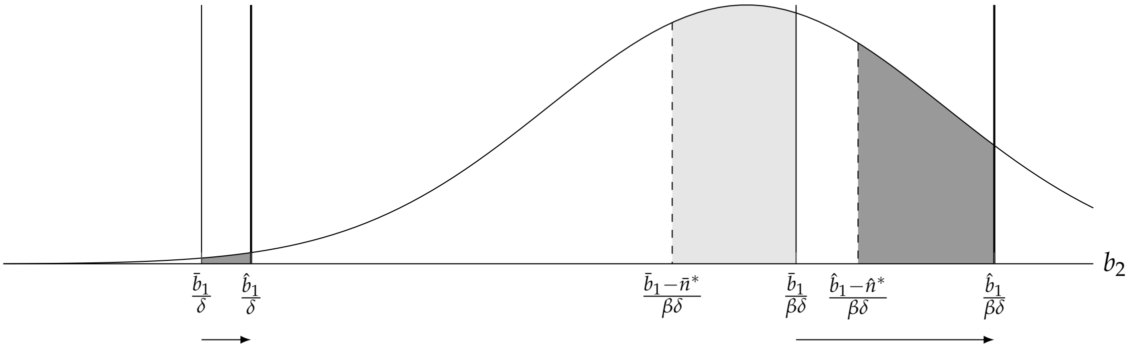

b_2
$\frac{\bar{b}_1}{\delta}$
$\frac{\hat{b}_1}{\delta}$
$\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}$
$\frac{\bar{b}_1}{\beta\delta}$
$\frac{\hat{b}_1-\hat{n}^*}{\beta\delta}$
$\frac{\hat{b}_1}{\beta\delta}$

Figure 8. $\hat{b}_1 > \bar{b}_1$: The immediate consequence from option B is relatively larger and so is the level of n^*. The proportion of b_2 values for which A's choice will conform with P's preference is now lower.

Corollary 4. *Consider game $\mathcal{G}$ satisfying Assumptions 1 and 2. A parallel rightward shift of $\mathcal{F}(b_2)$, which increases $E[b_2]$ from $\bar{b}_2$ to $\hat{b}_2$, where $\frac{\bar{b}_1}{\delta} < \bar{b}_2 < \hat{b}_2$, may induce P to invest less in n. However, such a shift will always result in greater compliance.*

Proof. See Appendix B. □

An increase in the magnitude of the expected future consequence can lead to a lower level of n^*. The intuition behind this result is straightforward. As the increase in $E[b_2]$ renders option B less attractive, the principal will eventually be discouraged from investing in n. The reason is that its instrumentality dwindles. As the agent becomes more likely to avoid B anyway, investing in n and assuming the cost of doing so gets progressively counter-productive.

Thus, the increase in $E[b_2]$ may be partially crowded out by the decrease in the incentive to instil a given level of n. The same trade-off ensues between the agent's extrinsic and intrinsic incentives to act in a particular way. In the face of higher exogenous motivation, his intrinsic preference dwindles, because it is no longer relevant.

It is worth noting that this is also true when the magnitude of the expected future consequence goes towards the opposite direction. The reasoning is the same as before. A reduction in $E[b_2]$ may induce the principal to compensate by increasing n. However, successive reductions will eventually discourage her from increasing n, as the preference discrepancy becomes progressively less relevant.

In line with the previous arguments, the agent's degree of compliance with the principal's preference depends on the initial distribution of b_2. If $E[b_2] > \frac{\bar{b}_1}{\delta}$ in the first place, then any subsequent increase will lead to higher compliance. Figure 9 presents a situation where a higher $E[b_2]$ results in both a lower n^* and a higher degree of compliance.

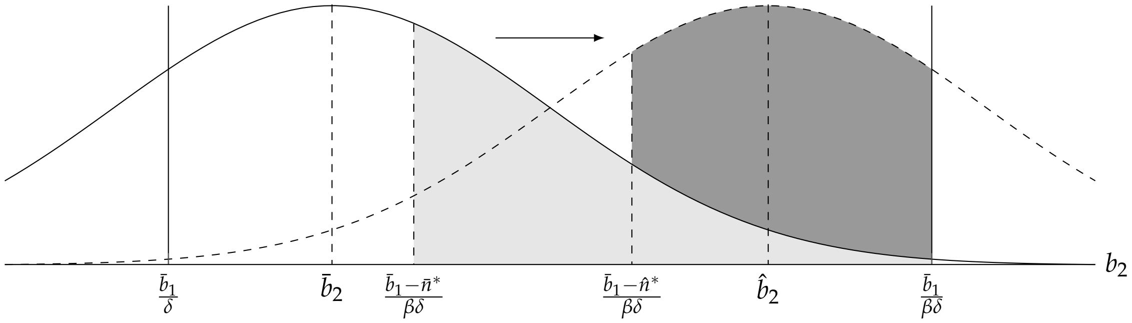


Figure 9. $\hat{b}_2 > \bar{b}_2$: The expected future consequence is larger, the level of n^* is lower, and the probability of compliance is higher.

Notice that the crowding-out of the intrinsic preference by the material cost is always accompanied by enhanced compliance. To see why, consider a situation where the expected cost of action B is such that the principal should optimally assign $n^* > 0$ to action F. If $E[b_2]$ increases, then the principal will only settle for a lower level of n if it confers a greater return than the previous one. Investing in n is not more expensive than it was before. If anything, she could still invest in it to the extent she did before. If she chooses to undercut her investment, it is because this is the optimal response.

Notice that Corollary 4 describes a variance-preserving shift. That is, it refers to a change in the distribution of b_2 to a higher expected value, but with the same *degree of uncertainty*. This is important for our analysis, as our conclusion that the increase in $E[b_2]$ always results in an increased degree of compliance does not necessarily hold if we allow for simultaneous changes in its variance. To see this, consider a situation where an exogenous shift affects both $\bar{b}_2$ and σ^2. Since n^* is affected by both, the effects of this change may actually counteract each other. We explore this possibility in the following Proposition.

Proposition 4. *Consider game $\mathcal{G}$ satisfying Assumptions 1 and 2. Suppose that an exogenous shock changes the distribution of b_2 to one that has a higher mean and a higher variance. Such a shock may induce P to invest less in n and may also lead to a lower degree of compliance.*

Proof. See Appendix C for a proof by example. □

Proposition 4 highlights the potential conflict between two effects that result from the distributional change. One of these effects comes as a result of the higher expected future consequence. The other follows from the increased uncertainty about that consequence. The net effect on n^* and the degree of compliance can be surprising.

As it has already been argued (see Corollary 4), an increase in $E[b_2]$ may reduce the principal's incentive to invest in n. However, the principal's incentive is crowded out due to the fact that, given the new $E[b_2]$, even a lower n makes the agent more likely to comply (by choosing action F). Thus, the increase in $E[b_2]$ (given b_1) leads to a higher degree of compliance.

The increase in σ^2, on the other hand, may induce n^* to fall even further. This is because, as the future consequence becomes more volatile, the marginal return that the principal receives by increasing n gets progressively lower. Intuitively, the increased uncertainty implies that the most likely b_2 values are now less probable. As a result, it becomes difficult for the principal to pinpoint a level for n that is highly likely to be optimal.[18] Given that the cost of providing n has not changed, the principal may find it better to reduce n in response to higher uncertainty.

The final outcome may, thus, resemble the one illustrated in Figure 10. This is a case where a shift towards a higher but more volatile $E[b_2]$ results in a reduced probability of the agent choosing in

[18] Recall that the optimal level of n would render the agent exactly indifferent between options B and F.

accordance with the principal's preference. Consequently, apart from crowding out n, the change also renders the agent more vulnerable to present bias. This result is all the more striking when considered in light of the intuition that a higher $E[b_2]$ on its own would have the exact opposite effect.

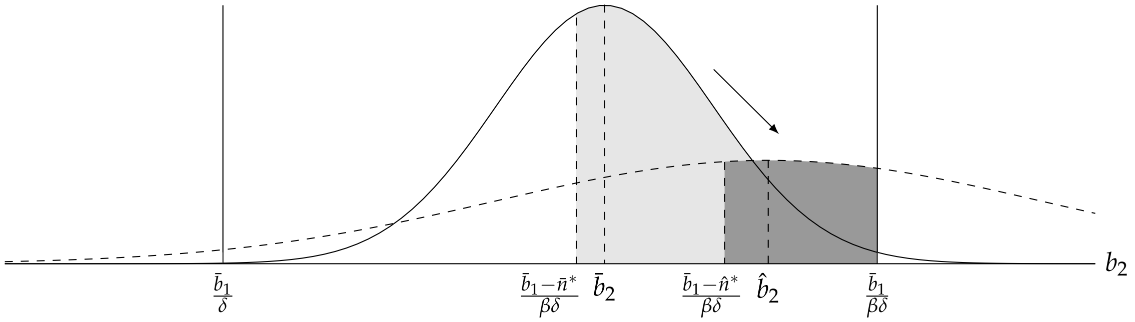


Figure 10. $\hat{b}_2 > \bar{b}_2, \hat{\sigma}^2 > \bar{\sigma}^2$: The expected future consequence is larger and more uncertain. The level of n^* and the degree of compliance are both lower.

We have thus far examined the changes in n^* and the degree of compliance induced by changes in b_1 and the distribution of b_2. As a final remark, we note that n^* varies monotonically with $C^{'}(.)$, to which it is inversely related. That is, other things being equal, an increase in the marginal cost of instilling n always leads to a lower level for the intrinsic preference and vice versa. In particular, there is no crowding-out related to the principal's incentives: a reduction in $C^{'}(n)$ will render her unambiguously more willing to provide a higher n^*. The same is true with respect to compliance.

3. Discussion

This section explores some of the potential applications of our framework. It discusses firstly some implications for public policy, where the government can be viewed as the principal and a representative citizen as the agent. It then examines cultural transmission under a parent–child interpretation. It concludes with a discussion of self-control problems as viewed from the perspective of the intertemporal self. Note that the examples reported below are not exhaustive. The domain of application of our theory is much more general and includes all instances where one-shot decisions can have consequences at multiple points in time.[19]

3.1. Public Policy

We turn to some consequences of our analysis for the design of public policy. Our aim is to demonstrate that, owing to the strategic interplay analysed in Section 2, the results of policy measures may be very different from those originally expected. To do so, we use examples of policies that may prove inefficient, given the policymaker's stated goals.

Consider, thus, a policy aimed at encouraging more people to save some of their income, e.g., an increase in the interest rate, taking effect at $t = 1$. Such a policy will have an effect on the amount of period-2 consumption one has to forfeit in order to spend more money in period 1. In the context of our model, it amounts to an exogenous increase in $E[b_2]$. Should we expect that this policy will be successful, and to what extent? One factor that may limit the policy's effectiveness is the change in the culture of parsimony that its announcement initiates. As Corollary 4 points out, a greater $E[b_2]$ may crowd out private and public investment in promoting frugality. Thus, even in the absence

[19] In fact, our paper is relevant to an even wider class of studies. Consider, for example, the model proposed by [29], where practices of tax evasion, as soon as they have been established, are perpetuated as history-dependent cultural norms. Our paper can be seen as complementary to his, due to the fact that we focus on the mechanism by which cultural indoctrination takes place across generations.

of additional effects stemming from the announcement of the policy, the resulting increase in the proportion of savers may not be as high as initially expected.

Suppose, now, that the government aims to discourage tax avoidance while in the midst of an austerity programme. To do so, it imposes stronger sanctions to perpetrators. However, owing to the need for austerity, it is also required to cut back on audits. What does the resulting situation look like? The announcement of stricter penalties (higher $E[b_2]$) is set to increase compliance, although it is also expected to discourage a culture of duty to pay one's taxes (lower n). The reduction in oversight, however, results in these penalties being more unlikely than before. As a result, it mitigates both the sense of social responsibility and compliance. Proposition 4 suggests that the resulting effect on taxes may well be negative.

Lastly, consider a policy that aims to reduce carbon emissions. One way of doing so would be to collect research on the adverse consequences for the environment and, thus, the society's future prospects. Then, this research would be disseminated, perhaps in the form of short advertisements, in a bid to increase environmental awareness among the population. Our analysis shows that there may be a caveat in this reasoning. Specifically, if the research appears inconclusive, so that many possible future scenarios seem likely but none is deemed particularly probable, the policy may backfire. Furthermore, as Proposition 4 points out, this can be true even if the additional information results in the situation appearing more dire on average. Thus, our framework suggests that caution must be exercised in the release of information as part of a policy measure.

The three examples outlined above highlight the trade-off between people's (exogenous) material incentives and their (endogenous) intrinsic motivation. By providing extrinsic incentives, public policies may end up crowding out intrinsic motivation. In doing so, they are compromising, at least partly, their own effects (this result is consistent with [30], among others). Our analysis indicates that caution needs to be exercised when assessing the potential effects of a proposed policy measure.

3.2. Cultural Transmission

The transmission of cultural values across generations has traditionally been modelled by means of dynamic games played by successive agents. In constructing such a process, Ref. [10] applies the imperfect-empathy setup of [11]. The result is a model of parent–child interaction. The assumptions made are that (a) parents can affect the deep preferences of their children and (b) parents try to maximise a notion of utility of their children that departs from pure material welfare. This general framework of parent–child interaction (with alternatives to imperfect empathy) is becoming increasingly popular as a means of explaining social dynamics and cultural change.[20]

In this interpretation, the principal in our model can assume the role of the parent, while the agent that of the child. It is then straightforward to see how parental indoctrination can be modelled as a principal–agent game, where the former takes an interest in the utility of the latter. Notice that, in the parent–child interpretation, it is not necessary that the players suffer from present bias, only that the agent's discount factor is different than that of the principal.

3.3. Self Control

We can also evaluate the scope of our framework through the perspective of the intertemporal self (in the spirit of [46]). To this end, consider a game $\mathcal{G}$ that is being played among the various instances of the same person, acting at different points in time. It is natural to assume that in this case $\alpha = \beta$. Then, our analysis focuses on the action of her self at $t = 0$ and the choice of her self at $t = 1$. Suppose that this person is initially characterised solely by preferences over outcomes and that she also exhibits present bias. Suppose, further, in line with the previous setup, that, while she knows

[20] See, for example, [3,31–41], among others, consider the deployment of strategic bequests by altruistic parents. [4,42–45] provide a variety of views on the usefulness of the general approach.

about her bias, she cannot eliminate it per se. Then, in trying to maximise her intertemporal utility, she might optimally set $n \in [0, |(1-\beta)b_1|)$.

How can such a result be interpreted? From the point of view of the self, at $t = 0$, it is (weakly) optimal to commit to preferring an action over another. She knows that, if she is equipped only with materialistic preferences, then it is probable that, in the face of temptation, she will make an ill-preferred choice. To reduce this probability, she may want to commit to a particular code of conduct, so as to enhance the appeal of the alternative option.

An appealing feature of this account of preference formation is its general applicability. Note that the aforementioned code of conduct can be grounded on various premises, such as moral principles (see [47,48]), social norms (see [49,50]), reputation (see [51–54]), and habitual or conventional decision-making (see [55–57], as well as [58]).[21] All such concerns can be shown to be instrumental from a purely materialist viewpoint. Thus, such preferences can also emerge through an evolutionary process, assuming that present bias is also at play (through a reasoning similar to [22]).

4. Conclusions

We propose a game-theoretic model of endogenous preference formation, where an action-based inclination is optimally chosen to counterbalance present-biased proclivities. We build on the idea that preferences are, to some extent, malleable. We then investigate the relationship between material incentives and intrinsic motivation. Our analysis indicates that the relationship between the two is non-monotonic. Our results are especially relevant to the domain of policy analysis.

The theory presented here describes how the endowment of an intrinsic preference can be optimal from a materialistically rational perspective. We depict the dialectics between parameter variations and individual incentives and show how the effects of the former can sometimes crowd out the latter. These effects are important, both with respect to cultural transmission and to the exercise of self-control. The effectiveness of a policy is demonstrated to depend, at least to some extent, on it providing the right mix of incentives to individuals.

The paper does not consider the intergenerational dynamics that ensue in such a context. This is a fascinating research question in its own right. Here, instead, we propose two main arguments: that preferences for actions can be rationally assigned and that they should be taken into account when considering the effects of changes in the underlying economic environment.

Funding: This research received no external funding

Acknowledgments: The author is grateful to Silvia Sonderegger and Daniele Nosenzo for their unwavering support and invaluable guidance, as well as to two anonymous referees for their thoughtful comments. The financial support provided by the Economic and Social Research Council (ESRC) is gratefully acknowledged.

Conflicts of Interest: The author declares no conflicts of interest.

[21] See [58–63] for critical appraisals of various accounts.

Appendix A. *P*'s Problem

Consider the objective function of P as defined by Expression (5). Upon satisfaction of the first-order condition:

$$\begin{aligned} F.O.C.: -\frac{1}{\beta\delta}\left(b_1-\delta\frac{b_1-n}{\beta\delta}\right)f\left(\frac{b_1-n}{\beta\delta}\right)-\frac{C'(n)}{\alpha\delta}=0\Rightarrow \\ \Rightarrow \left[(1-\beta)b_1-n\right]=\frac{\beta^2}{\alpha f(\frac{b_1-n}{\beta\delta})}C'(n)\Rightarrow \\ \Rightarrow n^*=(1-\beta)b_1-\frac{\beta^2}{\alpha f(\frac{b_1-n^*}{\beta\delta})}C'(n^*). \end{aligned} \tag{A1}$$

The second-order condition for a strict maximum suggests that:

$$-\frac{1}{\beta\delta}[(1-\beta)b_1-n^*]f'\left(\frac{b_1-n^*}{\beta\delta}\right)-f\left(\frac{b_1-n^*}{\beta\delta}\right)-\frac{\beta^2}{\alpha}C''(n^*)<0. \tag{A2}$$

Notice that inequality (A2) is not necessarily satisfied for every n^* that satisfies (A1). Assumption 2 ensures that the n^* that satisfies (A1) maximises P's utility function. Assumption 1 guarantees that this maximum point is unique.

Appendix B. Parameter Variations

We now investigate how variations in the parameters of the model affect the level of n and the rate of A's adherence to P's preference. We focus firstly on b_1. Recall that according to Equation (A1):

$$\left[(1-\beta)b_1-n\right]f\left(\frac{b_1-n}{\beta\delta}\right)-\frac{\beta^2}{\alpha}C'(n)=0. \tag{A3}$$

Deriving (A3) with respect to b_1, we find:

$$\frac{1}{\beta\delta}[(1-\beta)b_1-n]f'\left(\frac{b_1-n}{\beta\delta}\right)+(1-\beta)f\left(\frac{b_1-n}{\beta\delta}\right). \tag{A4}$$

Deriving (A3) with respect to n, we find:

$$-\frac{1}{\beta\delta}[(1-\beta)b_1-n]f'\left(\frac{b_1-n}{\beta\delta}\right)-f\left(\frac{b_1-n}{\beta\delta}\right)-\frac{\beta^2}{\alpha}C''(n). \tag{A5}$$

Consider an exogenous shift from $\bar{b}_1$ to $\hat{b}_1$, where $|\bar{b}_1|<|\hat{b}_1|$. Let $db_1\equiv|\hat{b}_1|-|\bar{b}_1|$ and $dn^*\equiv\hat{n}^*-\bar{n}^*$. Note that P will respond to the change in b_1 by adjusting n according to (A3). Therefore, (A4) and (A5) together add up to:

$$\begin{aligned} \left[\frac{1}{\beta\delta}[(1-\beta)b_1-n]f'\left(\frac{b_1-n}{\beta\delta}\right)+(1-\beta)f\left(\frac{b_1-n}{\beta\delta}\right)\right]db_1- \\ -\left[\frac{1}{\beta\delta}[(1-\beta)b_1-n]f'\left(\frac{b_1-n}{\beta\delta}\right)+f\left(\frac{b_1-n}{\beta\delta}\right)+\frac{\beta^2}{\alpha}C''(n)\right]dn=0. \end{aligned}$$

Thus, it is true that:

$$\frac{dn^*}{db_1}=\frac{\frac{1}{\beta\delta}[(1-\beta)b_1-n^*]f'\left(\frac{b_1-n^*}{\beta\delta}\right)+(1-\beta)f\left(\frac{b_1-n^*}{\beta\delta}\right)}{\frac{1}{\beta\delta}[(1-\beta)b_1-n^*]f'\left(\frac{b_1-n^*}{\beta\delta}\right)+f\left(\frac{b_1-n^*}{\beta\delta}\right)+(\beta^2/\alpha)C''(n^*)}. \tag{A6}$$

The sign of $\frac{dn^*}{db_1}$ depends on the sign and magnitude of $f'\left(\frac{b_1-n}{\beta\delta}\right)$. To see this, recall firstly that inequality (A2) implies that $f'\left(\frac{b_1-n^*}{\beta\delta}\right)$ has a lower bound:

$$f'\left(\frac{b_1-n^*}{\beta\delta}\right) > -\frac{\beta\delta f\left(\frac{b_1-n^*}{\beta\delta}\right) + (\beta^3\delta/\alpha)C''(n^*)}{(1-\beta)b_1 - n^*}.$$

This means that the denominator of the fraction on the right-hand side of Equation (A6) is positive for every n^* that constitutes a maximum (assuming that $\beta < 1$). The numerator, on the other hand, will be negative if:

$$f'\left(\frac{b_1-n^*}{\beta\delta}\right) < -\frac{\beta\delta(1-\beta) f\left(\frac{b_1-n^*}{\beta\delta}\right)}{(1-\beta)b_1 - n^*}.$$

Taking the above into account, we can discern the following cases:

$$\frac{dn^*}{db_1} = \begin{cases} y > 0, & if \, \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} > -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}, \\ y < 0, & if \, \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} < -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}, \\ y = 0, & if \, \frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} = -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}. \end{cases} \tag{A7}$$

It is worth noting that, when $\frac{dn^*}{db_1} < 0$, n^* will fall to zero following an increase in b_1. The reason is that, from Assumption 2, it can be seen that $f\left(\frac{b_1-n^*}{\beta\delta}\right)$ is decreasing more rapidly than $C(n)$. Thus, as is the case in our baseline scenario, in response to an increase in b_1 P will either increase n^*, or eliminate it altogether. We can describe the relationship between changes in b_1 and changes in n^* in a general proposition. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Suppose that $\bar{b}_1$ is replaced with $\hat{b}_1$, where $|\hat{b}_1| > |\bar{b}_1|$. Such a change will, ceteris paribus, lead to:

- $\hat{n}^* > \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} > -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$,
- $\hat{n}^* < \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} < -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$,
- $\hat{n}^* = \bar{n}^*$, if $\frac{f'\left(\frac{b_1-n^*}{\beta\delta}\right)}{f\left(\frac{b_1-n^*}{\beta\delta}\right)} = -\frac{\beta\delta(1-\beta)}{(1-\beta)b_1-n^*}$.

The first of these cases corresponds to Corollary 3. It suggests that, so long as the percentage change in the frequency of the cut-off point is above a certain threshold, P will have an incentive to increase n^* in response to increases in b_1.

Changes in b_1 also have a bearing on compliance, which, according to Corollary 3, may be negative. Following our definition of compliance (2), we can measure its variations as changes in the cumulative probability that A's choice wil *not* conform with P's preference. As this probability dwindles, the degree of compliance increases.

Let NC be the cumulative probability that the choice of A will be different from P's preference. Then, $NC = \mathcal{C}^{\mathcal{F}}\left(\frac{b_1-n^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{b_1}{\delta}\right)$, where $\mathcal{C}^{\mathcal{F}}(.)$ is the cumulative distribution function of distribution $\mathcal{F}(.)$. Consider, then, the change in this difference in response to a change in b_1:

$$\frac{\partial\left(\mathcal{C}^{\mathcal{F}}\left(\frac{b_1-n^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{b_1}{\delta}\right)\right)}{\partial b_1} = \left(1 - \frac{\partial n^*}{\partial b_1}\right)\frac{1}{\beta\delta} f\left(\frac{b_1-n^*}{\beta\delta}\right) - \frac{1}{\delta} f\left(\frac{b_1}{\delta}\right). \tag{A8}$$

Given that $\frac{\partial n^*}{\partial b_1} < 1$ (from Equation (A6)), the first term of the right-hand side of (A8) is always positive. Therefore, for a sufficiently low $f\big(\frac{b_1}{\delta}\big)$, an increase in b_1 will lead to a lower degree of compliance.

To clarify this argument further, we also provide a numerical example. Consider game $\mathcal{G}$ with $\bar{b}_1 = 4$, $\hat{b}_1 = 6$, $C(n) = 4n$, $\delta = 1$, $\alpha = 1$, $\beta = 0.25$, and $\mathcal{F}(b_2, \sigma^2) = \mathcal{N}(14, 2)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ, variance σ^2, and probability density function $f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$. Let $\bar{n}^*$ be the equilibrium level of n corresponding to $\bar{b}_1$ and $\hat{n}^*$ the one corresponding to $\hat{b}_1$. Then, from Equation (A1), solving for $\bar{n}^*$:

$$\begin{aligned}\bar{n}^* &= (1-\beta)\bar{b}_1 - \frac{\beta^2}{\alpha f\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big)} C^{'}(\bar{n}^*) = \\ &= 3 - \frac{0.0625}{f\big(\frac{4-\bar{n}^*}{0.25}\big)} 4.\end{aligned}$$

Solving out, we obtain $\bar{n}^*$ approximately equal to 1. On the other hand, solving for $\hat{n}^*$:

$$\begin{aligned}\hat{n}^* &= (1-\beta)\hat{b}_1 - \frac{\beta^2}{\alpha f\big(\frac{\hat{b}_1 - \hat{n}^*}{\beta\delta}\big)} C^{'}(\hat{n}^*) = \\ &= 4.5 - \frac{0.0625}{f\big(\frac{4-\hat{n}^*}{0.25}\big)} 4.\end{aligned}$$

Again, solving out, we obtain $\hat{n}^*$ approximately equal to 2.875. We see that P has increased n^* in response to the rise in b_1. With respect to compliance, it is easy to see that the cumulative probability of disagreement between the two players has increased. In particular, under $\bar{b}_1$, this probability is equal to:

$$\mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{\bar{b}_1}{\delta}\right) = \mathcal{C}^{\mathcal{F}}(12) - \mathcal{C}^{\mathcal{F}}(4) \approx 0.159. \tag{A9}$$

Under $\hat{b}_2$, on the other hand, it becomes:

$$\mathcal{C}^{\mathcal{F}}\left(\frac{\hat{b}_1 - \hat{n}^*}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{\hat{b}_1}{\delta}\right) = \mathcal{C}^{\mathcal{F}}(12.5) - \mathcal{C}^{\mathcal{F}}(6) \approx 0.227. \tag{A10}$$

Thus, compliance decreases following the increase of b_1 from $\bar{b}_1$ to $\hat{b}_1$.

We now turn to variations in b_2 and their effect on n^*. In what follows, $b_1 = \bar{b_1}$. In accordance with Assumptions 1 and 2, let $b_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, where $\mathcal{F}(.)$ is quasi-concave. To start with, suppose that a variance-preserving shift occurs, from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{H}(\hat{b}_2, \bar{\sigma}^2)$, where $\hat{b}_2 > \bar{b}_2 > 0$. In other words, the distribution shifts towards higher values of b_2, making option B less appealing than before. Let $\bar{n}^*$ denote the equilibrium n under $\mathcal{F}(.)$ and $\hat{n}^*$ that under $\mathcal{H}(.)$. In addition, let $f(.)$ denote the probability density function of distribution $\mathcal{F}(.)$ and $h(.)$ that of distribution $\mathcal{H}(.)$. It is straightforward to verify from Equation (A3) that an increase (decrease) of the density of the cut-off point that has resulted from a change in the distribution will lead to an increase (decrease) in n^*. The reason is that such a change adjusts the importance of $C^{'}(n^*)$ in the determination of n^*. In other words, it is true that if $h\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big) > f\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big)$, then $\hat{n}^* > \bar{n}^*$, while if $h\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big) < f\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big)$, then $\hat{n}^* < \bar{n}^*$. In addition, from (A3), the following two equations are true:

$$(1-\beta)\bar{b}_1 - \bar{n}^* - \frac{\beta^2}{\alpha f\big(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\big)} C^{'}(\bar{n}^*) = 0,$$

$$(1-\beta)\bar{b}_1 - \hat{n}^* - \frac{\beta^2}{\alpha h\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)} C^{'}(\hat{n}^*) = 0.$$

Therefore, it follows that:

$$\bar{n}^* - \hat{n}^* = \frac{\beta^2}{\alpha h\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)} C^{'}(\hat{n}^*) - \frac{\beta^2}{\alpha f\left(\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}\right)} C^{'}(\bar{n}^*). \tag{A11}$$

Then, comparing $\bar{n}^*$ with $\hat{n}^*$, one can see that:

$$\bar{n}^* > \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) > h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right) \frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)}.$$

The converse is also true:

$$\bar{n}^* < \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) < h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right) \frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)}.$$

Suppose, now, that $\frac{d^2C(n)}{dn^2} \geq 0$, that is, that the cost function is weakly convex in n. In this case, $\bar{n}^* > \hat{n}^* \Rightarrow \frac{C^{'}(\bar{n}^*)}{C^{'}(\hat{n}^*)} > 1$ and vice versa. Thus:

$$\bar{n}^* > \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) > h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right),$$

$$\bar{n}^* < \hat{n}^* \Rightarrow f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) < h\left(\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta}\right).$$

It thus becomes apparent that Equation (A11) implies an upper and a lower bound for the density of the new cut-off point, $h\left(\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}\right)$. That is, the density of a cut-off point that has resulted from an increase in n can never be lower than that of the initial cut-off point. Conversely, the density of a cut-off point that has resulted from a reduction in n will never surpass that of the initial cut-off point. This result implies that a parallel (variance-preserving) shift in the distribution of b_2, such as the one described above, always enhances compliance.

To see why this is the case, consider such a change, whereby rule $f : \mathbb{R}^+ \to \mathbb{R}^+$ is replaced by $h : \mathbb{R}^+ \to \mathbb{R}^+$ such, that $h(b_2) = f(b_2 - \Delta)\ \forall b_2$, where $\Delta > 0$. Recall that $\bar{n}^*$ is the equilibrium level of n under $f(.)$ and $\hat{n}^*$ that under $h(.)$. Then, it is true that:

$$\frac{\bar{b}_1 - \hat{n}^*}{\beta\delta} \leq \frac{\bar{b}_1 - \bar{n}^*}{\beta\delta} + \Delta. \tag{A12}$$

To see this, one can start from $\hat{n} = \bar{n}^* - \beta\delta\Delta$ and show that this is, in fact, not equal to $\hat{n}^*$.[22] Recall that, if $\hat{n}$ was an equilibrium level under $h(.)$, then it would need to satisfy (A3). However:

[22] If it were, the cut-off point $\frac{\bar{b}_1-\hat{n}^*}{\beta\delta}$ would be in the same relative position given $h(.)$ with that of $\frac{\bar{b}_1-\bar{n}^*}{\beta\delta}$ under $f(.)$.

$$
\begin{aligned}
\big[(1-\beta)\bar{b}_1 - \hat{n}\big] h\left(\frac{\bar{b}_1 - \hat{n}}{\beta\delta}\right) - \frac{\beta^2}{\alpha} C^{'}(\hat{n}) = \\
&= \big[(1-\beta)\bar{b}_1 - \bar{n}^* + \beta\delta\Delta\big] h\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta} + \Delta\right) - \frac{\beta^2}{\alpha} C^{'}(\bar{n}^* - \beta\delta\Delta) = \\
&= \big[(1-\beta)\bar{b}_1 - \bar{n}^* + \beta\delta\Delta\big] f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) - \frac{\beta^2}{\alpha} C^{'}(\bar{n}^* - \beta\delta\Delta) = \\
&= \frac{\beta^2}{\alpha}\big[C^{'}(\bar{n}^*) - C^{'}(\bar{n}^* - \beta\delta\Delta)\big] + \beta\delta\Delta f\left(\frac{\bar{b}_1 - \bar{n}^*}{\beta\delta}\right) > 0.
\end{aligned}
$$

It is obvious that (A3) is not satisfied by $\hat{n} = \bar{n}^* - \beta\delta\Delta$. Therefore, P has an incentive to further increase n, thereby increasing the probability that A will make her preferred choice in the next period.

We can organise our findings with respect to variance-preserving distributional shifts in another general proposition. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Consider a shift from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{H}(\hat{b}_2, \bar{\sigma}^2)$, where $0 < \bar{b}_2 < \hat{b}_2$. Let $f(.)$ denote the probabilty density function of distribution $\mathcal{F}(.)$ and $h(.)$ denote the probability density function of distribution $\mathcal{H}(.)$. Then, such a change will, ceteris paribus, lead to:

- $\hat{n}^* < \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) < f(\frac{b_1 - \bar{n}^*}{\beta\delta})$,
- $\hat{n}^* > \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) > f(\frac{b_1 - \bar{n}^*}{\beta\delta})$,
- $\hat{n}^* = \bar{n}^*$, if $h(\frac{b_1 - \bar{n}^*}{\beta\delta}) = f(\frac{b_1 - \bar{n}^*}{\beta\delta})$.

If, additionally, $\frac{b_1}{\delta} < \bar{b}_2$, then, ceteris paribus, the probability that A's choice will comply with P's preference increases as $E[b_2]$ grows larger (as in Figure A1).

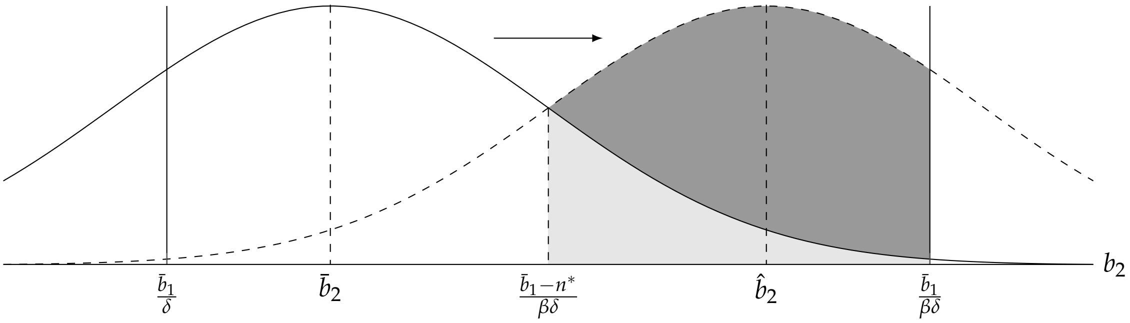


Figure A1. $\hat{b}_2 > \bar{b}_2$: The expected future consequence is relatively larger, but the level of n^* is the same.

Appendix C. Example of a Distributional Shift

What if both the mean and the variance of b_2 increase as a result of the distributional shift? This is the case pertaining to Proposition 4, which states that such a change may decrease both n^* and compliance. We show how this can be the case through an example situation. Consider game $\mathcal{G}$ with $b_1 = 4$, $C(n) = 2n$, $\delta = 1$, $\alpha = 1$, $\beta = 0.5$, $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2) = \mathcal{N}(5.5, 0.4)$, and $\mathcal{J}(\hat{b}_2, \hat{\sigma}^2) = \mathcal{N}(7, 1)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with mean μ, variance σ^2, and probability density function $g(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$. Let $\bar{n}^*$ be the equilibrium level of n under distribution $\mathcal{F}(.)$ and $\hat{n}^*$ that is under $\mathcal{J}(.)$. Then, from Equation (A1), solving for $\bar{n}^*$:

$$
\begin{aligned}
\bar{n}^* &= (1-\beta)b_1 - \frac{\beta^2}{\alpha f(\frac{b_1 - \bar{n}^*}{\beta\delta})} C^{'}(\bar{n}^*) = \\
&= 2 - \frac{0.25}{f(\frac{4 - \bar{n}^*}{0.5})} 2.
\end{aligned}
$$

Solving out, we obtain $\bar{n}^*$ approximately equal to 1.38. On the other hand, solving for $\hat{n}^*$:

$$\hat{n}^* = (1-\beta)b_1 - \frac{\beta^2}{\alpha j\left(\frac{b_1-\hat{n}^*}{\beta\delta}\right)}C^{'}(\hat{n}^*) =$$

$$= 2 - \frac{0.25}{j\left(\frac{4-\hat{n}^*}{0.5}\right)}2.$$

Again, solving out, we obtain $\bar{n}^*$ approximately equal to 0.67. Thus, the level of n^* has decreased as a result of the distributional shift. Regarding compliance, let $\mathcal{C}^{\mathcal{N}}(.)$ denote the cumulative distribution function of $\mathcal{N}(.)$. Then, under $\mathcal{F}(5.5, 0.4)$, the share of b_2 values for which A would conform with P's preference in the case of conflict was:

$$\mathcal{C}^{\mathcal{F}}\left(\frac{b_1}{\beta\delta}\right) - \mathcal{C}^{\mathcal{F}}\left(\frac{b_1-\bar{n}^*}{\beta\delta}\right) \approx 1 - 0.258 = 0.742.$$

Under $\mathcal{J}(7,1)$, the share of b_2 values for which A will conform with P's preference in the case of conflict becomes:

$$\mathcal{C}^{\mathcal{J}}\left(\frac{b_1}{\beta\delta}\right) - \mathcal{C}^{\mathcal{J}}\left(\frac{b_1-\hat{n}^*}{\beta\delta}\right) \approx 0.841 - 0.345 = 0.496.$$

Thus, both n^* and compliance decrease as a result of the distributional shift. The results are illustrated in Figure 10, in Section 2.3. In general terms, the proposition may be stated as follows. Consider game $\mathcal{G}$ with $\bar{b}_1$, $\bar{b}_2 \sim \mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$, and $\bar{n}^*$. Consider a shift from $\mathcal{F}(\bar{b}_2, \bar{\sigma}^2)$ to $\mathcal{J}(\hat{b}_2, \hat{\sigma}^2)$, where $\frac{1}{\delta} < \bar{b}_2 < \hat{b_2}$ and $\bar{\sigma}^2 < \hat{\sigma}^2$. Let $f(.)$ denote the probability density function of distribution $\mathcal{F}(.)$ and $j(.)$ denote the probability density function of distribution $\mathcal{J}(.)$. Then, $\exists f, j : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\hat{n}^* < \bar{n}^*$ and the degree of compliance is lower.

References

1. Hobbes, T. *De Cive (The Citizen)*; Appleton-Century-Crofts: New York, NY, USA, 1949.
2. Stigler, G.J.; Becker, G.S. De gustibus non est disputandum. *Am. Econ. Rev.* **1977**, *67*, 76–90. [CrossRef]
3. Adriani, F.; Sonderegger, S. Why do parents socialize their children to behave pro-socially? An information-based theory. *J. Public Econ.* **2009**, *93*, 1119–1124. [CrossRef]
4. Algan, Y.; Cahuc, P.; Shleifer, A. *Teaching Practices and Social Capital*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2011.
5. Acemoglu, D.; Jackson, M.O. *History, Expectations, and Leadership in the Evolution of Social Norms*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2011.
6. Bowles, S. Endogenous preferences: The cultural consequences of markets and other economic institutions. *J. Econ. Lit.* **1998**, *36*, 75–111.
7. Laffont, J.J.; Martimort, D. *The Theory of Incentives: The Principal-Agent Model*; Princeton University Press: Princeton, NJ, USA, 2009.
8. Prendergast, C. The provision of incentives in firms. *J. Econ. Lit.* **1999**, *37*, 7–63. [CrossRef]
9. Edmans, A.; Gabaix, X. Executive compensation: A modern primer. *J. Econ. Lit.* **2016**, *54*, 1232–1287. [CrossRef]
10. Tabellini, G. The scope of cooperation: Values and incentives. *Q. J. Econ.* **2008**, *123*, 905–950. [CrossRef]
11. Bisin, A.; Verdier, T. The economics of cultural transmission and the dynamics of preferences. *J. Econ. Theory* **2001**, *97*, 298–319. [CrossRef]
12. Ainslie, G. Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control. *Psychol. Bull.* **1975**, *82*, 463–496. [CrossRef] [PubMed]
13. Ainslie, G. *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person;* Cambridge University Press: Cambridge, UK, 1992.

14. Meier, S.; Sprenger, C. Present-biased preferences and credit card borrowing. *Am. Econ. J. Appl. Econ.* **2010**, *2*, 193–210. [CrossRef]
15. Benhabib, J.; Bisin, A.; Schotter, A. Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games Econ. Behav.* **2010**, *69*, 205–223. [CrossRef]
16. Laibson, D. Golden eggs and hyperbolic discounting. *Q. J. Econ.* **1997**, *112*, 443–477. [CrossRef]
17. O'Donoghue, T.; Rabin, M. Doing It Now or Later. *Am. Econ. Rev.* **1999**, *89*, 103–124. [CrossRef]
18. Gul, F.; Pesendorfer, W. Temptation and Self-Control. *Econometrica* **2001**, *69*, 1403–1435. [CrossRef]
19. Bénabou, R.; Tirole, J. Self-confidence and personal motivation. *Q. J. Econ.* **2002**, *117*, 871–915. [CrossRef]
20. Bénabou, R. The economics of motivated beliefs. *Rev. Econ. Polit.* **2015**, *125*, 665–685. [CrossRef]
21. Bénabou, R.; Tirole, J. Mindful economics: The production, consumption, and value of beliefs. *J. Econ. Perspect.* **2016**, *30*, 141–64. [CrossRef]
22. Samuelson, L.; Swinkels, J. Information, evolution and utility. *Theor. Econ.* **2006**, *1*, 119–142.
23. Becker, G.S. A theory of social interactions. *J. Polit. Econ.* **1974**, *82*, 1063–1093. [CrossRef]
24. Phelps, E.S. *Altruism, Morality, and Economic Theory*; Russell Sage Foundation: New York, NY, USA, 1975.
25. Cornes, R.; Sandler, T. Easy riders, joint production, and public goods. *Econ. J.* **1984**, *94*, 580–598. [CrossRef]
26. Cornes, R.; Sandler, T. *The Theory of Externalities, Public Goods, and Club Goods*; Cambridge University Press: Cambridge, UK, 1996.
27. Steinberg, R. Voluntary donations and public expenditures in a federalist system. *Am. Econ. Rev.* **1987**, *77*, 24–36.
28. Andreoni, J. Impure altruism and donations to public goods: A theory of warm-glow giving. *Econ. J.* **1990**, *100*, 464–477. [CrossRef]
29. Varvarigos, D. Cultural norms, the persistence of tax evasion, and economic growth. *Econ. Theory* **2017**, *63*, 961–995. [CrossRef]
30. Bohnet, I.; Frey, B.S.; Huck, S. More order with less law: On contract enforcement, trust, and crowding. *Am. Polit. Sci. Rev.* **2001**, *95*, 131–144. [CrossRef]
31. Lindbeck, A.; Nyberg, S. Raising children to work hard: Altruism, work norms, and social insurance. *Q. J. Econ.* **2006**, *121*, 1473–1503.
32. Adriani, F.; Sonderegger, S. Signaling about norms: Socialization under strategic uncertainty. *Scand. J. Econ.* **2018**, *120*, 685–716. [CrossRef]
33. Adriani, F.; Sonderegger, S. The Signaling Value of Punishing Norm-Breakers and Rewarding Norm-Followers. Unpublished work, 2018.
34. Lizzeri, A.; Siniscalchi, M. *Parental Guidance and Supervised Learning*; Technical Report, Discussion Paper; Center for Mathematical Studies in Economics and Management Science: Evanston, IL, USA, 2006.
35. Doepke, M.; Zilibotti, F. *Occupational Choice and the Spirit of Capitalism*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2007.
36. Doepke, M.; Zilibotti, F. *Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission*; Technical Report; Forschungsinstitut zur Zukunft der Arbeit: Bonn, Germany, 2012.
37. Cosconati, M. *Parenting Style and the Development of Human Capital in Children*; Job Market Paper; University of Pennsylvania: Philadelphia, PA, USA, 2009.
38. Bhatt, V.; Ogaki, M. Tough Love and Intergenerational Altruism. *Int. Econ. Rev.* **2012**, *53*, 791–814. [CrossRef]
39. Bernheim, B.D.; Shleifer, A.; Summers, L.H. The strategic bequest motive. *J. Polit. Econ.* **1985**, *93*, 1045–1076. [CrossRef]
40. Lindbeck, A.; Weibull, J.W. Intergenerational aspects of public transfers, borrowing and debt. *Scand. J. Econ.* **1986**, *88*, 239–267. [CrossRef]
41. Wilhelm, M.O. Bequest behavior and the effect of heirs' earnings: Testing the altruistic model of bequests. *Am. Econ. Rev.* **1996**, *86*, 874–892.
42. Pinker, S. *The Blank Slate: The Modern Denial of Human Nature*; Penguin: London, UK, 2003.
43. Turkheimer, E. Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **2000**, *9*, 160–164. [CrossRef]
44. Heckman, J.J.; Stixrud, J.; Urzua, S. *The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2006.
45. Cappelen, A.W.; List, J.A.; Samek, A.; Tungodden, B. *The Effect of Early Education on Social Preferences*; Working Paper 22898; National Bureau of Economic Research: Cambridge, MA, USA, 2016.

46. Fudenberg, D.; Levine, D.K. A Dual-Self Model of Impulse Control. *Am. Econ. Rev.* **2006**, *96*, 1449–1476. [CrossRef] [PubMed]
47. Alger, I.; Weibull, J.W. A generalisation of Hamilton's rule: Love others how much? *J. Theor. Biol.* **2012**, *299*, 42–54. [CrossRef] [PubMed]
48. Alger, I.; Weibull, J.W. Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching. *Econometrica* **2013**, *81*, 2269–2302.
49. Bicchieri, C. *The Grammar of Society: The Nature and Dynamics of Social Norms*; Cambridge University Press: Cambridge, UK, 2006.
50. Bicchieri, C. Norms, Preferences, and Conditional Behavior. *Pol. Phil. Econ.* **2010**, *9*, 297–313. [CrossRef]
51. Binmore, K. Game theory and the social contract. In *Game Equilibrium Models II*; Springer Nature: Berlin, Germany, 1991; pp. 85–163.
52. Binmore, K. *Natural Justice*; Oxford University Press: Oxford, UK, 2005.
53. Binmore, K. Why do people cooperate? *Pol. Phil. Econ.* **2006**, *5*, 81–96.
54. Gächter, S.; Falk, A. Reputation and Reciprocity: Consequences for the Labour Relation. *Scand. J. Econ.* **2002**, *104*, 1–26. [CrossRef]
55. Lewis, D.K. *Convention: A Philosophical Study*; Blackwell Publishers: Hoboken, NJ, USA, 1969.
56. Sugden, R. Spontaneous Order. *J. Econ. Perspect.* **1989**, *3*, 85–97. [CrossRef]
57. Sugden, R. Thinking as a Team: Towards an Explanation of Nonselfish Behavior. *Soc. Philos. Pol.* **1993**, *10*, 69–89. [CrossRef]
58. Binmore, K. Social norms or social preferences? *Mind* **2010**, *9*, 139–157. [CrossRef]
59. Binmore, K. Modeling Rational Players: Part I. *Econ. Philos.* **1987**, *3*, 179–214. [CrossRef]
60. Binmore, K. Modeling Rational Players: Part II. *Econ. Philos.* **1988**, *4*, 9–55. [CrossRef]
61. Binmore, K. Do conventions need to be common knowledge? *Topoi* **2008**, *27*, 17–27. [CrossRef]
62. Elster, J. Social Norms and Economic Theory. *J. Econ. Perspect.* **1989**, *3*, 99–117. [CrossRef]
63. Paternotte, C.; Grose, J. Social Norms and Game Theory: Harmony or Discord? *Br. J. Philos. Sci.* **2012**, *64*, 551–587. [CrossRef]