FACULTY OF SOCIAL SCIENCES
Department of Economics
University of Copenhagen

Seminar Paper (Economics of Privacy)

Rasmus Bjørn

Knowledge is Privacy

A theoretic analysis of how restricting privacy to promote prosocial behaviour may hinder correct aggregation of stochastically evolving societal preferences.

Supervisor: Christoph Schottmüller

ECTS points: 7.5.

Date of submission: 09/11/2017

Keystrokes: 35.893

Knowledge is Privacy

A theoretic analysis of how restricting privacy to promote prosocial behaviour may hinder correct aggregation of stochastically evolving societal preferences.

A seminar paper in *Economics of Privacy* at the University of Copenhagen by

R. Bjørn

November 2017

"Those who don't study history are doomed to repeat it. Yet those who *do* study history are doomed to stand by helplessly while everyone else repeats it."

Abstract.

This paper explores the theoretical foundations of restricting privacy to promote *prosocial behaviour*. While this may help combat free-riding, a cost may be that policy-makers are less certain which types of prosocial behaviour citizens truly value. I explore this in the context of provision of a public good with reputation-concious agents. I formalize how societal preferences are formed over time by allowing a principal to estimate the value of the public good by observing agent behaviour. For a higher level of visibility (or, conversely, a lower level of privacy) contributions to the good are enhanced, but knowledge of the value of the public good declines. This shows a trade-off between eliminating free-riding behaviour and correct policy-making. This suggests that politicians should be careful in adjusting privacy to advocate 'proper behaviour', particularly when it is uncertain what 'proper behaviour' truly is. However, even if policy is misguided in the short term, a rational principal should adjust her policy over time.

Contents

1 Introduction: Privacy might make politicians smarter **3**

2 Related literature **4**
2.1 The odd concept of self-less behaviour (to economists) 4
2.2 Public goods, paternalism, puzzled principals and finally privacy 5

3 Model **6**
3.1 Contribution: An extension with time and evolving preferences 6
3.2 The model at hand 6
3.3 Simplifying assumptions 8

4 Characterizing the equilibrium(s) **9**
4.1 Agents: Equilibrium behaviour 9
4.2 Principal: Expectation structure & equilibrium behaviour 10
4.3 Discussion: Endogenous privacy 13

5 Results: Let's a-go exploring! **14**
5.1 Illustrating the privacy trade-off 14
5.2 What happens if societal values change suddenly? 16

6 Conclusion: Visibility should be applied carefully as a policy tool **17**

7 References **18**

A Appendix: Proofs and wonderful little oddities **19**
A.1 Agents' equilibrium behaviour (proof of proposition 4.1) 19
A.2 The failed aspirations of history-concious agents (argument for assumption) . . . 20
A.3 Principal's expectation structure (proof of proposition 4.2) 21

B Appendix: MATLAB code (available at GitHub) **22**

C Appendix: Extra figures **23**

1. Introduction: Privacy might make politicians smarter

Why do you vote? When asked people might typically state that it is important to vote. In an aggregate context, it is hard to argue gainst this. Yet the probability that your vote affects the outcome is minuscule, so self-interest is an unlikely motivational candidate. People might venture a sense of civic virtue or pride - that is, an innate satisfaction from participating in the democratic process - as their motivation. Alternatively, it may be that people thinks the general consensus is that voting is important, and fear social repercussions if they are not seen to participate. Indeed, one study found that introducing postal voting in Switzerland, which should make voting easier, actually decreased turnout in small communities (Funk 2005). In reality, it is likely that both intrinsic and reputational concerns affect motivation.

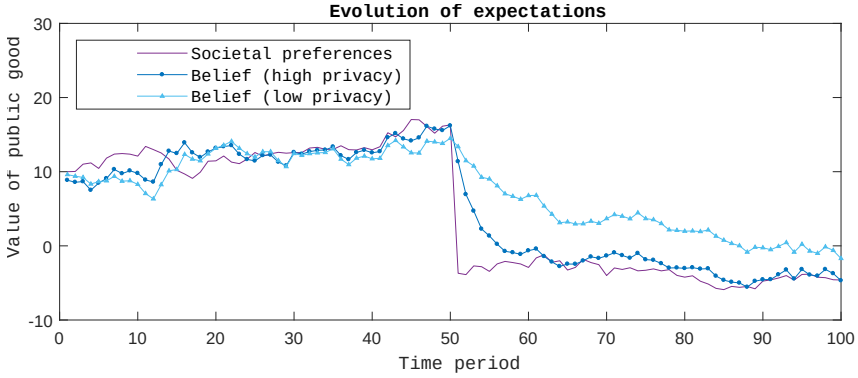
Now consider this scenario: You are the mayor of a small city, and must decide whether the city should fund a theatre. To figure out whether it is something people want you hold a town hall meeting, and put the question to the crowd. Immediately an influential citizen grabs the microphone, and proclaims that culture is necessary for civilisation to exist, and anyone who disagrees is a savage. Everyone agrees. Are you convinced everyone wants a theatre?

These cases suggests the theme explored in this paper: When policy-makers observe values in society, they cannot distinguish between people who truly believe in these values, and people who just wants to appear as if they do. If policy-makers attempt to infer what is good behaviour from the preferences they observe in society, this presents a problem..

This may be particularly problematic if values change over time. Sometimes practices or beliefs which have been relatively common are suddenly deeply scorned (Ali and Bénabou 2016, p. 4). If policy-makers are uncertain why people behave, they will adjust slowly to such changes, even if agents adjust quickly. My model formalizes how policy-makers adjust their perception of societal preferences over time. As a sneak peak at my results, consider figure 1.1. Details will be introduced later. For now, notice that if true preferences change rapidly, then policy-makers notice this far slower if they have been using visibility as a tool to maintain prosocial behaviour.

The next section relates my work to the existing literature. Section 3 formalizes the model I use and section 4 derives the main results. Section 5 explores these results using simulated games and figures like figure 1.1. Finally, the conclusions sums up on the lessons for privacy: While reducing the privacy of what people do and believe might increase free-riding and reduce prosocial behaviour, it could in theory make public policy more enlightened.

Figure 1.1: *Privacy may affect how well policy-makers decide what is in the public interest.*



2. Related literature

My analysis and model is a direct extension of Ali and Bénabou (2016), where i allow preferences to develop stochastically over time, and formalize how a principal learns about societal preferences over time. This places me in the literature srpawled by sprawled by Bénabou and Tirole (2006) (which Ali and Bénabou (2016, p. 7) was in turn built on), who developed a theory of prosocial behaviour where people are driven by a mix of directly prosocial motivations, self-interest and image concerns, which I expand upon in the first section below.

The main insight from this theory is that 'rewards or punishments (whether material or image-related) create doubt about the true motive for which good deeds are performed.' (Bénabou and Tirole 2006, p. 1652), which typically results in under provision of public goods (although it may also result in over provision, as shown in Daughety and Reinganum (2010)). The second section below considers the wider placement of this style of questions in the economic literature..

2.1. The odd concept of self-less behaviour (to economists)

The concept of self-interested behaviour likely comes easy to most economists. However, since the concept that people might be altruistic is slightly foreign to (some) economists, I will now spend a few paragraphs expanding on the two other featured concepts. Both of these serve as alternative explanations to *prosocial behaviour*[1], and may appear identical at the surface.

Altruism. I assume that people derive some intrinsic motivation from prosocial behaviour. The possibility of such motivation has been understood for a while. For instance, an older survey states that 'there appears to be a 'paradigm shift' away from the earlier position that behaviour that appears to be altruistic must, under closer scrutiny, be revealed as reflecting egoistic motives' (Piliavin and Hong-Wen 1990). This need not be altruism in a strict sense - scholars have also argued that work choice and performance may be affected by whether people see a prosocial side of it (Grant and Berg 2012)[2].

Reputation. People may instead merely want to *appear* prosocial. That is, people may be motivated to do good in part because others observe them doing it. Theoretically, this means one must also allow people actual prosocial motivations. If this was not the case, agents could not pretend to have such motivations. Yet we also need to set boundaries, since strong enough intrinsic motivations would eliminate free-rider behaviour by itself.

It can be difficult to isolate these two motivations in empirical work. Studies of prosocial motivation often rely on factor analysis on questionnaire data (see e.g. James L Perry (1996), J. L. Perry (1997) or Berthelsen, Bjørn, and Larsen (2013)). However, such evidence could be equally consistent with reputational concerns. Attempting instead to isolate the effect of reputation, Ariely, Bracha, and Meier (2009) explore image motives in an experimental context and find that monetary incentives are less effective when activity is public, reflecting image concerns. Extending results beyond the laboratory, DellaVigna, List, and Malmendier (2009) found that if people knew when fundraising volunteers would knock on their door, donations declined. They could reflect that some people give to donations mainly for reputational reasons.

[1] Prosocial behaviour is understood as 'behaviour which the actor expects will benefit the person or persons to whom it is directed' (Brief and Motowidlo 1986, p. 711), where the agent does not expect *material* rewards.

[2] Within political science the concept of public service motivation is used to describe how prosocial motivations may explain why people choose careers of public service (J. Perry and Hondeghem 2008, pp. 4-6).

2.2. Public goods, paternalism, puzzled principals and finally privacy

Public good. The basic concept I consider is that of the *public good*, taught in many a microeconomic course, where the key feature is that consumption of the good by some agents does not dilute its value to other agents (Mas-Colell, Whinston, and Green 1995, p. 359). I consider the concept in a rather broad sense: It may reflect anything from a public project like a road to societal values such as minority rights and even concepts such as racism which most would now say were always bad, but were politicians where once called upon to enforce them.

Private provision will typically be Pareto-inefficient due to *free-rider* problems (ibid., p. 351). Theoretical mechanisms may solve this, but they are often unrealistic, and rely on perfect knowledge about the value of the good (ibid., p. 364). Such knowledge does not exist in practice, and as one knows from the study of *mechanism design*, the fficient outcome can rarely be obtained if information must be elicited from agents[3](ibid., pp. 857-916). Even so, mechanism design allows us to search for the best possible outcome (Hurwicz, Maskin, and Myerson 2007, p. 2).

In this paper, I craft costs and intrinsic motivations so that the public good is insufficiently provided, and consider one mechanism which may alleviate this situation: Visibility of actions. Given the motivation structure above, higher visibility (i.e. a lower level of privacy) increases provision of the public good. If the true value of the public good was known, the efficient level could likely be achieved. However, as opponents of *paternalism* are prone to mention, true societal values are rarely known in practice. This is formalized in the approach taken here, as, agents only observe a signal on the true value of the good, and the principal is only allowed to infer the value based on the actions of agents.

Societal preferences. The key insight from Ali and Bénabou (2016), is that the use of visibility as a policy tool dilutes the information available to policy-makers. If society's concept of good behaviour is based on what agents do, and behaviour is partly motivated by something other than what they think is right, policy-makers have a hard time figuring out what to do. The analysis focuses on this, rather than on fine-tuning privacy to achieve the first-best outcome.

Thus, the main contribution of the paper is to the game-theoretic literature on how policy-makers form their concept of societal welfare[4]. Since societies exist more or less indefinitely, this is a never-ending process. I highlight this by extending the game of Ali and Bénabou (ibid.) with a time-dimension, and formalizing how the policy-makers view of societal values relate to the true preferences. This allows me to explore how persistent errors in societal values are over time. Particularly, I explore how a higher level of privacy actually reduces errors in societal preferences, because the principal may trust observed behaviour to reflect underlying preferences.

Privacy. Finally, I provide a contribution to the literature on the *economics of privacy*. This literature deals with a highly diverse set of issues[5], and a main insight is that 'In some [scenarios] privacy protection can decrease individual and societal welfare; in others, privacy protection enhances them' (Acquisti, Taylor, and Wagman 2015, p. 42). This is consistent with the findings in this paper, where stronger privacy allows policy-makers better information, but

[3]The efficient outcome may be obtained in quasi-linear environments, but will usually break other logical conditions such as budget balance, voluntary participation or plausibility of the mechanism

[4]While the existence of a social choice/welfare functions Fare often taken as a given in modelling, the proper formation of it depends crucially on observability of preferences (Mas-Colell, Whinston, and Green 1995, p. 807)

[5]See Acquisti, Taylor, and Wagman (2015) for a survey

increases free-riding.

3. Model

The following section lean heavily on the model of Ali and Bénabou (2016). I first present the main part of my extension of their model, and then present the used model in its entirety.

3.1. Contribution: An extension with time and evolving preferences

The basic setting is that of a *public good*, which agents and a principal decide to contribute to based on their inference of its value.Ali and Bénabou (ibid.) consider this as a one-time game. However, they themselves are well aware that the trade-offs of privacy are particularly interesting when 'societal attittudes... are prone to signficant change' (ibid., p. 24). They model this based on the variance of values, but note also that values 'change over time, sometimes quite radically and very fast' (ibid., p. 5). I model this explicitly by considering a *repeated game*.

Consider a measure of time $t \in \{0, 1, ..., T\}$. Within each time period t, agents and the principal decide on contributions to the public good. For simplicity, there is no overlap of generations so each agent features only once[6]. Similarly, while the principal can technically be assumed to survive generations, I shall assume that she only maximizes utility based on the current period. Intuitively, this corresponds to the notion that current agents elect a principal. The result is that only "history" in the form of previous periods affect the game.

The second extension I implement is that societal values evolve stochastically over time[7]. I $t = 0$ correspond to the one-off game, so that everyone believes initial values are distributed as $\theta_0 \sim N(\bar{\theta}_0, \sigma^2_{\theta_0})$. In future periods, this evolves stochastically as a *random walk* with I.I.D walk components $\eta_t \sim N(\bar{\eta}, \sigma^2_\eta)$:

$$\theta_t = \begin{cases} \theta_0 \\ \theta_{t-1} + \eta_t \end{cases} = \begin{cases} \theta_0, & t = 0 \\ \theta_0 + \sum_{k=1}^{t} \eta_k, & t \in [1, T] \end{cases} \tag{3.1}$$

For any given period $t > 0$, θ_t will be a sum of independent normally distributed variables. It then follows that it also follows a normal distribution with summed parameters, so $\theta_t \sim N(\bar{\theta}_0 + t\eta, \sigma^2_{\theta_0} + t\sigma^2_\eta)$, which we may alternatively write as $\theta_t \sim N(\bar{\theta}_t, \sigma^2_{\theta_t})$.

By imposing I.I.D on η_t, I restrict myself to three interesting cases: Either we have $\bar{\eta} = 0$ where one should expect no systematic change over time, or we have trending social values with $\bar{\eta} \neq 0$.[8]. Both cases have easy real world examples (consider for instance unemployment insurance, which remains contentious, versus minority rights, which tend to be seen as more and more favourable). Finally, if $\bar{\eta} = 0$ and $\sigma^2_\eta = 0$, societal preferences are constant over time.

3.2. The model at hand

The model has two types of players: Agents and the principal. I consider each in turn. Unless otherwise specified, all averages (e.g. $\bar{\theta}_0$) and variances (e.g. $\sigma^2_{\theta_0}$) are *common knowledge*.

[6] Ali and Bénabou (2016, p. 24) suggest extending to an *overlapping generations framework*. Their intent, however, is to study *lifecycle effects*, where I merely intend to explore the adaption to changing norms

[7] This was also suggested by Ali and Bénabou (ibid., p. 24), but again for an OLG analysis.

[8] Trends are constant. Alternatively, we might have allowed $\bar{\eta}$ to change over time, perhaps stochastically.

3.2.1. Agents: What does the citizens want?

Type. I assume a continuum of small agents ($i \in [0,1]$) in each period t. They each select a contribution level $a_i \in \mathbb{R}$ at cost $C(a_i) = \frac{1}{2}a_i^2$. The type of each agent will consist of three factors $(v_{i,t}, \theta_{i,t}, \mu_{i,t})$ reflecting intrinsic motivation, signal about the public good and reputational motivation. Each term is independent from each other and, since there is no generational overlap, also independent over time. They are best understood in terms of the agent's utility function.

Pay-off. The first terms is featured in the agent's *direct* (non-reputational) utility $U_{i,t}$. I assume agents derive an *intrinsic motivation* from contributing to the good, which has both a stochastic component $v_i \sim \mathcal{N}(\bar{v}, s_v^2)$ and the common shift factor θ_t, reflecting that people prefer to contribute to useful goods.

$$U_{i,t}(v_{i,t}, \theta_t, w; a_{i,t}, \bar{a}_t, a_{P,t}) = \underbrace{(v_{i,t} + \theta_t)a_{i,t}}_{\text{Intrinsic motivation}} + \underbrace{(w + \theta_t)(\bar{a}_t + a_{P,t})}_{\text{Value from public good}} - \underbrace{C(a_i)}_{\text{Cost}} \tag{3.2}$$

Agents also derive value from the total provision of the public good $\bar{a}_t + a_{P,t}$, which consist of total private contributions and the principal's contribution. I assume $w < \bar{v}$, which ensures that intrinsic motivation does not solve the free-rider problem by itself.

The second term $\mu_{i,t}$ is featured in the *reputational* utility. Each agent cares about the inference of others about $v_{i,t}$ and wishes to appear prosocial (i.e. as a good citizen rather than a free rider). Other agents do not observe whether a given contribution $a_{i,t}$ was motivated by prosocial motivation $v_{i,t}$ or reputational concerns $\mu_{j,t}$, but they will forecast this based on $\theta_{j,t}$, $\mu_{j,t}$ and the average contribution $\bar{a}_t$. Agent's *ex ante* expected social image will therefore be:

$$R(a_{i,t}, \theta_{i,t}, \mu_{i,t}, \bar{\boldsymbol{a}}_{t-1}) = E_{\bar{a}_t, \theta_{-i,t}, \mu_{-i,t}}\left[\int_0^1 E\left[v_{i,t} | a_{i,t}, \bar{a}_t, \theta_{j,t}, \mu_{j,t}, \bar{\boldsymbol{a}}_{t-1}\right] \mathrm{d}j \middle| \theta_{i,t}, \mu_{i,t}, \bar{\boldsymbol{a}}_{t-1}\right] \tag{3.3}$$

$\mu_{i,t}$ captures how much each agent captures about this social image, through a net utility pay-off of $\mu_{i,t} x_t \left(R(a, \theta_{i,t}, \mu_{i,t}) - \bar{v}\right)$. I assume that this preference is distributed across individuals as $\mu_{i,t} \sim N(\mu_t, s_\mu^2)$, which itself varies across a common prior $\bar{\mu}$[9], so that $\mu_t \sim N(\bar{\mu}, \sigma_\mu^2)$[10]. Finally, $x_t \geq 0$ is the degree of *visiblity* in the model: A higher degree of visibility enhances reputational motivations. Combining these parts, the agents solve the following problem[11]:

$$\max_{a_{i,t} \in \mathbb{R}} \left(\underbrace{E\left[U_{i,t}\left(v_{i,t}, \theta_t, w; a_{i,t}, \bar{a}_t, a_{P,t}\right) | \theta_{i,t}\right]}_{\text{Direct pay-off}} + \underbrace{x_t \mu_{i,t}\left[R(a_{i,t}, \theta_{i,t}, \mu_{i,t}) - \bar{v}\right]}_{\text{Reputational pay-off}} \right) \tag{3.4}$$

Knowledge. The type of each agent is privately known to him. The final part of his type is a private signal about the value of the public good with an error $\epsilon_{i,t} \sim N(0, s_\theta^2)$. Note that the prior depends on t as $\bar{\theta}_t = \bar{\theta}_0 + t\eta$ and $\sigma_{\theta_t}^2 = \sigma_{\theta_0}^2 + t\sigma_\eta^2$ due to the random walk in (3.1).

$$\theta_{i,t} = \theta_t + \epsilon_{i,t} \sim N(\bar{\theta}_t, s_\theta^2 + \sigma_{\theta_t}^2) \tag{3.5}$$

[9] Assume that $\bar{\mu}$ is large enough that a high fraction desires a positive reputation

[10] Allowing μ_t to vary over time simplifies the problem, as one cannot use variation over time to infer μ_t.

[11] Utility is separable in direct and reputational motivation for simplicity (Ali and Bénabou 2016, p. 12)

3.2.2. Principal

Pay-off. The principal may derive utility from two sources. First, she gains utility from the total utility observed by agents, representing a (partially) beneficial principal (weighted by $\lambda \in [0,1]$). Secondly, she may observe her own private benefits from the supply of the public good.

$$\begin{aligned} V_t(\bar{a}_t, a_{P,t}, \theta_t) = \lambda \left[\alpha \int_0^1 (v_{i,t} + \theta_t) a_{i,t} \, \mathrm{d}i + (w + \theta_t)(\bar{a}_t + a_{P,t}) - \int_0^1 C(a_{i,t}) \, \mathrm{d}i \right] \\ + (1 - \lambda) \left[b(w + \theta_t)(\bar{a}_t + a_{P,t}) - k_P C(a_{P,t}) \right] \end{aligned} \tag{3.6}$$

$\alpha \in [0,1]$ captures the extent to which the principal internalizes intrinsic motivation. Reputational motivations are not present, as these sum to zero across agents as $\int_0^1 R(a_{i,t}, \theta_{i,t}) \, \mathrm{d}i = \bar{v}$. k_P defines the Principals direct costs from contributing the good relative to other agents, while $b \in \mathbb{R}$ represents private benefits she derives from the total supply of the public good.

Note that this includes, as special cases both a completely self-less principal ($\lambda = 1$) and a purely selfish one ($\lambda = 0$). It also encompasses a standard social planner ($\lambda = 0.5, b = 0$) who values agents and her own costs of provision equally.

Knowledge. While the principal does not observe a signal, she may instead observe $\bar{a}_t$ and use this to infer a signal om θ_t. As an extension in my model, she also observes "history" in the form of all previous the sequence $\{\bar{a}_0, \bar{a}_1, ..., \bar{a}_{t-1}\}$. This allows her to better estimate θ_t over time, but is counteracted by the fact that θ_t also evolves stochastically.

3.2.3. Structure of the game

Within each time period t, the game will unfold as follows.

1. If $t > 0$, the principal observes "history" in the form of the vector[12] $\bar{\boldsymbol{a}}_{t-1} = (\bar{a}_0, \bar{a}_1, ..., \bar{a}_{t-1})'$.
2. Agents observe their private type $(v_{i,t}, \theta_{i,t}, \mu_{i,t})$ and choose their contribution $a_{i,t} \in \mathbb{R}$.
3. The principal observes $\bar{a}_t$, and chooses her contribution $a_{P,t} \in \mathbb{R}$.
4. The total supply $\bar{a} + a_P$ is enjoyed by everyone.

The game is repeated T times. Following Ali and Bénabou (2016) I keep the problem tractable by focusing on Perfect Bayesian Equilibria where agent's contributions are linear in their type.

3.3. Simplifying assumptions

I have made quite a few simplifying assumptions to keep the problem tractable and focus focus on the feature of main interest: How does the principal learn from the behaviour of agents?

Forgetful agents. The extension of the model greatly complicates analysis. Particularly, if agents are allowed to observe history, it becomes technically difficult to compute later conditional expectation of θ_t based on their behaviour[13]. To simplify analysis, I therefore assume that agents are *forgetful* in the sense that they do not observe previous iterations of the game. This leads them to disregard $\bar{\boldsymbol{a}}_{t-1}$ and $\boldsymbol{a_P}_{t-1}$, even though they could be used to glean information from the signals θ_i received by previous generations. However, the focus of this paper is not how *agents* are affected by new information.

[12] Throughout the paper, I use bold symbols to denote vectors and matrices.
[13] Appendix A.2 partially develops such a model, to highlight the technical difficutlies.

Principal receives no signal. Ali and Bénabou (ibid.) include a private signal similar to agents' for the principal. I do not. It would have been decently simple to include it[14], but the only result would have been that the principal's expectations would be more precise. Forcing the principal instead to rely on agent's behaviour for her inference exacerbates the trade-off between visibility as a policy tool and behaviour as a correct signal which this paper is interested in.

No-one looks forward. I assume that everyone maximizes utility *within* periods. This is solely to simplify analysis, and I discuss that a forward-looking principal might be interesting in section 4.3. However, one could argue that it is not unrealistic to presume that both people and policy-makers main consider the interests of current generations when setting policy.

Exogenous privacy. One of the key features of Ali and Bénabou (ibid.) is that they endogenize privacy by placing the level of x_t under the principal control. Section 4.3 argues that while it would be interesting to consider the endogenous evolution of privacy over time in this model, the derivations becomes too technically cumbersome to solve in this paper. Instead, I compare the main results for different levels of privacy in section 5.1.

4. Characterizing the equilibrium(s)

I now characterize equilibrium behaviour by first agents and then the principal. The agents' problem remains quite simple due to the assumptions just covered, and reproduces the result in Ali and Bénabou (ibid.). The principal's expectation structure becomes slightly complicated but remains quite elegant due to the niceties of the normal distribution.

4.1. Agents: Equilibrium behaviour

Agents, solving (3.4), will set contributions based on the first order condition:

$$C'(a_{i,t}) = v_{i,t} + \mathrm{E}[\theta_t|\theta_{i,t}] + x_t\mu_{i,t}\frac{\partial R(a_{i,t},\theta_{i,t},\mu_{i,t})}{\partial a_{i,t}} \tag{4.1}$$

Each term reflects one of the agent's basic motivations: His intrinsic motivation, his belief about the value of the public good[15] and his reputational concerns.

4.1.1. The agent's expectation structure

The key uncertainty in the model is θ_t. Agents have a prior belief that $\theta_t \sim N\left(\hat{\theta}_t, \sigma_t^2\right)$, but also observes a signal $\theta_{i,t} = \theta_t + \epsilon_{i,t} \sim N(\bar{\theta}_t, s_\theta^2 + \sigma_{\theta_t}^2)$. This can be seen as a simple version of the principal's expectation structure, and it may be wise to take the time to see the logic.

Lemma 4.1. In any period t, θ_t and $\theta_{i,t}$ form a bivariate normal distribution

$$\begin{pmatrix}\theta_t\\ \theta_i\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\theta}_t\\ \bar{\theta}_t\end{pmatrix}, \begin{pmatrix}\sigma_{\theta_t}^2 & \sigma_{\theta_t}^2\\ \sigma_{\theta_t}^2 & \sigma_{\theta_t}^2 + s_\theta^2\end{pmatrix}\right) \tag{4.2}$$

[14]The principal would just include more signals in her inference of θ_t in proposition 4.9.
[15]Note however that θ_t is featured here because I assumed it affected intrinsic motivation.

An agent's expectation on θ_t will be

$$\mathrm{E}[\theta_t|\theta_{i,t}] = (1-\rho_t)\,\bar{\theta}_t + \rho_t\theta_{i,t} \tag{4.3}$$

where $\rho_t \equiv \sigma^2_{\theta_t} / \left(\sigma^2_{\theta_t} + s^2_\theta\right)$ is the *signal-to-noise* ratio of $\theta_{i,t}$

Proof. Any linear combination of θ_t and $\theta_{i,t}$ will follow a univariate normal distribution, which establishes a bivariate normal distribution[16]. Covariances follow from the independence between θ_t and $\epsilon_{i,t}$: $\mathrm{Cov}(\theta_t, \theta_{i,t}) = \mathrm{Cov}(\theta_t, \theta_t + \epsilon_{i,t}) = \mathrm{V}(\theta_t, \theta_t) = \sigma^2_t$. The conditional mean of a bivariate vector (X_1, X_2) may be written as $\mathrm{E}[X_1|X_2 = a] = \mathrm{E}[X_1] + \mathrm{Cov}\,(X_1, X_2)\,/\,(\mathrm{V}\,(X_2))\,(a - \mathrm{E}[X_2])$ (Bain and Engelhardt 1992, p. 187). Applying this yields the lemma. □

4.1.2. Optimal behaviour

The third term of the FOC in (4.1) captures the expected effect on reputation of altering $a_{i,t}$. Fortunately for us, and perhaps due to clever model design, this effect can be shown to be constant (see appendix A.1), which leads to a unique outcome. The result follows exactly that of Ali and Bénabou (2016, p. 13).

Proposition 4.1 (Equilibrium behaviour)**.** For a given $x_t \geq 0$, there is a unique linear equilibrium in period t where an agent of type $(v_{i,t}, \theta_{i,t}, \mu_{i,t})$ chooses

$$a_{i,t} = v_{i,t} + \rho_t\theta_i + (1-\rho_t)\bar{\theta}_t + x_t\mu_{i,t}\xi_t(x_t) \tag{4.4}$$

where $\rho_t = \sigma^2_{\theta_t}/(s^2_\theta + \sigma^2_{\theta_t})$ and $\xi_t(x_t)$ is defined in (A.6). The aggregate contribution will be

$$\bar{a}_t = \bar{v} + \rho_t\theta_t + (1-\rho_t)\bar{\theta} + x_t\mu_t\xi_t(x_t) \tag{4.5}$$

Proof. See appendix A.1. □

Contributions increase with higher intrinsic motivation $v_{i,t}$, a higher signal $\theta_{i,t}$ or a stronger image concern μ_i. They are also increasing in x_t, since $\xi_t(x_t)$ is increasing in x_t (ibid., p. 15).

$\xi_t(x_t)$ reflects the marginal effect on expected image one contributes an extra unit and, strikingly, it is *the same* for all agents. The reason is that an observer attempting to infer $v_{i,t}$ does not need to estimate separately $\theta_{i,t}$ and $\mu_{i,t}$, but only the linear combination $\rho_t\theta_{i,t}+\mu x_t\xi_t(x_t)$. To do this she would rely on the sufficient statistic $\rho_t\theta + \mu_t x_t\xi_t(x_t)$ (which may be derived from $\bar{a}_t$), which does not depend on her signals $\theta_{j,t}$ and $\mu_{j,t}$ (ibid., p. 14).

4.2. Principal: Expectation structure & equilibrium behaviour

4.2.1. The principal forms a multivariate normal distribution of signals...

The following section covers the main contribution of my model. The principal observes no private signal, but must rely solely on the behaviour of agents to infer the value of the public good. She does so by observing the history vector $\bar{\boldsymbol{a}}_t = (\bar{a}_0, \bar{a}_1, ..., \bar{a}_t)'$.

[16] More formally, note that we may write $\begin{pmatrix}\theta_t \\ \theta_i\end{pmatrix} = \begin{pmatrix}\sigma_{\theta_t} & 0 \\ \sigma_{\theta_t} & s_\theta\end{pmatrix}\boldsymbol{W} + \begin{pmatrix}\bar{\theta}_t \\ \bar{\theta}_t\end{pmatrix}$ where $\boldsymbol{W}$ is a 2×1 vector of independent $N(0,1)$ variables. This is an operational definition of a multivariate normal distribution (MIT 2008).

Lemma 4.2. Observing $\bar{a}_t$, the principal may obtain a noisy but unbiased signal of θ_t.

$$\begin{aligned}\hat{\theta}_t &\equiv \frac{1}{\rho}\left(\bar{a} - \bar{v} - (1-\rho)\bar{\theta} - x_t\bar{\mu}\xi_t(x_t)\right) = \theta_t + \left(\frac{x_t\xi_t(x_t)}{\rho_t}\right)(\mu_t - \bar{\mu}) \\ &\sim N\left(\bar{\theta}_t, \sigma^2_{\theta_t} + \left(\frac{x_t\xi_t(x_t)}{\rho_t}\right)^2\sigma^2_\mu\right) = N\left(\bar{\theta}_t, \sigma^2_{\theta_t} + \gamma^2_t(x_t)\sigma^2_\mu\right)\end{aligned} \tag{4.6}$$

Proof. Invert (4.1), and replace the unknown μ_t with the expected value $\bar{\mu}$. □

Consider first the simple case of $t = 0$. Then she only observes one signal, $\hat{\theta}_0$, which is an unbiased signal akin to $\theta_{i,t}$ so that it forms a bivariate normal distribution with θ_t.

$$\begin{aligned}\begin{pmatrix}\theta_0 \\ \hat{\theta}_0\end{pmatrix} &\sim N\left(\begin{pmatrix}\bar{\theta}_0 \\ \bar{\theta}_0\end{pmatrix}, \begin{pmatrix}\sigma^2_{\theta_0} & \sigma^2_{\theta_0} \\ \sigma^2_{\theta_0} & \sigma^2_{\theta_0} + \gamma^2_t(x_t)\sigma^2_\mu\end{pmatrix}\right) \\ &\Rightarrow \mathrm{E}[\theta_|\theta_i] = \left(1 - \frac{\sigma^2_{\theta_0}}{\sigma^2_{\theta_0} + \gamma^2_t(x_t)\sigma^2_\mu}\right)\bar{\theta}_0 + \frac{\sigma^2_{\theta_0}}{\sigma^2_{\theta_0} + \gamma^2_t(x_t)\sigma^2_\mu}\hat{\theta}_0\end{aligned} \tag{4.7}$$

This already reflects they key characteristic: The signal becomes more noisy if privacy is used as a policy tool to affect behaviour, creating uncertainty about the value of the public good.

In general, the principal observes $\bar{\boldsymbol{a}}_t = (\bar{a}_0, \bar{a}_1, ..., \bar{a}_t)'$, and therefore computes a vector of signals $\hat{\boldsymbol{\theta}}_t = \left(\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_t\right)$. These will be independently drawn signals on θ_t, so each linear combination of these also forms a univariate normal distribution. It follows that they form a *multivariate* normal distribution (see appendix A.3 for a formal proof).

Proposition 4.2 (Principal's expectation.)**.** In any period t, θ_t and $\hat{\boldsymbol{\theta}}_t = \left(\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_t\right)$ form a multivariate normal distribution

$$\begin{pmatrix}\theta_t \\ \hat{\boldsymbol{\theta}}_t\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\theta}_t \\ \bar{\boldsymbol{\theta}}_1\end{pmatrix}, \begin{pmatrix}\sigma^2_{\theta_t} & \boldsymbol{\sigma}^{\mathbf{2}\,\prime}_{\boldsymbol{\theta_t}} \\ \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\theta_t}} & \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_t}(x_t)\end{pmatrix}\right) \tag{4.8}$$

where $\boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\theta_t}} = \left(\sigma^2_{\theta_0}, \sigma^2_{\theta_1}, ..., \sigma^2_{\theta_t}\right)'$ is the covariance between each signal and θ_t and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_t}(x_t)$ is the covariance matrix for the signals $\hat{\theta}_t$. Its elements are computed in appendix A.3.
The principal's expectation on θ_t will be

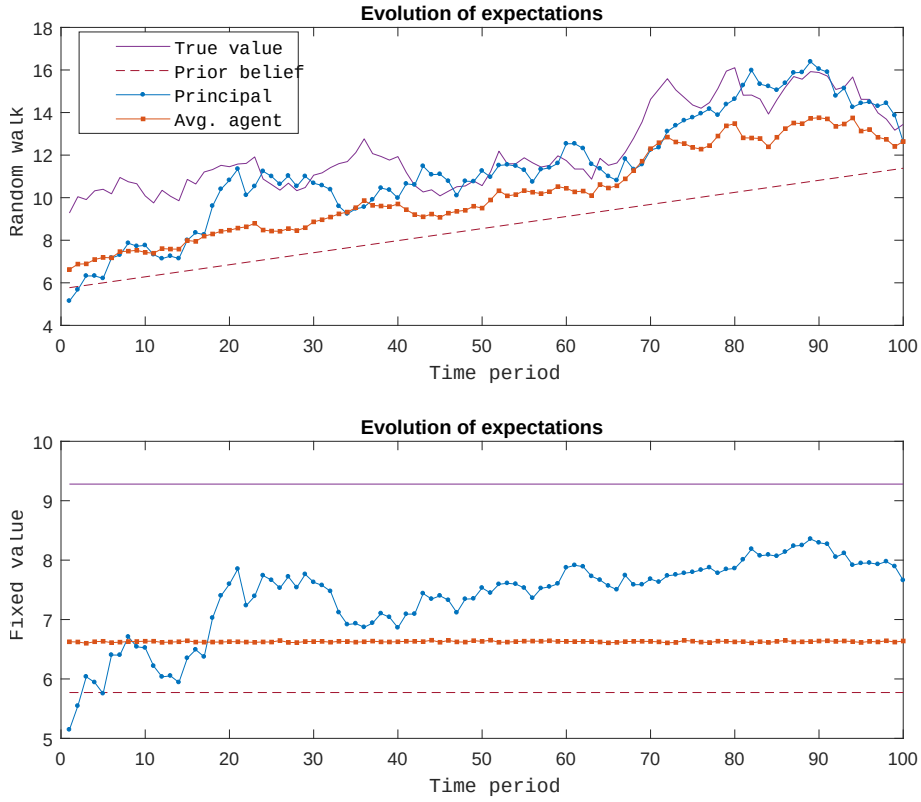
$$\mathrm{E}[\theta_t|\hat{\boldsymbol{\theta}}_t] = \bar{\theta}_t + \boldsymbol{\sigma}^{\mathbf{2}\,\prime}_{\boldsymbol{\theta_t}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\hat{\boldsymbol{\theta}}_t}(x_t)\left(\hat{\boldsymbol{\theta}}_t - \bar{\boldsymbol{\theta}}_t\right) \tag{4.9}$$

Proof. See appendix A.3 for a proof of both multivariate normality and the expectation. □

The bivariate case is a special case of this, as is the agent's problem. The proof of normality holds for these cases as well. Multiplying out the matrices, each each signal is again weighted by a signal-to-noise ratio based on the covariance and variance of the individual signal.

To illustrate the concepts, consider figure 4.1. The lower part shows a simply case where $\theta_t = \theta_0$ in all periods. Here, it is extremely clear that the principal improves her prediction over time, while the agents do not. In the upper part θ_t evolves as a random walk with a positive trend. This is an example of the general case I have considered. Note that the principals expectation tends to follow the movements of the random walk, but that she still suffers considerable noise.

Figure 4.1: *The expectation structure*



4.2.2. ... and then she sets her optimal contribution

Consider the $x_t \geq 0$ the principal sets at the start of the period fixed for now. She then sets her contribution according to a first order condition derived from (3.6).

$$C'(a_{P,t}) = \frac{\left(w + \mathrm{E}\left[\theta_t | \hat{\boldsymbol{\theta}}_t\right]\right)(\lambda + (1-\lambda)b)}{(1-\lambda)k_P} \tag{4.10}$$

Proposition 4.3 (Optimal contribution)**.** For a given $x_t \geq 0$ the principal will contribute

$$a_{P,t} = \frac{\left(w + \bar{\theta}_t + \boldsymbol{\sigma}^{\mathbf{2}}_{\hat{\boldsymbol{\theta}}_t}{}' \boldsymbol{\Sigma}^{-\mathbf{1}}_{\hat{\boldsymbol{\theta}}_t}(x_t)\left(\hat{\boldsymbol{\theta}}_t - \bar{\boldsymbol{\theta}}_t\right)\right)\varphi}{(1-\lambda)k_P} \tag{4.11}$$

where $\varphi \equiv (\lambda + (1-\lambda)b)$ is principal's total gain per (efficiency) unit added to the total supply of the public good, no matter the source.

Proof. Insert $C'(a_{P,t}) = a_{P,t}$ and (4.9) in (4.10). □

Greater visibility x_t dilutes the signal $\hat{\theta}_t$, as a larger contribution $\bar{a}_t$ might reflect reputational concerns, rather than the value of the good. However, a higher x_t also works against the free-rider incentive, since it motivates people to contribute. This illustrates the trade-off of privacy.

$\hat{\theta}_t$ increases in $\bar{a}_t$, reflecting that people have received higher average signals. This shows that my solution follows the same logic as Ali and Bénabou (2016, p. 19), who argue that one

may divide the optimal contribution into a base level $\underline{a}_{P,t}$ and a matching rate on $\bar{a}_t$. Due to the more complex expectation structure here however, the division cannot be elegantly performed.

4.2.3. First-best outcomes

One may consider how these results relate to the both outcomes considered optimal for either society at large, or for an agent in the model. Below, three interesting cases are covered.

Claim 4.1. If θ_t was known, the following quantities would be preferred by the specified actor:

$$\begin{aligned} \text{Aggregate agents:} \qquad & \bar{a}_t = \bar{v} + w + 2\theta_t \\ \text{Principal:} \qquad & a_{P,t} = (w + \theta_t)\,\varphi/((1-\lambda)k_P) \end{aligned} \tag{4.12}$$

Proof. They follow directly from differentiating (3.4) and (3.6), noting that reputational concerns cancel out in the aggregate since $\int_0^1 R(a_{i,t}, \theta_{i,t})\,\mathrm{d}i = \bar{v}$ by the law of iterated expectations. □

The preferences of aggregate agents and the principal may be understood as the true societal preferences. If $x_t = 0$[17], then the public good is underprovided (since agents are atomistic, they would free-ride completely if they had no intrinsic motivation and contribute nothing).

4.3. Discussion: Endogenous privacy

Note that the level of privacy x_t could be placed under the principal's control, as it is in Ali and Bénabou (ibid.). The principal may anticipate both the behaviour of agents and of herself at the later stage at the beginning of each period. Based on the history vector $\bar{\boldsymbol{a}}_{t-1} = (\bar{a}_0, \bar{a}_1, ..., \bar{a}_{t-1})'$ she may form $\mathrm{E}[\theta_t|\hat{\boldsymbol{\theta}}_{t-1}$, and relies on her priors for other variables.

4.3.1. The derivations would be particularly cumbersome...

Both $\bar{a}_t$ and $a_{P,t}$ will depend on x. She does not know these quantities at this stage, and must use her interim beliefs. Again, these will reflect the more complicated (and precise) information structure available to the principal.

Claim 4.2. At the start of each period t, the principal may form the *interim* beliefs

$$\begin{aligned} \mathrm{E}[\bar{a}_t|\bar{\boldsymbol{a}}_{t-1}] = \tilde{\bar{a}}_t = \bar{v} + (1-\rho_t)\bar{\theta} + x_t\bar{\mu}\xi_t(x_t) \\ + \rho_t\left(\bar{\theta}_{t-1} + \bar{\eta} + \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\theta_{t-1}}}{}'\boldsymbol{\Sigma}^{\mathbf{-1}}_{\hat{\boldsymbol{\theta}}_{\mathbf{t-1}}}(x_{t-1})\left(\hat{\boldsymbol{\theta}}_{t-1} - \bar{\boldsymbol{\theta}}_{t-1}\right)\right) \\ \mathrm{E}[a_{P,t}|\bar{\boldsymbol{a}}_{t-1}] = \tilde{a}_{P,t} = \frac{\left(w + \bar{\theta}_t + \boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\theta}_t}{}'\boldsymbol{\Sigma}^{\mathbf{-1}}_{\hat{\boldsymbol{\theta}}_t}(x_t)\left(\tilde{\hat{\boldsymbol{\theta}}}_t - \bar{\boldsymbol{\theta}}_t\right)\right)\varphi}{(1-\lambda)k_P} \end{aligned} \tag{4.13}$$

where the unknown last element of $\tilde{\hat{\boldsymbol{\theta}}}_t$ is replaced by her expectation.

Proof. They follow from proposition 4.1 and 4.3 and the principal's expectation structure. □

[17] If $x_t \geq 0$ it is less apparent what would happen, as we would need to adjust the expectation structure in appendix A.1 to ignore signals. However, it would be possible that the provision was either too low, efficient over even too high.

To derive a first order condition from (3.6), one must differentiate each of these with regard to x_t. As one may see, this is quite complicated since we need to differentiate again the matrix expression in the expectation. This is complicated by the presence of the inverse matrix, which does not have a closed solution for a general t. Due to this, and the fact that even for the close game the derivations are quite complicated (Ali and Bénabou 2016, p. 27), I do not endogenize privacy in this paper. This also actually leaves us some pages to explore the results!

4.3.2. ... but, what might happen?

Before doing so, I may offer some conjecture on what would occur if privacy was endogenized. Principals will equate the marginal benefit of increasing pro-social behaviour (the first derivative of first derivative of $a_{i,t}$) with the costs of dilluting their own information about the public good (throught the variance of $\hat{\theta}_t$) which causes her to set both $a_{P,t}$ and now x_t inefficiently.

I suspect that principals will tend to set a low level of visibility in early periods until they (believe they) have a good estimate of preferences and then crank up privacy to encourage proper behaviour[18]. The next section suggests that expectations do stabilize over time.

It would be interesting to model this in a future paper, to check if my conjecture holds up. It might be particularly interesting to consider a forward-looking principal. A principal would then only face an immediate cost from lowering visibility (more free-riding) but would see a gain in *all* future periods from the improved signal.

5. Results: Let's a-go exploring!

The previous section established the theoretical results from my extension of the model through proposition 4.1-4.3. This section will explore the structure of these results. As we've seen, the expressions are somewhat difficult to evaluate due to the presence of large inverse matrices. Therefore we rely on simulating the model in MATLAB[19] with meaningful parameter values.

5.1. Illustrating the privacy trade-off

First, let us look at the basic version of the model with no twists. In doing so, I may also define some decently meaningful parameter values. Define $\theta_0 = 10$ but $\bar{\theta}_0 = 5$, so that everyone initially *underestimates* the true value of the social good. Let $\sigma^2_{\theta_0} = 10$ so agents don't consider the value too unlikely (6 pct. chance of a higher value). Assume a slight trend in the random walk, but with high uncertainty, $\eta_t \sim N(0.03, 0.5)$, and a decently high variance in the signals $s^2_\theta = 15$.

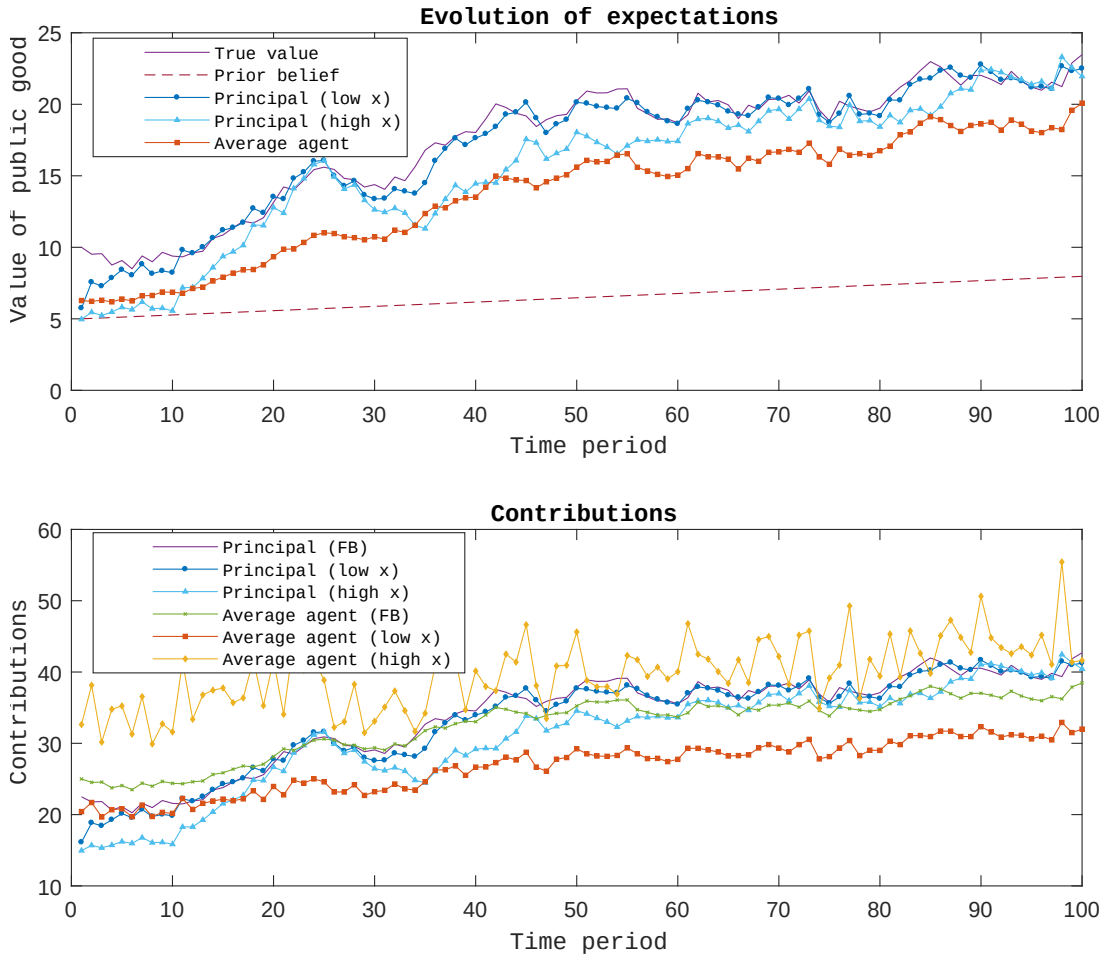
$$\begin{aligned} \text{Agents:} \quad & \bar{v} = 5, \quad \sigma^2_v = 5, \quad w = 5, \quad \bar{\mu} = 10, \quad \sigma^2_\mu = 5, \quad s\mu^2 = 5, \quad s^2_\theta = 1 \\ \text{Principal} \quad & \lambda = \frac{1}{2}, \quad b = \frac{1}{2}, \quad k_p = 1, \quad \alpha = \frac{1}{2} \end{aligned} \tag{5.1}$$

For the agent, this corresponds to initially valuing each part of intrinsic motivation equally, and fairly strong reputational concerns. The principal values equally her own and agents' costs, like a standard social planner (ibid., p. 11). She also obtains some private gains from higher contributions, and internalizes half of the agents' intrinsic motivation.

[18] This may depend on regularity conditions. Particularly, we will need a relatively large initial value θ_0 compared to random walk components, since if the latter dominates the process, she will never have a stable estimate of θ_t

[19] All code is available at GitHub. See appendix B

Figure 5.1: *The default game*



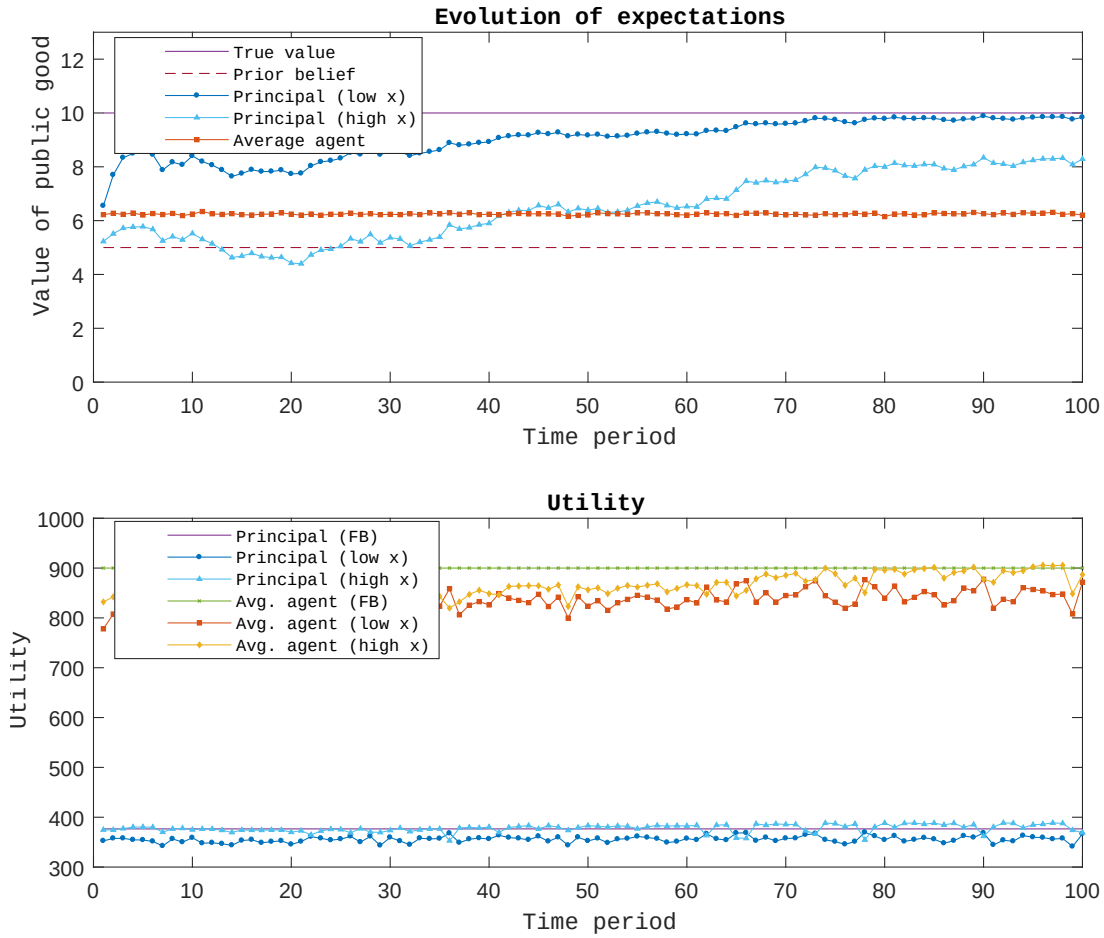
Finally, since this paper investigates privacy after all, consider two levels of visibility $x_t = x \in \{0.3, 3\}$. A random simulation is shown in figure 5.1. It directly illustrates the trade-off of privacy in our model. In the upper picture, we see that expectations are clearly more precise for the lower level of visibility. However, looking at the lower figure we see that the price of better information is rampant free-riding, as contributions are far below the optimal level for agents.

These effects may be seen even more clearly in the special case where preferences are constant $\theta_t = \theta$ (i.e.$\bar{\eta} = 0, \sigma_{\theta_t}\eta^2 = 0$. Notice in figure 5.2 that in the first periods there is a clear trend where the principal in higher privacy environment improves her forecast, but the pace is far slower in the environment of high visibility. The lower graph now shows utilities, and again shows that the case of higher information results in worse underprovision of the public good.

The principal's expectation stabilizes after some time[20]. This is even more clear in figure C.1 in the appendix which extends the time period (to $T = 1000$). This leads us back to the conjecture in section 4.3: A principal, once she notices that her expectation no longer changes, might conclude she has a good grasp of the societal preferences, and increase x_t to promote prosocial behaviour. The next section suggest why this might be dangerous.

[20]This happens quite easily when θ_t is constant. However, as long as θ_0 is sufficiently important compared to the trend, one could expect this to happen in the general case as well

Figure 5.2: *Permanent societal values*

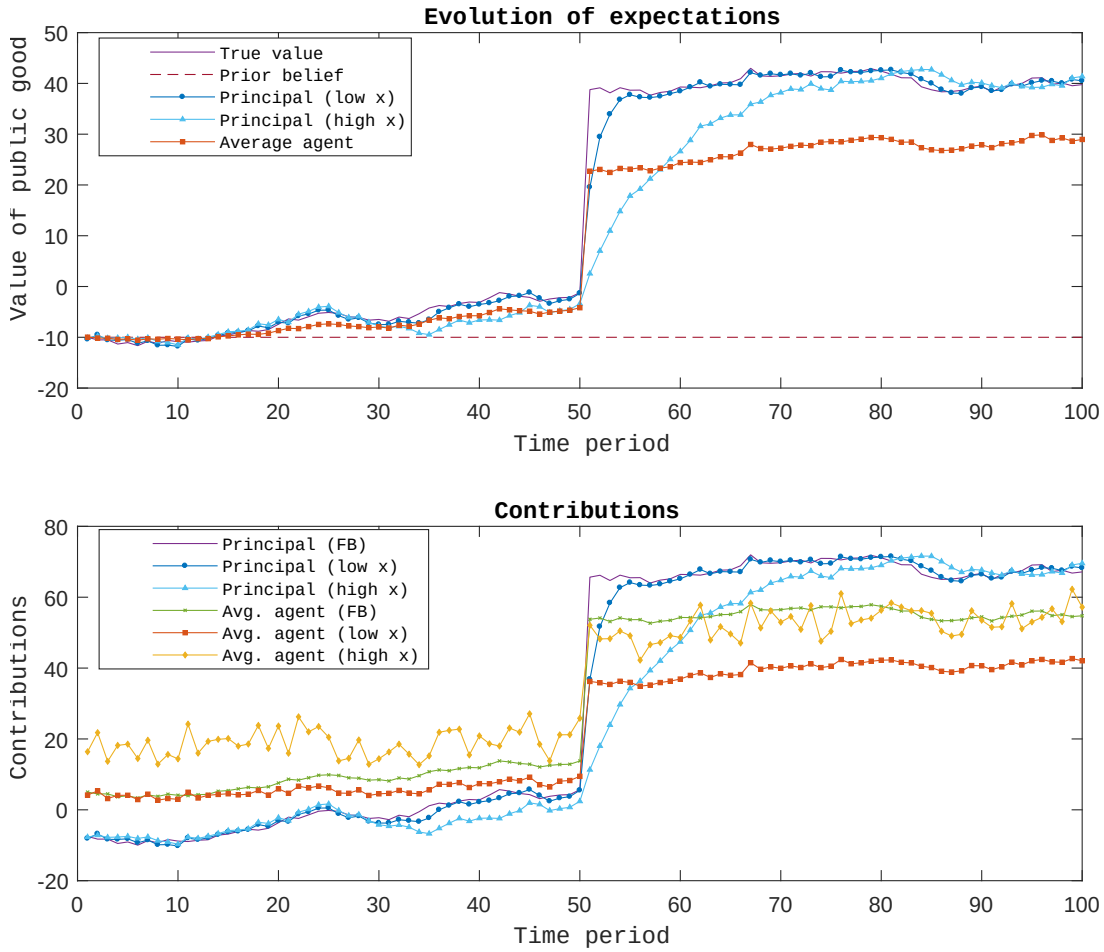


5.2. What happens if societal values change suddenly?

How would policy-makers react if societal values changed suddenly and unexpectedly? Ali and Bénabou (2016, p. 3) note that '[f]ull reversals of societal preferences, where some behaviour goes ... from intensely stigmatized to widely acceptable, like divorce, cohabitation ('living in sin'), homosexuality or drug use, are relatively common in modern societies and sometimes quite sudden'.To illustrate this, consider that we are initially in a case of a public "bad" $\theta_0 = -5$ and that expectations are quite well-calibrated with $\bar{\theta} = -5$ and no perceived trend in the random walk $\bar{\eta} = 0$. Then impose a large positive shock, say $\eta_5 0 = -4 \times \bar{\theta}_0 = 20$ (with a believed probability of approximately zero). An example of this is shown in figure 5.3.

There is a clear difference between how the principals react. In the context of higher privacy, the principal responds far more swiftly to the changed values. Ali and Bénabou (ibid., p. 3) argue that the use of visibility as a policy tool leads to 'rigid and maladaptive public policy'. As my formalization of the principal's learning clearly shows here, this is indeed true, as principals adapt slower to changing norms. However, it is also apparent than even at a higher level of visibility, the principal does still adapt to changing norms (she is rational after all). Figure C.2 in the appendix compares a far larger difference in privacy ($x_t \in \{1, 100\}$) and shows that even though the principal adapts in more slowly, she still does so. So while one should be cautious about visibility as a policy tool, one should not expect it to halt societal progress indefinitely.

Figure 5.3: *Suddenly, a public 'bad' becomes a public good*



6. Conclusion: Visibility should be applied carefully as a policy tool

The main lesson is that if *visibility of actions* is carelessly applied as a tool to increase prosocial behaviour, it may lead to rigid public policy which does not reflect the values of current society. This is not an argument against the tool per se - the previous section argued that failure to apply it may lead to sub-optimal behaviour - but it should be applied carefully.

Particular care should be taken when values change suddenly and massively. If politicians are unsure whether values are shifting, they should stay away from visibility as a policy tool. Doing so should make it far more easy to figure out whether values are indeed changing.

I formalize how perceptions of public preferences change over time. One lesson here is although policy may be rigid in the short term, it should adapt over time even with weak privacy. Governments may repress progress for a time, but they should rarely succeed in the long run.

The most interesting extension might be to endogenize the level of privacy under the control of a principal who maximize intertemporal utility. As conjectured, one such might reduce visibility in times of uncertainty, and use it increasingly as a policy tool when expectations stabilize.

A final caveat is that the paper here, and the ones cited, are entirely theoretical. Even in the face of elegant analytical solutions, it is always prudent to await empirical judgement before trusting the results completely. Another extension might be a clever research design, which compare the informational content of policy in different environments of privacy.

7. References

Acquisti, Alessandro, Curtis R. Taylor, and Liad Wagman (2015). 'The Economics of Privacy'. In: *SSRN Electronic Journal* 54, pp. 442–492. ISSN: 1556-5068. URL: http://www.ssrn.com/abstract=2580411.

Ali, S Nageeb and Roland Bénabou (2016). 'Image Versus Information'. In: *NBER Working Paper* 22203.

Ariely, Dan, Anat Bracha, and Stephan Meier (2009). 'Doing Good or Doing Well?: Imge Motivation adn Monetary Incentives in Behaving Proscially'. In: *Am. Econ. Rev.* 99, pp. 544–555. ISSN: 0002-8282.

Bain, Lee J. and Max Engelhardt (1992). *Introduction to Probability and Mathematical Statistics*. 2nd. Brooks/Cole. ISBN: 9780534380205.

Bénabou, Roland and Jean Tirole (2006). 'Incentives and Prosocial Behavior'. In: *American Economic Review* 96.5, pp. 1652–1678. ISSN: 1098-6596. arXiv: arXiv:1011.1669v3.

Berthelsen, Martin Frost, Rasmus Bjørn, and Mads Bahn Larsen (2013). 'Succes , selvudvikling eller samfundspligt ? Et studie af tre motivationstyper i kommunalt regi'. Bachelor's thesis. University of Copenhagen.

Brief, A. P. and S. J. Motowidlo (1986). 'Prosocial Organizational Behaviors.' In: *Academy of Management Review* 11.4, pp. 710–725. ISSN: 0363-7425. URL: http://amr.aom.org/cgi/doi/10.5465/AMR.1986.4283909.

Daughety, Andrew F and Jennifer F Reinganum (2010). 'Public Goods, Social Pressure, and the Choice Between Privacy and Publicity'. In: *American Economic Journal: Microeconomics* 2.2, pp. 191–221. ISSN: 1945-7669. URL: http://pubs.aeaweb.org/doi/10.1257/mic.2.2.191.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier (2009). 'Testing for Altruism and Social Pressure in Charitable Giving'. In:

Funk, Patricia (2005). 'Theory and Evidence on the Role of Social Norms in Voting'. In: *SSRN Electronic Journal*. ISSN: 1556-5068.

Grant, A. and J. Berg (2012). 'Prosocial motivation at work'. In: *The Oxford handbook of positive organizational scholarship*. ISBN: 9780199734610 (hbk.) :

Hurwicz, Leonid, Eric Maskin, and Roger Myerson (2007). 'Mechanism Design Theory'. In: *Nobel Prize* October, pp. 0–28.

Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R. Green (1995). *Microeconomic theory*. URL: https://global.oup.com/academic/product/microeconomic-theory-9780195102680?pubdatemonthto=%7B%5C&%7Dpubdatemonthfrom=%7B%5C&%7Dpubdateyearfrom=%7B%5C&%7Dpubdatemonthfrom%7B%5C_%7Ddefault=select%20month%7B%5C&%7Dauthor=Andreu%20Mas-Colell%7B%5C&%7Dtitle=%7B%5C&%7Dpubdatemonthto%7B%5C_%7Ddefault=select%20month%7B%5C&%7Dpubdateyearto=%7B%5C&%7Dbic=%7B%5C&%7Dsub.

MIT (2008). *MULTIVARIATE NORMAL DISTRIBUTIONS*. Tech. rep. Lecture 15, pp. 1–9.

Mitra, Sujit Kumar (1980). 'Generalized Inverse of a Matrix and Its Applications'. In: *Technometrics* 15.1, pp. 471–512. ISSN: 01697161. URL: http://linkinghub.elsevier.com/retrieve/pii/S0169716180800459.

Perry, J. L. (1997). 'Antecedents of Public Service Motivation'. In: *Journal of Public Administration Research and Theory*. ISSN: 1053-1858.

Perry, James L (1996). 'Measuring Public Service Motivation : An Assessment'. In: *Journal of Public Administration Research and Theory*.

Perry, J.L. and A. Hondeghem (2008). 'Editor's Introduction'. In: *Motivation in Public Management: The Call of Public Service*. Ed. by Oxford University Press. Oxford, pp. 1–14.

Piliavin, Jane Allyn and Charng Hong-Wen (1990). 'Altruism: A review of recent theory and research'. In: *Annual Review of Sociology*. ISSN: 03600572.

Wang, Ruye (2006). *Marginal and conditional distributions of multivariate normal distribution*. URL: http://www.ccs.neu.edu/home/vip/teach/MLcourse/3%7B%5C_%7Dgenerative%7B%5C_%7Dmodels/lecture%7B%5C_%7Dnotes/Marginal%20and%20conditional%20distributions%20of%20multivariate%20normal%20distribution.pdf.

A. Appendix: Proofs and wonderful little oddities

A.1. Agents' equilibrium behaviour (proof of proposition 4.1)

This proof follows exactly the structure of Ali and Bénabou (ibid., p. 25), only adding t subscripts. We restrict out attention to linear strategies in type as below. The total contributions follows from the fact that agents form a continuum $[0,1]$). I subtract $a_{i,t}-\bar{a}_t$ and rearrange.

$$\begin{aligned} a_{i,t} &= A\mu_{i,t} + Bv_{i,t} + C\theta_{i,t} + D \\ \bar{a}_t &= A\mu_t + B\bar{v} + C\theta_t + D \\ a_{i,t}-\bar{a}_t &= A(\mu_{i,t}-\mu) + B(v_{i,t}-\bar{v}) + C(\theta_{i,t}-\theta_t) \Leftrightarrow \\ Bv_{i,t} &= B\bar{v} + (a_{i,t}-\bar{a}_t) - \left(C\epsilon_{i,t}^{\theta_t} + A\epsilon_{i,t}^{\mu}\right), \qquad \epsilon_{i,t}^{\theta} = \theta_{i,t}-\theta_t, \quad \epsilon_{i,t}^{\mu_t} = \mu_{i,t}-\mu_t \end{aligned} \tag{A.1}$$

The vector $\left(Bv_{i,t}, a_{i,t}-\bar{a}_t, \bar{a}_t, \theta_{j,t}, \mu_{j,t}, \left(C\epsilon_{i,t}^{\theta_t} + A\epsilon_{i,t}^{\mu_t}\right)\right)$ will be jointly normal distributed. Every linear combination of these components is a linear combination of a set of independent normal random variables and therefore will follow a univariate normal distribution (which establishes a multivariate normal distribution). Note that $\bar{a}_t, \theta_j$ and μ_j are uncorrelated with both $(a_{i,t}-\bar{a}_t)$ and $\left(C\epsilon_{i,t}^{\theta_t} + A\epsilon_{i,t}^{\mu_t}\right)$ and since these variables are jointly normal, it follows that they are independent. Then we do not need to condition on these variables so that $\mathrm{E}\left[v_i|a_{i,t},\bar{a}_t,\theta_{j,t},\mu_{j,t}\right] = \mathrm{E}\left[v_i|a_{i,t}-\bar{a}_t\right]$. It follows that we may, without loss of information, reduce the full multivariate distribution of the components to a bivariate distribution of $(v_i, a_{i,t}-\bar{a}_t)'$ (basically, all other covariance terms will be zero, and thus uninformative).

$$\begin{pmatrix} v_i \\ a_i-\bar{a} \end{pmatrix} \sim N\left(\begin{pmatrix} \bar{v} \\ 0 \end{pmatrix}, \begin{pmatrix} s_v^2 & Bs_v^2 \\ Bs_v^2 & A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2 \end{pmatrix}\right) \tag{A.2}$$

The covariance term is derived from (A.1) where $v_{i,t}$ features for $a_{i,t}-\bar{a}_t$ with weight B. The variance is found similarly, noting that all terms for $a_{i,t}-\bar{a}_t$ are independent, so the covariance terms in the variance expression will be zero. Note that θ_t and μ here are found by integrating the signals. Thus, the realizations are featured rather than the stochastic variables (which is why they do not contribute variance). Apply the rule for conditional means to get:

$$\mathrm{E}[v_i|a_{i,t},\bar{a}_t,\theta_{i,t},\mu_{j,t}] = \mathrm{E}\left[v_i|a_{i,t}-\bar{a}_t\right] = \bar{v} + \frac{Bs_v^2}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}\left(a_{i,t}-\bar{a}_t\right) \tag{A.3}$$

We insert this expression in (3.3) and obtain:

$$\begin{aligned} R(a_{i,t},\theta_{i,t},\mu_{i,t}) &= \mathrm{E}\left[\mathrm{E}\left[v_{i,t}|a_{i,t}-\bar{a}_t\right]\middle|\theta_{i,t},\mu_{i,t}\right] \\ &= \mathrm{E}\left[\left(\bar{v} + \frac{Bs_v^2}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}\left(a_{i,t}-\bar{a}_t\right)\right)\middle|\theta_{i,t},\mu_{i,t}\right] \\ &= \bar{v} + \frac{Bs_v^2}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}\left(a_{i,t} - A\,\mathrm{E}[\mu_t|\mu_{i,t}] - B\bar{v} - C\,\mathrm{E}[\theta_t|\theta_{i,t}] - D\right) \end{aligned} \tag{A.4}$$

We could insert expressions for the expectation of μ_t and θ_t, but they won't depend on a_i, and thus won't matter. Since we take the FOC wrt. a_i, the entire parentheses simply yields a 1, and

we only need the terms outside. Insert in the FOC in (4.1) and, noting that $C'(a_i) = a_i$, get:

$$a_{i,t} = v_{i,t} + (1-\rho_t)\bar{\theta}_t + \rho_t\theta_i + x_t\mu_{i,t}\left(\frac{Bs_v^2}{A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2}\right) \tag{A.5}$$

We may then note that, using the definitions of the linear constants, $B = 1$, $C = \rho_t$, $D = (1-\rho_t)\bar{\theta}_t$ and $A = x_t s_v^2 / \left(A^2s_\mu^2 + B^2s_v^2 + C^2s_\theta^2\right)$. Substituting in $A = x_t\xi_t(x_t)$ we obtain:

$$\xi_t(x_t) = \frac{s_v^2}{x_t^2\xi_t(x)^2s_\mu^2 + s_v^2 + \rho_t s_\theta^2} \tag{A.6}$$

Given x_t, $\xi_t(x)$ is the unique solution to this for $\xi(x_t) = \xi$. To see there is just one solution, note that the RHS is continuous and decreasing in $\xi_t(x)$, clealy cutting the diagonal at a unique solution. Inserting this and the defined constants in the FOC above, we obtain the proposition.

A.2. The failed aspirations of history-concious agents (argument for assumption)

Here I provide an argument for the simplifying assumption of forgetful agents. I go easy on the formal proofs and show the intuition for why the problem complicates greatly.

Agents now skedaddle to the library, and also observe $\bar{\boldsymbol{a}}_{t-1} = (\bar{a}_0, \bar{a}_1, ..., \bar{a}_{t-1})'$. In $t = 0$, they observe $\bar{\boldsymbol{a}}_{t-1} = \emptyset$, and choose behaviour just as in proposition 4.1. In the second period however, they observe $\bar{\boldsymbol{a}}_0 = (\bar{a}_0)'$, which they invert for an unbiased signal $\hat{\theta}_t$ as defined in (4.6).

The agent now has three pieces of evidence: His original prior $\bar{\theta}_1$, his own signal $\theta_{i,1}$ and $\hat{\theta}_0$. Following the logic in appendix A.3, they form a multivariate normal distribution.

$$\begin{pmatrix}\theta_1\\ \theta_{i,1}\\ \hat{\theta}_0\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\theta}_1\\ \bar{\theta}_1\\ \bar{\theta}_0\end{pmatrix}, \begin{pmatrix}\sigma_{\theta_1}^2 & \sigma_{\theta_1}^2 & \sigma_{\theta_0}^2\\ \sigma_{\theta_1}^2 & \sigma_{\theta_1}^2 + s_\theta^2 & \sigma_{\theta_0}^2\\ \sigma_{\theta_0}^2 & \sigma_{\theta_0}^2 & \sigma_{\theta_0}^2 + \gamma_0^2(x_0)\sigma_\mu^2\end{pmatrix}\right) \tag{A.7}$$

We may write this a partition, just as in appendix A.3. Define the signals available to an agent as $\boldsymbol{\theta}_{i,1} \equiv \left(\theta_{1,t}, \hat{\theta}_0\right)'$. We may then write the distribution as the partition:

$$\begin{pmatrix}\theta_1\\ \boldsymbol{\theta}_{i,1}\end{pmatrix} \sim N\left(\begin{pmatrix}\bar{\theta}_1\\ \bar{\boldsymbol{\theta}}_1\end{pmatrix}, \begin{pmatrix}\sigma_{\theta_1}^2 & \boldsymbol{\sigma}_{\boldsymbol{\theta_1}}^{\boldsymbol{2}\,\prime}\\ \boldsymbol{\sigma}_{\boldsymbol{\theta_1}}^{\boldsymbol{2}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta_1}}(x_0)\end{pmatrix}\right) \tag{A.8}$$

$$\begin{aligned}\mathrm{E}[\theta_i|\boldsymbol{\theta}_{i,1}] &= \bar{\theta}_1 + \boldsymbol{\sigma}_{\boldsymbol{\theta_1}}^{\boldsymbol{2}\,\prime}\boldsymbol{\Sigma}_{\boldsymbol{\theta_1}}{}^{-1}(x_0)\left(\boldsymbol{\theta}_{i,1} - \bar{\boldsymbol{\theta}}_1\right)\\ &= \bar{\theta}_1 + \begin{pmatrix}\sigma_{\theta_1}^2 & \sigma_{\theta_0}^2\end{pmatrix}\begin{pmatrix}\sigma_{\theta_1}^2 + s_\theta^2 & \sigma_{\theta_0}^2\\ \sigma_{\theta_0}^2 & \sigma_{\theta_0}^2 + \gamma_0^2(x_0)\sigma_\mu^2\end{pmatrix}^{-1}\left(\begin{pmatrix}\theta_{i,1}\\ \hat{\theta}_0\end{pmatrix} - \begin{pmatrix}\bar{\theta}_1\\ \bar{\theta}_0\end{pmatrix}\right)\end{aligned} \tag{A.9}$$

Given this expectation, one may repeat the proof in appendix A.1 to show that the first order condition for reputation is the same when we also condition on $\bar{a}_0$ (basically, $\mathrm{E}[v_i|a_{i,1} - \bar{a}_1, \bar{a}_0] = \mathrm{E}[v_i|a_{i,1} - \bar{a}_1)$ and while it does affect $\mathrm{E}[\mathrm{E}[v_i|a_{i,1} - \bar{a}_1, \bar{a}_0]|\theta_{i,1}, \mu_{i,1}, \bar{a}_0]$ it does not affect the first derivative wrt. $a_{i,t}$. This leads to equilibrium behaviour of

$$\begin{aligned}a_{i,1} &= v_{i,1} + \left(\bar{\theta}_1 + \boldsymbol{\sigma}_{\boldsymbol{\theta_1}}^{\boldsymbol{2}\,\prime}\boldsymbol{\Sigma}_{\boldsymbol{\theta_1}}{}^{-1}(x_1)\left(\boldsymbol{\theta}_{i,1} - \bar{\boldsymbol{\theta}}_1\right)\right) + x_1\mu_{i,1}\xi_1(x_1)\\ \bar{a}_1 &= \bar{v}_1 + \left(\bar{\theta}_1 + \boldsymbol{\sigma}_{\boldsymbol{\theta_1}}^{\boldsymbol{2}\,\prime}\boldsymbol{\Sigma}_{\boldsymbol{\theta_1}}{}^{-1}(x_1)\left(\boldsymbol{\theta}_1 - \bar{\boldsymbol{\theta}}_1\right)\right) + x_1\mu_1\xi_1(x_1)\end{aligned} \tag{A.10}$$

This is where the problems arise. To obtain a new signal $\hat{\theta}_1$ agents (or the principal) need to invert (A.10). But as $\boldsymbol{\sigma_{\theta_1}^2}'\boldsymbol{\Sigma_{\theta_1}}^{-1}$ has full rank and we have more columns than rows, the *left inverse* does not exist (Mitra 1980, p. 602), which means we cannot simply isolate θ_1. There are 2 unknowns (θ_1 and θ_0, or more precisely θ_0 and η_1), too many unknowns for a single equation.

There will only become more unknowns as t increases, and just the single equation each time. One may insert expectations for previous θ_t to obtain an unbiased signal of current θ_t, but as all signals would then feature iteratively going forward, it would greatly complicate the covariance structure between signals. All in all, I did not consider including this element operationally useful, as the paper mainly focuses on allowing the principal's knowledge to develop over time.

A.3. Principal's expectation structure (proof of proposition 4.2)

Consider that we may write $\hat{\theta}_t = \theta_t + \left(\frac{x_t\xi_t(x_t)}{\rho_t}\right)(\mu_t - \bar{\mu}) = \theta_t + \hat{\epsilon}_t$. Since μ_t is drawn independently from a normal distribution, it follows that $\theta_t \perp\!\!\!\perp \hat{\epsilon}_t$ and $\hat{\epsilon}_t \perp\!\!\!\perp \hat{\epsilon}_k$ for $k \neq j$. This establishes that at its core, the vector is drawn from independent normal distributions. Define now $\hat{\boldsymbol{\epsilon}}_t = (\hat{\epsilon}_0, \hat{\epsilon}_1, ...\hat{\epsilon}_t)'$ and $\boldsymbol{\theta}_t$ as a $t \times 1$ matrix of θ_t's. Consider now that we may write

$$\begin{pmatrix} \theta_t \\ \hat{\boldsymbol{\theta}}_t \end{pmatrix} = \begin{pmatrix} \theta_t \\ \boldsymbol{\theta_t} + \boldsymbol{\epsilon}_t \end{pmatrix} = \begin{pmatrix} \sigma_{\theta_t} & 0 & \cdots & 0 \\ \sigma_{\theta_0} & \left(\frac{x_0\xi_0(x_0)}{\rho_t}\right)\sigma_\mu & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{\theta_t} & 0 & \cdots & \left(\frac{x_t\xi_t(x_t)}{\rho_t}\right)\sigma_\mu \end{pmatrix} W + \begin{pmatrix} \bar{\theta}_t \\ \bar{\boldsymbol{\theta}}_t \end{pmatrix} \tag{A.11}$$

where W is a vector of independent $N(0,1)$ variables. This is one of the more operational definitions of a multivariate normal distribution, and therefore establishes that the vector follow such a distribution (MIT 2008). The idea here is that every element may be written as a linear combination of independent standard normal distributions, which is then transformed to the actual distributions by multiplying the standard deviation and adding the mean. For each, a similar first column is included to provide θ_t, and then independent errors are added to each $\hat{\theta}_t$.

Given a multivariate n-dimensional variable $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, one may partition this into two subvectors, and compute the conditional mean of one subvector on the other:

$$\begin{pmatrix} \overbrace{\boldsymbol{x}_1}^{q\times 1} \\ \underbrace{\boldsymbol{x}_2}_{(n-q)\times 1} \end{pmatrix} \sim N\left(\begin{pmatrix} \overbrace{\boldsymbol{\mu}_1}^{q\times 1} \\ \underbrace{\boldsymbol{\mu}_2}_{(n-q)\times 1} \end{pmatrix}, \begin{pmatrix} \overbrace{\boldsymbol{\Sigma}_{11}}^{q\times q} & \overbrace{\boldsymbol{\Sigma}_{12}}^{q\times(n-q)} \\ \underbrace{\boldsymbol{\Sigma}_{21}}_{(n-q)\times q} & \underbrace{\boldsymbol{\Sigma}_{22}}_{(n-q)\times(n-q)} \end{pmatrix}\right) \tag{A.12}$$

$$\mathrm{E}[\boldsymbol{x}_1|\boldsymbol{x}_2] = \boldsymbol{\mu_1} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x_2} - \boldsymbol{\mu_2})$$

Wang (2006) provides a proof of this result. Applying the above to the distribution yields the expectation stated in the lemma. We then need to determine the elements in the matrices.

The i'th element of $\boldsymbol{\sigma_t}^2$ will be the covariance between θ_t and $\hat{\theta}_i$, where we always have $i \leq t$. We may compute here that $\mathrm{Cov}(\theta_t, \hat{\theta}_i) = \mathrm{Cov}(\theta_i + \sum_{k=i+1}^{t}\eta_k, \theta_i + \hat{\epsilon}_i) = \mathrm{V}(\theta_i, \theta_i) = \sigma_{\theta_i}^2$, so:

$$\boldsymbol{\sigma_t}^2 = \left(\sigma_{\theta_0}^2, \sigma_{\theta_1}^2, ..., \sigma_{\theta_t}^2\right)' \tag{A.13}$$

$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_t}(x_t)$ is the covariance matrix of $\hat{\theta}_t$. It follows that the diagonal elements will be the variance of the corresponding θ_t, defined in (4.6) as $\sigma^2_{\theta_t} + \gamma^2_t(x_t)\sigma^2_\mu$.

The non-diagonal elements will then show the covariance between various $\hat{\theta}_t$. We may compute $\text{Cov}(\hat{\theta}_i, \hat{\theta}_j) = \text{Cov}(\theta_i + \hat{\epsilon}_i, \theta_j + \hat{\epsilon}_j) = \text{Cov}(\theta_i, \theta_j) = \min[\text{V}(\theta_i), \text{V}(\theta_j)] = \min[\sigma^2_{\theta_i}, \sigma^2_{\theta_j}]$. This reflects that they will have the terms of θ_i in common until we reach the lowest of i and j. Since the variance is increasing, we can just put the minimum function outside. So, for $i < j$ it is $\sigma^2_{\theta_i}$.. In the covariance matrix below, it means that elements left of the diagonal will be follow the sequence $(\sigma^2_{\theta_0}, \sigma^2_{\theta_1}, ...\sigma^2_{\theta_{t-1}}$, while to the right they will simply be $\sigma^2_{\theta_t}$, so that:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_t}(x_t) = \begin{pmatrix} \sigma^2_{\theta_0} + \gamma^2_0(x_0)\sigma^2_\mu & \sigma^2_{\theta_0} & \cdots & \sigma^2_{\theta_0} & \sigma^2_{\theta_0} & \sigma^2_{\theta_0} \\ \sigma^2_{\theta_0} & \sigma^2_{\theta_1} + \gamma^2_1(x_1)\sigma^2_\mu & \cdots & \sigma^2_{\theta_1} & \sigma^2_{\theta_1} & \sigma^2_{\theta_1} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \sigma^2_{\theta_0} & \sigma^2_{\theta_1} & \cdots & \sigma^2_{\theta_{t-2}} & \sigma^2_{\theta_{t-1}} + \gamma^2_{t-1}(x_{t-1})\sigma^2_\mu & \sigma^2_{\theta_{t-1}} \\ \sigma^2_{\theta_0} & \sigma^2_{\theta_1} & \cdots & \sigma^2_{\theta_{t-2}} & \sigma^2_{\theta_{t-1}} & \sigma^2_{\theta_t} + \gamma^2_t(x_t)\sigma^2_\mu \end{pmatrix} \tag{A.14}$$

$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_t}(x_t)$ will be a $(t+1) \times (t+1)$ matrix in a period t. Inverting it is therefore somewhat of a hassle for $t \geq 2$, where the matrix becomes 3×3. I only handle these cases in the simulated games using MATLAB.

B. Appendix: MATLAB code (available at GitHub)

All code used in the paper is available at GitHub: `https://github.com/rbjoern/Public_economics/tree/master/Code`. The following files are provided:

1. **EoP_Matlab_v2.m**. This is a file most suited to work in, which provides a script which simulates the game. This is also the place to look for errors, as other files mirror this one.
2. **EoP.m**. A class file, which mirrors the working file into a function, so it may more easily be run with different parameter values. Necessary for the files below.
3. **Game_Default**. A file which runs the function above. No real gimmicks. It might be a good place to start if one wants to run some different parameter values.
4. **Game_VaryingPrivacy**. A file which loops over the simulating function twice for different values of x_t and provides figure 5.1. These files are suited for recreating figures.
5. **Game_NoTrend**. A file which loops over the simulating function twice for different values of x_t and provides figure 5.2 and C.1 (if one adjusts input values).
6. **Game_BigShock**. A file which loops over the simulating function twice for different values of x_t and provides figure 5.3 and C.2 (if one adjusts input values).
7. **Game_IntroFigure**. A file which loops over the simulating function twice for different values of x_t and provides figure 1.1.
8. **Game_Withandwithoutrend**. A file which loops over the simulating function twice with and without a trend and provides 4.1.

C. Appendix: Extra figures

Figure C.1: *Permanent societal values (with a longer time frame)*

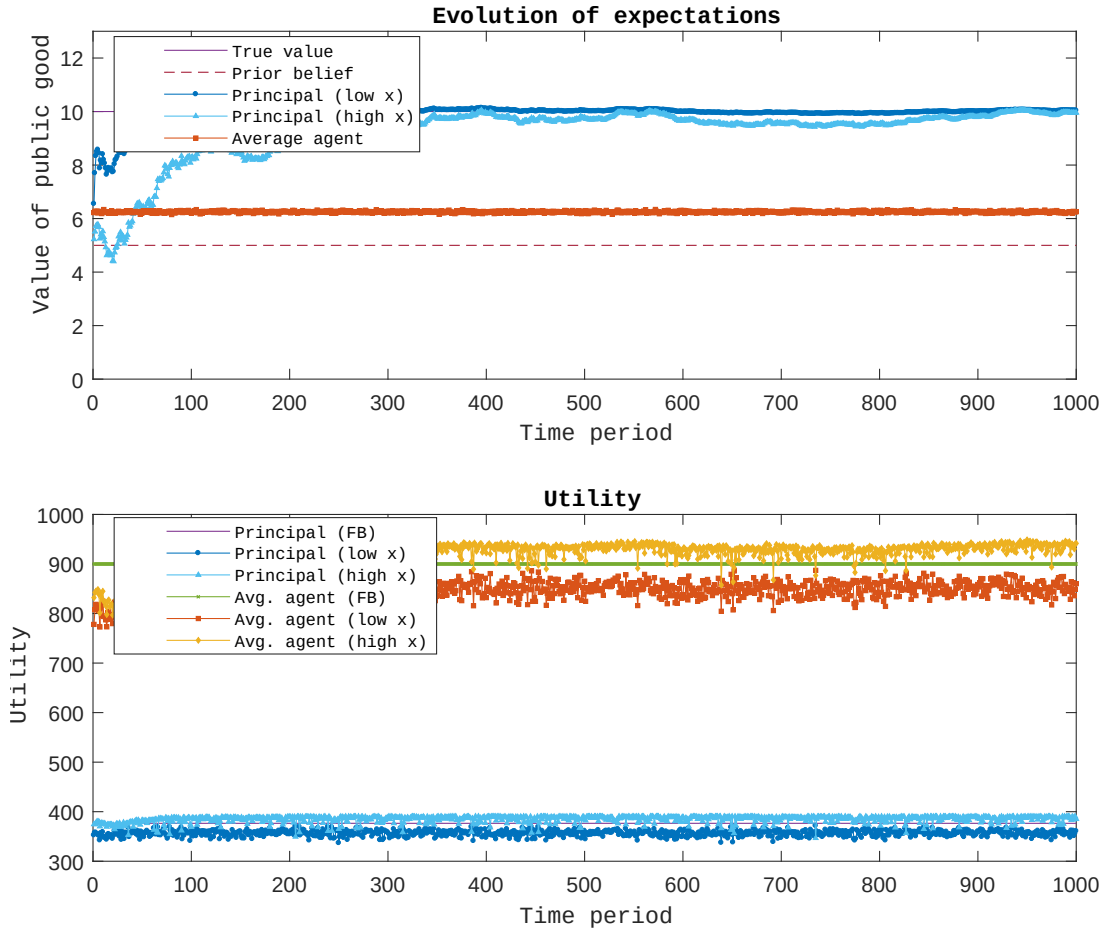


Figure C.2: *Suddenly, a public 'bad' becomes a public good (with a higher 'high' x)*