Hallman, Alice; Spiro, Daniel

Working Paper

A Theory of Hypocrisy

CESifo Working Paper, No. 9734

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Hallman, Alice; Spiro, Daniel (2022) : A Theory of Hypocrisy, CESifo Working Paper, No. 9734, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/260864

CESifo WORKING PAPERS

9734
2022

April 2022

A Theory of Hypocrisy

Alice Hallman, Daniel Spiro

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9734*

A Theory of Hypocrisy

Abstract

This paper explains the occurrence of hypocrisy – when the by-society most despised types pretend to be the most revered types. Real-world phenomena include pedophile priests, sex-offender feminists and seemingly very busy dispensable office workers. Building on the signaling framework of Bernheim (1994) – where payoffs consist of an intrinsic cost of falsifying yourself, and a concern for social esteem – we show conditions for emergence of hypocrisy in equilibrium. In such equilibria the most despised types along with the most revered types behave normatively, others do not. Thus, in equilibrium there are "rumors" about those acting the most normatively – society infers that they are either truly normative or despised, but one cannot know who is who. This is to be distinguished from "conformity" – where the most normative and almost-normative types fully follow a social norm. Whether conformity or hypocrisy will arise in equilibrium depends on the cost of falsification, and the number of hypocrites depends on the weight of social esteem. Our theory thus shows how cultural parameters map into equilibrium culture.

JEL-Codes: D700, D910, Z100.

Keywords: social esteem, hypocrisy, conformity, social norm.

Alice Hallman
Department of Economics
Uppsala University / Sweden
alice.hallman@nek.uu.se

Daniel Spiro
Department of Economics
Uppsala University / Sweden
daniel.spiro.ec@gmail.com

We are grateful for comments from Ola Andersson, Gustav Karreskog, Torben Mideksa, Alan Sola and seminar participants at Uppsala University. The authors gratefully acknowledge funding from Handelsbanken Research Foundations grant number P18-0142.

1 Introduction

You love to look earnest and inform the world that it's the ' duty of responsible business men to be strictly moral as an example to the community.' In fact you're so earnest about morality, old Georgie, that I hate to think how essentially immoral you must be underneath.

Sinclair Lewis (Babbitt, p. 69, 1922)

The purpose of this paper is to explain the existence of hypocrisy – "a situation in which someone pretends to believe something (...) that is the opposite of what they do or say at another time" *(Cambridge Dictionary)*. Lewis, in our introductory quote, describes the key ingredient of our proposed equilibrium: the more normative a person makes an effort to appear, the less normative she might be in private. Casual examples include: pedophiles who invest in education and public service to get into the clergy; rapists who are in the career of law enforcement;[1] and dispensable office workers who go out of their way to seem busy. There is also experimental evidence of hypocrisy. In so-called 'lying games', people lie to the maximum once they do lie.[2] Psychological experiments show that, in societies where homophobia is a norm, disproportionately many of those claiming to be homophobic have homosexual tendencies (Weinstein et al, 2012; Adams et al, 1996; D'Augelli, 2006).[3] These observations constitute a form of hypocrisy by the definition above. Similar evidence concern women's stated sexual preferences and actual sexual arousal (Morokoff, 1985).

From these examples it is clear that the issue of hypocrisy applies when society cares about something people say, do or feel *in private*. We will refer to this as a person's type. But since the private type is unobservable, society infers it from the person's public action. This suggests that, in order to understand hypocrisy as an equilibrium phenomenon, we need a

[1] In 2010, the former Swedish chief of police and then advisor on gender equality and sexual harassment to the National Police Directorate, was arrested and charged with multiple sex offenses, including the rape of a 14-year-old child. He had been fighting for women's rights for years and had gotten the nickname 'captain skirt,' as he was described to be obsessed with women's issues. He was then sentenced to prison due to being convicted on multiple charges of rape and sex offenses. Catholic Church sexual abuse cases have, in the 20th and 21st centuries involved many allegations, investigations, trials, convictions, and revelations about decades of attempts by Church officials to cover up reported incidents. Recently, Catholic priests in Argentina were sentenced to 45 years for child abuse (Guardian, 2019).

[2] In the experiments subjects throw a die (Kajackaite and Gneezy, 2015; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018; Khalmetski and Sliwka, 2019). Knowing that the experimenter cannot know their actual result and that they get a certain amount per dice eye, they report an outcome of the throw.

[3] Casual evidence of conservative politicians who are found visiting gay establishements are available to anyone surfing the web.

signaling model. Note also from these examples that the puzzle of hypocrisy is not about situations where the whole population claims to be normative. Such situations are better described as *conformity* as per the seminal "Theory of conformity" by Bernheim (1994). Likewise, hypocrisy is not when near-normative types behave fully normatively. Also this is easily explained by the signalling theory of Bernheim (1994). Instead, hypocrisy is a situation where the *least* normative types ("anormative") signal that they are fully normative, while others do not do so. An explaination of hypocrisy thus has to answer three subquestions:

A. Why do the most anormative types behave normatively?

B. Why do not all others behave normatively?

C. If the anormative types behave normatively, why aren't they fully revealed as such?

The purpose of the paper is to answer these questions as an equilibrium outcome. Since we are interested in situations where the anormative types, but not all others, behave normatively we need the theory to distinguish between different levels of normativeness. Hence, we will use the framework of Bernheim (1994) that has exactly this feature. But we reverse his result so that, instead of the nearly normative behaving normatively, the anormative behave normatively. That is, we provide a theory of hypocrisy.

In the theory, like in many other theories of social norms (e.g., Kuran 1989; Lindbeck et al. 1999; Michaeli and Spiro 2015; Nyborg and Rege 2003) the payoff of an individual consists of two parts: the intrinsic utility which is decreasing the more one misrepresents onself; and the social esteem which is decreasing if a person's type is inferred to be distant from the norm. As it turns out, hypocrisy is a natural equilibrium once one departs from a particular functional-form assumption – that the intrinsic utility is concave – which lies behind the conformity result of Bernheim (1994). If the intrinsic utility is convex, then hypocrisy arises in equilibrium instead. Hence, our paper provides a prediction for when conformity and when hypocrisy will arise in societies. We discuss this further in the conclusions.

In order to show our results we stick closely to the framework of Bernheim (1994). For tractability we depart from it by assuming a simpler distributions of types, a simpler social-esteem function and that actions are taken on a grid instead of a continuum. These simplifications do not drive our result. What does drive our result vis-a-vis Bernheim (1994)

is that the curvature of intrinsic preferences is convex instead of concave. As such our paper is closely related to a series of research papers showing that the structure of preferences has important implications for outcomes (Osborne, 1995; Eguia, 2013; Kamada and Kojima, 2014; Michaeli and Spiro, 2015, 2017; Chen et al., 2019; Michaeli, 2020). In comparison to these papers, ours is the first to explain why hypocrisy arises and why hypocrites get away with it.

2 The model

There is a continuum of players, normalized to a unit mass. A player has a privately known type $t \in T$. For tractability we depart from Bernheim (1994) and define the set of types, T, as the discrete type space of $k+1$ equidistant points on $[0,1]$ s.t. $T = \{0, \frac{1}{k}, \frac{2}{k}, .., \frac{k-1}{k}, 1\}$. Types are distributed according to a discrete uniform distribution over T. There is then a mass $\frac{1}{k+1}$ of each type $t \in T$. The distribution of t is common knowledge. Each player selects some publicly observable action $x \in X$. We normalize X to T. Social esteem depends on types and not on actions. As types are unobservable, however, actions matter indirectly by being indicative of the type. Therefore, this model is a signaling game.

Before we go into the details about the types and describe how social esteem is rewarded, we describe the payoff of a player of type t choosing action x as the sum of two functions: the intrinsic value of action, $g(t-x)$; and how social esteem is rewarded to those who take action x, $s(x)$.

$$u(x,t) = g(t-x) + \lambda s(x) \tag{1}$$

As in Bernheim (1994), $g(.)$ is decreasing in $|t-x|$ and is symmetric around 0, $g(z) = g(-z)$, and $g(.)$ is continuous and twice differentiable in $x \neq t$. This captures that misrepresenting oneself is costly. The parameter $\lambda > 0$ determines the payoff-weight on social esteem, and will play a role in determining equilibrium. We will further describe how social esteem, $s(x)$ is determined from equilibrium actions below.

Except for differing in their bliss points, t, all players are identical. Let $t = 0$ be the most reverred type, from now on called the norm, and the type $t = 1$ the most despised type.

The esteem society holds for each type t is $-t$.[4] If types were observable, we would have that $s(t) = -t$. Since they are not, the type is inferred from actions. Social esteem of taking an action x is rewarded according to which types are believed to take action x. The esteem function s thus describes how types t are inferred from actions x, and how each (perceived) type is treated. All face the same information and form the same inference. Let $\phi(b, x)$ be the inference correspondence. It describes the posterior over all types, given the prior $t \sim \mathcal{U}\{0, 1\}$, i.e., after having observed an action x, the correspondence $\phi(b, x)$ describes for each point $b \in T$ the conditional probability that a type taking action x is of type b. Inferences are rational so that $\sum_{b \in T} \phi(b, x) = 1$ and updated using Bayes rule whenever possible. For example, if only type t' choose action x', then

$$\phi(b, x') = \begin{cases} 1, & \text{if } b = t' \\ 0, & \text{otherwise.} \end{cases}$$

The expected social esteem of an action is described by the following function,

$$s(x) = \sum_{b \in T} -b\phi(b, x). \tag{2}$$

Note that the inference function ϕ, though endogenous, is taken as given by each agent. It describes a v.N.M-type of payoff to actions, as in Bernheim (1994).

We use Bayesian Nash equilibrium as our equilibrium concept and Banks and Sobel's (1987) Divinity Criterion (also known as D1) as an equilibrium refinement to specify beliefs over non-equilibrium actions. Loosely speaking, D1 instructs to use the following protocol when assigning off-equilibrium beliefs. Suppose the receiver's best response is to take action $a(x, t)$ if knowing an off-equilibrium action x was taken by type t. To get D1 beliefs, one first collects the $a(x, t)$ for all the different types t into the set A. Next, one counts how many of the $a \in A$ that would yield a payoff increase for a type t' if t' took action x instead of her equilibrium action $x^*(t')$. D1 then instructs, that upon observing x, to believe it was taken

[4]This is a more restrictive assumption than Bernheim (1994) which we have chosen for tractability.

by the t' with the largest such count.[5] In applying D1 to our setting we follow Bernheim (1994). In our setting the equivalent of $a \in A$ is simply $-t \in -T$. Since all players value esteem in the same way, the "counting" boils down to how many of the possible t that imply $g(x - t') - \lambda t > g(x^*(t') - t') + \lambda s(x^*(t'))$ for a given t' and then identifying the type $t' \in T$ with the largest such count. Call this $\tilde{t}$, inferences are then given by

$$\phi(t', x) = \begin{cases} 1, & \text{if } t' = \tilde{t} \\ 0, & \text{otherwise.} \end{cases}$$

We denote by x a particular action $x \in X$, and by the correspondence $x(t)$ the actions taken by some $t \in T$. Hence, we write $x(t') = x'$ if the type t' only take the action x', and $x, x', x'' \in x(t'')$ if some share of the type t'' take each of the actions x, x' and x''.

3 Analysis

We start by replicating the conformity result of Bernheim (1994). He assumes that the cost of misrepresentation is concave. Then actions are monotone in types.

LEMMA 1 *Suppose $g(.)$ is strictly concave. There exists a λ^* such that if $\lambda \leq \lambda^*$ in the unique equilibrium all types separate at $x(t) = t$. If $\lambda > \lambda^*$ in the unique equilibrium*

1) $x(t)$ is weakly increasing in t;

2) there exists a $\bar{t} \leq 1$ such that $x(t) = 0$ for all $t \leq \bar{t}$;

3) all types $t > \bar{t}$ take actions spanning a range $x(t) \in [\bar{x}, 1]$ for some $\bar{x} > 0$;

4) beliefs for an off-equilibrium action x' is $\phi(b, x') = \begin{cases} 1, & \text{if } b = \bar{t} \\ 0, & \text{otherwise.} \end{cases}$

The main purpose of stating this result is to verify that Bernheim's conformity result holds in our setting too whenever his functional form assumption of a concave g holds. The proof, which can be found in Appendix A.1, is a straightforward translation of Bernheim (1994) to a setting of discrete actions. Naturally, if λ is small all types separate, as the

[5] If more than one type of sender could benefit from a deviation to an off-equilibrium message, the weaker Intuitive Criterion (Cho and Kreps, 1987) assumes that the receiver's belief assigns the same weight to all potential deviators (see Munoz-Garcia and Espinola-Arredondo, 2011, for further comparisons.).

potential gain in esteem is not worth the effort of trying to pool with more normative types. The intuition behind Berheim's conformity result is that a concave g implies a small intrinsic loss when misrepresentation is small, but that large misrepresentation bears a very high intrinsic loss. This means that, for types close to the prevailing norm, it is profitable to pretend to be fully normative – there is pooling at the norm as per the second bullet. The concave g also implies that types far from the norm will need to incur a very high cost of misrepresentation if they pretend to be fully normative. This may happen if λ is very high, but if not then types very far from the norm take a non-normative action as per the third bullet. The properties of $x(t)$ for this range of types can take two forms (outlined in the proof). Either, if λ is somewhat larger than λ^*, then those who do not follow the norm speak their minds. Alternatively, if λ is even larger, then instead the non-normative types try to pool with somewhat less non-normative types. This makes all such types take actions somewhat closer to the norm but not at the norm itself.

We next move to showing equilibrium properties when g is convex. This means that the marginal intrinsic cost of small misrepresentation is large, and that of a large misrepresentation is small. We may then expect that once a player has decided to not be honest, she might as well move substantially from her intrinsic bliss point. On the other hand, if the social esteem function is flat enough, players' might not be willing to take on any cost. For technical reasons, we normalize the function space so that $-g(1) < 1$.[6]

We formalize the implications of a convex intrinsic cost of actions, $g(t-x)$, in a few helpful and instructive lemmas.

LEMMA 2 *Suppose $g(.)$ is strictly convex.*

1) If $u(x,t) \geq u(t,t)$ then $u(x,t') > u(t,t')$ for $x \neq t$ and $t' \neq t$.

2) Whenever an off equilibrium action x' is observed, D1 selects $t = x'$.

PROOF: See Appendix A.2. *Q.E.D.*

The techical proof can be found in the appendix, we here present the intuition. Part

[6]For proof of Proposition 1, part b, claim 2, we need that $g(1) < g'(\frac{k-1}{k})$. This condition approaches strict convexity as k goes to infinity, and requires that the point a that solves $g''(a) = 0$ is interior to $(0, \frac{k-1}{k})$, which can be achieved by a normalization of the function.

1 of Lemma 2 follows from that because the marginal cost for any misrepresentation is higher than for further misrepresentation, a type t requires a larger net increase in social esteem to misrepresent herself and choosing action x than does any other t' for choosing misrepresentative action t rather than that same action x. The first important implication of this lemma is that whenever we observe a pool at some action x_p, this pool must include $t = x_p$. A second important implication, that follows from the first, is that whenever we observe an off-equilibrium action, the belief is that it is due to an accidental truth-telling, i.e., that the action was taken by a type with that bliss point, part 2 of Lemma 2.

LEMMA 3 *Suppose $g(.)$ is strictly convex. In any equilibrium, $s(x)$ is monotonically decreasing.*

PROOF: See Appendix A.3. *Q.E.D.*

The intuition for this is that whenever actions are such that rational inferences would make $s(x)$ increasing, some will strictly prefer to change their action, up until the point when $s(x)$ is decreasing. Then, whenever social esteem $s(x)$ is strictly decreasing in actions x, any strategy $x > t$ is strictly dominated. It directly follows that:

LEMMA 4 *Suppose $g(.)$ is strictly convex. In any equilibrium $x(t) \leq t$ for all $t \in T$.*

PROOF: First, note that if $s(x)$ is decreasing, any strategy $x > t$ is strictly dominated. This follows directly from the symmetry of $g(.)$. Since $u(t-a,t) > u(t+a,t) \Leftrightarrow g(a)+\lambda s(t-a) > g(-a)+\lambda s(t+a) \Leftrightarrow s(t-a) > s(t+a)$. It then follows that if there exists some t' such that $x(t') \neq t'$, then $x(t') < t'$. Therefore, in any equilibrium, $x(t) \leq t$ for all $t \in T$. *Q.E.D.*

This result is intuitively obvious. If an agent deviates from her bliss point, it will be towards the norm.

LEMMA 5 *Suppose $g(.)$ is strictly convex.*

1) If $u(x',t') \geq u(x,t')$ for all $x \in (x',t')$, then $u(x',t) \geq u(x,t)$ for $t > t'$.

2) If $s(x)$ is weakly decreasing, then if $t' = argmax_x u(x,t')$ then $t = argmax_x u(x,t)$ for all $t < t'$.

3) If $u(x,t') = u(x',t')$, where $x < x'$, then $u(x,t) < u(x',t)$ for all $t \in [x',t')$.

PROOF: See Appendix A.4. *Q.E.D.*

Note that this game does not satisfy the single-crossing property. In particular, part 3 of Lemma 5 exhibits what Chen et al. (2020) call the *reverse single-crossing property.*

There are two direct implications of part 2 of Lemma 5. First, whenever $g(.)$ is convex, there can never be a conformity type equilibrium à la Bernheim (1994), where types closest to the norm choose the normative action while those furthest from the norm are sincere. Second, it also directly follows that whenever concern for social esteem is small, $\lambda \leq 1$, the unique (fully separating) equilibrium is for each player to be sincere, which will be shown in Proposition 1.

Consider now what happens as we increase how much people care about social esteem, $\lambda > 1$. On the one hand, as people care more about esteem, they should make more efforts to appear to be normative. On the other hand, understanding this, 'good deeds' become less informative, and therefore the esteem rewarded to anyone who behaves normatively will be lower. Interestingly, even as λ goes to infinity, there cannot be an equilibrium where all players choose the norm action. This is since then all would get the same esteem, implying that some lower-than-average types would deviate to honesty. More generally,

LEMMA 6 *Under the Divinity criterion there is no fully pooling equilibrium.*

PROOF: Suppose there exists a fully pooling equilibrium. Then, $x(t) = x_p$ for some $x_p \in X$ and all $t \in T$. Then, $s(x_p) = -\mathbb{E}[t]$. By Lemma 2, the esteem for an off-equilibrium action t is $s(t) = -t$ hence for any $t < \mathbb{E}[t]$, $s(t) > s(x_p)$ each $t < \mathbb{E}[t]$ would gain from deviating to $x(t) = t$. *Q.E.D.*

4 Equilibrium

Aided by the previous lemmas we now arrive at our main result. Here we define "the norm" as the action that would only be taken by the best types if types were observable, i.e. $x = 0$. We define "anormative" as the least normative types ($t = 1$).

PROPOSITION 1 *Suppose $g(.)$ is strictly convex. If $\lambda < 1$ then the unique equilibrium is $x(t) = t$, for all $t \in T$. For any $\lambda > 1$ there exists a unique Bayesian Nash equilibrium. In this equilibrium:*

A) Some anormative take the norm action in public, $0 \in x(1)$,

B) No other type take the norm action in public, $0 \notin x(t), t \in (0, 1)$, and

C) The normative take the norm action in public, $x(0) = 0$.

PROOF: See Appendix B. *Q.E.D.*

The proposition establishes hypocrisy as an equilibrium and thus provides answers to the three parts of our research question.[7] (A) In equilibrium the anormative pretend to be fully normative. The reason for this is that the anormative get very low esteem if behaving honestly, and thus have much esteem to gain by pretending to be more normative.[8] The convexity of $g(.)$ implies that once an anormative person deviates from her blisspoint, the additional cost of a large misrepresentation is small. This makes anormative pretend to be fully normative. This has the consequence of lowering the social esteem of behaving normatively – society infers that when observing a normative statement it must have been taken by either a truly normative person or a very anormative person.

This is also what lies behind point (B), that no other types pretend to be fully normative. The lower esteem of acting normatively makes others less interested in taking that stance. More precisely, for types who are far from normative yet not fully anormative, lowering of the esteem at the norm means that for them the additional esteem of taking the action 0 compared to the action $1/k$ is lower than the additional intrinsic cost of doing so.[9] This means that such types will mispresent themselves, but not as much as the anormative.

[7]When $\lambda = 1$,a zero-mass of the most anormative are indifferent between $x = 1$ and $x = 0$. Hence a continuous number of equilibria exists with any zero-mass of $t = 1$ choosing $x = 1$ while the rest choose $x(t) = t$.

[8]Note that for any finite number of types and actions, $k \in (1, \infty)$, there always exists some mass of hypocrites $p > 0$ s.t. $s(x)$ is strictly decreasing, which is a necessary condition for existence of a hypocrisy equilibrium. Proof in appendix B.4.

[9]This is despite the convexity of $g(.)$. In equilibrium, the anormative types spread themselves over the action 0 and 1/k (and in some instances 2/k etc). The mass of anormative at each of this is precisely such that they are indifferent between 0 and 1/k, i.e. the slope of the esteem function between these points is the same as the slope of $g(0, 1) - g(1/k, 1)$ for the anormative type. Now, since other types are closer to 0 than the anormative are, $g(0, t) - g(1/k, t)$ is smaller for $t < 1$ making no other such types go all the way to the norm.

For types who are close to the norm, the convexity of $g(.)$ makes any small misrepresentation not worthwhile as it incurs a high cost. Put together we get, as visualized in Figure 1 (upper left panel), that types far from the norm misrepresent themselves by taking stances close to the norm and they do so in descending order – the least normative pretend to be the most normative, the second to least normative pretend to be almost normative and so on. $x(t)$ is thus a decreasing function for the range of such types. Types close to the norm, on the other hand, behave honestly and $x(t)$ is an increasing function for the range such types.

Finally, part (C) establishes that the truly normative also take the normative action ($x(0) = 0$). The question is why the truly normative do not deviate to avoid being confused with the anormative. The reason relies on two observations. The first is that in equilibrium $x(0)$ is the action that maximizes esteem as not all anormative choose that action. The second is that a small deviation is very costly since $g(.)$ is convex. Hence, any "escape" from 0 by the normative would be very costly. The fact that the truly normative behave normatively is the reason why the anormative who behave normatively are found out as hypocrites.

Having established hypocrisy as an equilibrium we now turn to highlighting a few further insights. The proposition shows that (given a convex $g(.)$) a necessary and sufficient condition for hypocrisy is that $\lambda > 1$. That is, agents need to put a sufficiently high weight on social esteem as otherwise it would not be worthwhile for the anormative to pretend.

COROLLARY 1 *The share of $t = 1$ who choose $x(1) = 0$ is increasing in λ.*

This follows directly from that at higher λ, more of the anormative will misrepresent themselves by spreading out over the normative and almost normative actions, $x(1) = \{0, \frac{1}{k}, ..\}$. By Lemma 5 part 3, if $min\{x(\frac{k-j}{k})\} < \frac{k-j}{k}$ then $x(\frac{k-j}{k}) \geqq \max\{x(\frac{k-j+1}{k})\}$, where j is increasing in λ and determined by the convexity of $g(.)$.

This corollary further highlights the importance of λ in creating hypocrisy. The higher is the weight on social esteem, the more "hypocrites" there will be in equilibrium.

The next two corollaries highlight the cultural differences between honest societies, conformity societies (here defined as the equilibrium in Lemma 1) and hypocrisy societies (here defined as the equilibrium in Proposition 1).

COROLLARY 2 *If λ is sufficiently small, the unique equilibrium is Honesty, $x(t) = t$ for all $t \in T$. If not, and $g(.)$ is strictly concave, the unique equilibrium is one of Conformity; and if $g(.)$ is strictly convex, the unique equilibrium is one of hypocrisy.*

PROOF: Follows directly from Lemma 1 and Proposition 1. *Q.E.D.*

COROLLARY 3 *In a hypocrisy equilibrium, an off-equilibrium action x is believed to be taken by $t = x$. In a Conformity equilibrium, an off-equilibrium action is believed to be taken by the least normative type in the pool.*

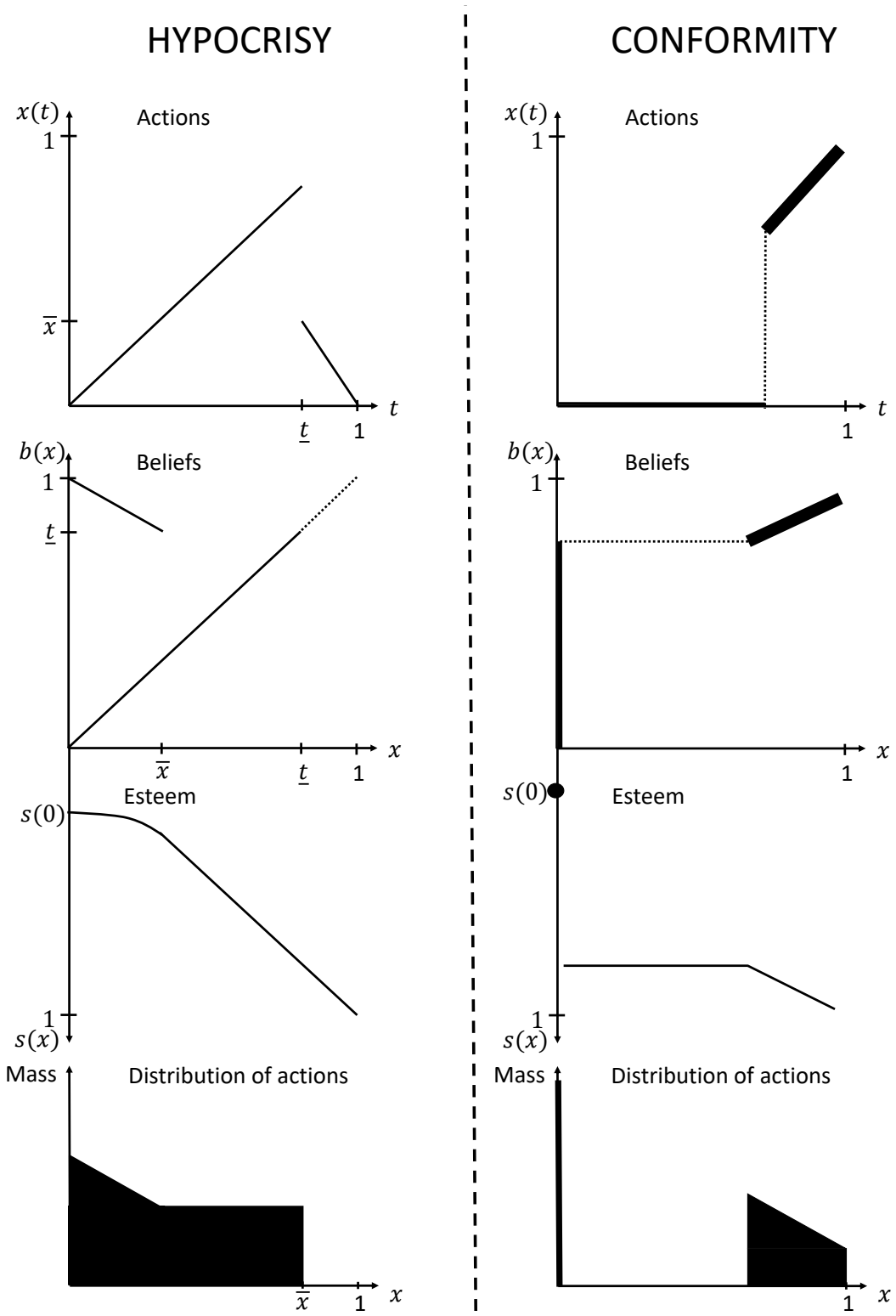
PROOF: Follows directly from Lemma 2 and Lemma 1.[10] *Q.E.D.*

A first thing to note is that, when the importance of social esteem is low, all societies will look similar – honesty is the only equilibrium independently of the intrinsic cost function. Apart from this similarity, the two corollaries highlight broad cultural differences between conformity-type and hypocrisy-type societies. The differences are illustrated in Figure 1 where conformity-type societes are on the right and hypocrisy-type societies are on the left. First (row-1 panels of the figure) the actions in conformity-type societies are monotone in type and all types move towards the norm while in hypocrisy-type societies actions are non-monotonic in type and only types far from the norm move towards it.

This implies, second (row-2 panels of the figure) that the societies will differ in terms of the observed actions. In conformity-type societies, the distribution of actions will be bimodal with a sharp peak at the norm itself and another more compressed mode at a distance to the norm, with a decreasing mass at each action. In hypocrisy-type societes the distribution will look more smooth, roughly like an exponential distribution with the peak at the norm (if the norm would have been in the center of the distribution of types, the distribution of actions would have looked roughly like a normal distribution).

[10] Note that there are no other off-equilibrium actions than in a neighborhood of the pool, see p. 856, Bernheim (1994).

FIGURE 1.— Outcomes in equilibrium



Notes: The left figures correspond a hypocrisy-type society. The right figures correspond a conformity-type society. Dotted line means off-equilibrium beliefs. In the lower right figure the peak at zero has been truncated for purposes of visibility.

Third (row 3) the societies differ also in terms of beliefs. In conformity-type societies,

beliefs of the observed normative action is that it is taken by the truly normative and the almost truly normative. In a hypocrisy-type society, upon observing a normative action, society infers that this must be either a truly normative type or a truly anormative type, but nothing in between. If the action is close but not precisely at the norm these two possibilities are attenuated – it is either somewhat normative or somewhat anormative types who take it.[11] Also when observing a mistaken action (an off-equilibrium action) the societies differ in what they believe. In a conformity-type society it is believed that this mistake is done by the least normative person who was meant to behave normatively. In a hypocrisy-type society the belief is that the mistake was done by the person whose type aligns with that action – that the person mistakenly spoke the truth – arguably a simpler belief than a conformity-type society.

Finally (row 4 of the figure) the societies differ in how esteem is rewarded according to actions. In conformity type societies there is a sharp decline in esteem when moving from the normative action to any adjacent action and after that the esteem is constant up until the second mode where esteem again falls. In hypocrisy-type societies, esteem is concave and smoothly falling as a function of the action.

5 An illustrative example

We here present a simple example to illustrate the main properties of the hypocrisy equilibrium. Suppose there are three types, $t \in \{0, \frac{1}{2}, 1\}$, i.e. $k = 2$. Hence, esteem rewarded to known types is

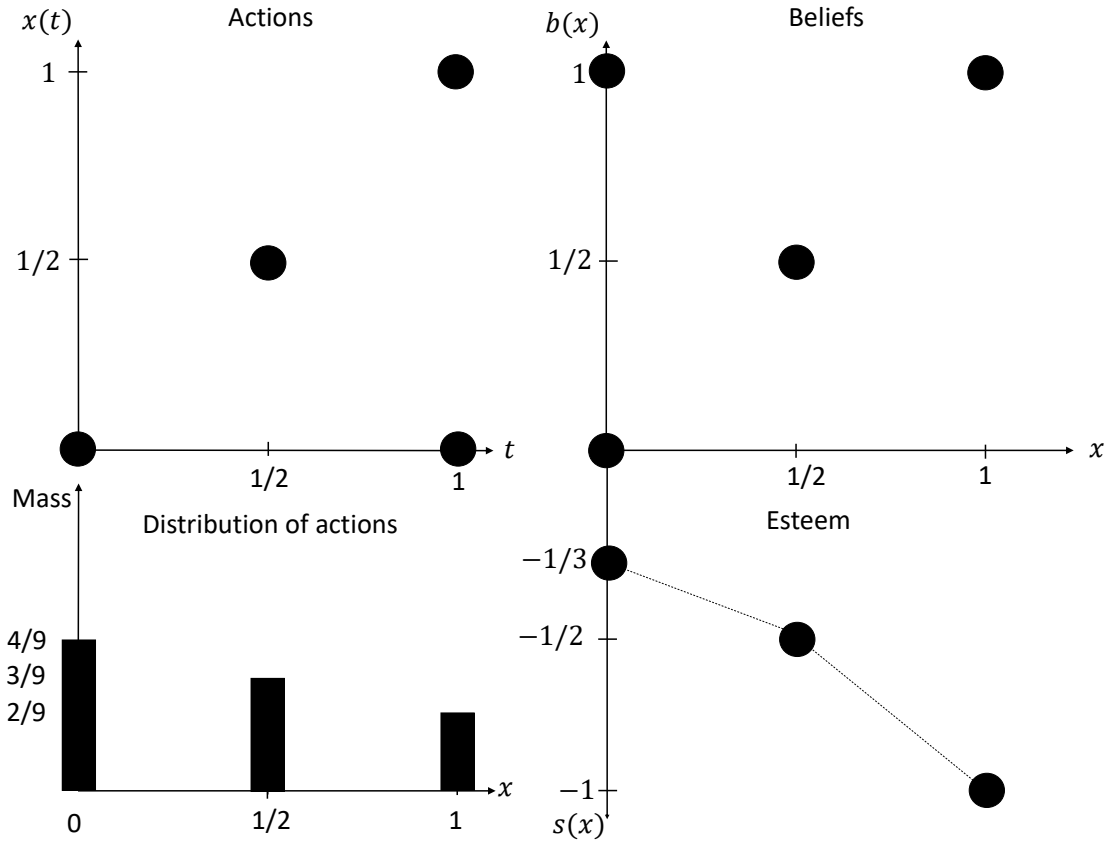
$$s(x) = \begin{cases} 0 & \text{if only } t = 0 \text{ takes action } x \\ 1/2 & \text{if only } t = 1/2 \text{ takes action } x \\ 1 & \text{if only } t = 1 \text{ takes action } x \end{cases}$$

[11] See the two lines converging up until the point $\bar{x}$.

and the available actions are $x \in \{0, \frac{1}{2}, 1\}$. Let $g = -\sqrt{|t-x|}$ and $\lambda = 4/3$. Hence, the payoff function is

$$u(x,t) = -\sqrt{|t-x|} + \frac{4}{3}s(x).$$

FIGURE 2.— Example equilibrium



Notes: The left figures correspond a hypocrisy-type society. The right figures correspond a conformity-type society. Dotted line means off-equilibrium beliefs. In the lower right figure the peak at zero has been truncated for purposes of visibility.

The equilibrium is illustrated in Figure 2. The upper left schedule shows the equilibrium actions $x(0) = 0$, $x(\frac{1}{2}) = \frac{1}{2}$ and $x(1) \in \{0, 1\}$. Most importantly, note that that some of the anormative types, $t = 1$, are Hypocrites as they take the action $x(1) = 0$. Beliefs (as displayed on the upper right) are thus that upon observing $x = 1$ it must be an anormative type; upon observing $x = 1/2$ it must be $t = 1/2$; and upon observing $x = 0$ it is either a truly normative person or an anormative person. To see how many hypocrites

there are in equilibrium, define by p the mass of type 1 that take $x(1) = 0$. This mass blends with the truly normative $t = 0$. The social esteem rewarded to action $x = 0$ depends on the conditional likelihood that the person taking $x = 0$ is truly normative or anormative hence $s(0) = \sum_{t \in T} -t \Pr(t|x = 0) = \frac{-p}{1+p}$. If p is too large, the esteem gain of behaving normtatively is too small for $t = 1$ to do so; and if p is too small the esteem gain of behaving normatively is very large implying all anormatives would want to do so. The equilibrium thus requires that the anormative are indifferent between hypocrisy and honesty i.e., $u(0, 1) = u(1, 1) \Leftrightarrow -1 - \frac{4}{3}\frac{p}{1+p} = -\frac{4}{3} \Leftrightarrow p = \frac{1}{3}$. The social equilibrium esteem function (displayed on the lower right) is then $s(X) = [-\frac{1}{3}, -\frac{1}{2}, -1]$. Clearly, esteem is decreasing in how anormative the action is, but it is also concave so that the marginal loss increases the more anormative the action is.[12]

The normative have, because of the hypocrites blending among them, lost out on esteem, $s(0) = -\frac{1}{4}$ but still hold high enough esteem to keep taking the norm action. Focusing on the middle types, $t = \frac{1}{2}$, they could earn $\frac{1}{4}$ in esteem by also choosing $x = 0$, but since g ic convex, such misrepresention albeit being small would cost them $\frac{1}{\sqrt{2}} > 1/4$. The distribution of actions (lower left) is decreasing in the deviation from the norm – the more anormative the action is the fewer take it, masking that the true distribution of types is in fact flat.

6 Concluding discussion

We have explained the emergence of hypocrisy as an equilibrium phenomenon, where the best people and the worst people take the best action, but one cannot distinguish the two. In the model individuals trade off the loss of social esteem when being perceived as non-normative with an intrinsic cost of misrepresentation. We show that hypocrisy arises when the loss of misrepresentation is convex – small falsification is very costly while large falsification is only marginally more costly. In this equilibrium anormative individuals pretend to be fully normative thus pool with the truly normative, while intermediately normative individuals reveal their true colors.

The intuition is that anormative types will lose much esteem if revealing themselves,

[12]The equivalent of liberal social sanctioning in the language of Michaeli and Spiro (2015).

they need to pretend to be someone else. When falsification loss is convex[13] they may as well pretend to be fully normative. This lowers the esteem of taking normative actions. The truly normative could potentially avoid being confused with the anormative types, by claiming to be less normative than they are, but since even such small falsification is very costly they choose not to.

Thus, in equilibrium there are "rumors" that those behaving normatively are either hypocrites or truly normative: in anti-gay environments homophobes are suspected of (but not known to) being gay, and in political settings the most eloquent politicians are suspected of being corrupt.

Hypocrisy is distinguished from conformity which is the equilibrium phenomenon in Bernheim (1994). Under Conformity the "almost normative" pretend to be fully normative while the anormative are revealed as such. Our theory emphasizes the role of culture and context for which of these equilibrium types that will arise – conformity arises in cultures and contexts where small pretence is costless while hypocrisy arises when small pretence entails a large cost and further pretence implies only a small additional loss. Thus, hypocrisy is driven by a feeling that if you "cannot be yourself" you may as well do whatever gives the highest esteem. We speculate that such feelings of falsification exists in contexts of moral conviction (such as whether to hide ones true religious beliefs) and deep identity (such as sexual preferences).

In equilibrium the number of hypocrites depends on the weight of social esteem – yet another cultural parameter that thus determines equilibrium culture. In societies or contexts where esteem is important, hypocrisy will be more prevalent.

[13]The equivalent of concave cost in the settings of Michaeli and Spiro (2015); Chen et al. (2019); Michaeli (2020).

References

Banks, J. S. and J. Sobel (1987). Equilibrium selection in signaling games. *Econometrica: Journal of the Econometric Society*, 647–661.

Bernheim, B. D. (1994). A theory of conformity. *Journal of political Economy 102*(5), 841–877.

Chen, C.-H., J. Ishida, and W. Suen (2020). Signaling under double-crossing preferences. *ISER Discussion Paper* (1103).

Chen, D. L., M. Michaeli, and D. Spiro (2019). Non-confrontational extremists. *TSE Working Paper*.

Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics 102*(2), 179–221.

Eguia, J. X. (2013). On the spatial representation of preference profiles. *Economic Theory 52*(1), 103–128.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise – an experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review 108*(2), 419–53.

Guardian (2019). Roman catholic priests in Argentina sentenced to 45 years for child abuse. https://www.theguardian.com/world/2019/nov/25/roman-catholic-priests-argentina-sentenced-45-years-child-abuse-school-deaf . Accessed: 2019-11-25.

Kajackaite, A. and U. Gneezy (2015). Lying costs and incentives. *UC San Diego Discussion Paper*.

Kamada, Y. and F. Kojima (2014). Voter preferences, polarization, and electoral policies. *American Economic Journal: Microeconomics 6*(4), 203–36.

Khalmetski, K. and D. Sliwka (2019). Disguising lies – image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics 11*(4), 79–110.

Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public choice 61*(1), 41–74.

Lindbeck, A., S. Nyberg, and J. W. Weibull (1999). Social norms and economic incentives in the welfare state. *The Quarterly Journal of Economics 114*(1), 1–35.

Michaeli, M. (2020). Grouping, in-group bias and the cost of cheating. *Games and Economic Behavior 121*, 90–107.

Michaeli, M. and D. Spiro (2015). Norm conformity across societies. *Journal of public economics 132*, 51–65.

Michaeli, M. and D. Spiro (2017). From peer pressure to biased norms. *American Economic Journal: Microeconomics 9*(1), 152–216.

Munoz-Garcia, F. and A. Espinola-Arredondo (2011). The intuitive and divinity criterion: Interpretation and step-by-step examples. *Journal of Industrial Organization Education 5*(1), 1–20.

Nyborg, K. and M. Rege (2003). On social norms: the evolution of considerate smoking behavior. *Journal*

of Economic Behavior & Organization 52(3), 323–340.

Osborne, M. J. (1995). Spatial models of political competition under plurality rule: A survey of some explanations of the number of candidates and the positions they take. *Canadian Journal of economics*, 261–301.

A Proofs

A.1 Proof of Lemma 1

We start by stating some useful results and then in the subsections prove the statements of Lemma 1.

LEMMA 7 *Suppose g is strictly concave. In any signaling equilibrium, if $t > t'$, then $x(t) \geq x(t')$.*

PROOF: The lemma is equivalent to Bernheim's (1994), Theorem 1. Since his proof relies only on the strict concavity of g and on inequalities in utilities, it applies directly as is to our restrictions on space. *Q.E.D.*

LEMMA 8 *If $g(.)$ is strictly concave then $\bar{t}$ is increasing in λ.*

PROOF: By Lemma 7, if $t < t'$ then $x(t) \leq x(t')$, hence whenever there exists some $t : x(t) = 0$ then there exists a largest type t' s.t. $0 \in x(t')$ (both t and t' might be 0). Let $\bar{t} \equiv sup(t : x(t) = 0)$. At the boundary, where a zero mass of $\bar{t}$ choose the action 0, by monotonicity in actions (Lemma 7), $s(0) = -\frac{(\bar{t}-\frac{1}{k})}{2}$. Then, type $\bar{t}$ does not want to deviate if $u(0,\bar{t}) \geq u(\bar{t},\bar{t}) \Leftrightarrow \lambda > \underline{\lambda}(\bar{t}) \equiv \frac{-g(\bar{t})}{s(0)-s(\bar{t})} = \frac{-2kg(\bar{t})}{k\bar{t}+1}$. We note that $\frac{\partial\underline{\lambda}(\bar{t})}{\partial\bar{t}} = \frac{\partial}{\partial\bar{t}}\frac{-2kg(\bar{t})}{k\bar{t}+1} = \frac{2}{\bar{t}^2}(g(\bar{t}) - \bar{t}g'(\bar{t}))$ and that we can write $g(\bar{t}) - \bar{t}g'(\bar{t}) = -g(0) + g(\bar{t}) + g'(\bar{t})(0 - \bar{t})$. As any strictly concave differentiable function f is bounded from above by its first-order Taylor approximation: $f(y) < f(x) + f'(x)[y - x]$, where $y \neq x$, hence for any $\bar{t} > 0$, $g(\bar{t}) - \bar{t}g'(\bar{t}) > 0$ and we can conclude that $\frac{\partial\underline{\lambda}(\bar{t})}{\bar{t}} > 0$. Note that $s(0)$ decreases smoothly with an increase of the share of types $\bar{t}$ that chooses action $x(\bar{t}) = 0$ and hence that $\underline{\lambda}(\bar{t})$ increases also within-type. Hence, as λ increases, a larger mass of types $\bar{t}$ chooses action 0, and if λ increases further, $\bar{t}$ increases. *Q.E.D.*

LEMMA 9 *There exists no separating equilibrium where $x(t) \neq t$ for some t.*

PROOF: Note that the action space has the same number of points as the type space. Hence, if any one type chooses an action $x(t) \neq t$, then in order to separate, types $t' = x(t)$ must choose some other action $x(t') \neq t'$. Furthermore, some other type $t'' \neq t$ has to choose $x(t'') = t$. There is no way to create such a chain of actions that agree with Lemma 7 of monotonicity in actions. *Q.E.D.*

Denote a pool by x_p and all types who choose this pool $t \in T(x_p)$. Let t_h be the largest type in the pool, and t_l the smallest type, s.t. $t_l < t_h$ and $t_l, t_h \in T(x_p)$. If all t_h choose the pool, let $\hat{x} \equiv min\{x(t_h + \frac{1}{k})\}$. If not all t_h choose the pool, let $\hat{x} \equiv min\{x(t_h) : x(t_h) > x_p\}$.

LEMMA 10 *When there exists an off-equilibrium action $x' \in (x_p, \hat{x})$ then D1 selects $\phi(t_h, x') = 1$ for all $x \in (x_p, \hat{x})$, where $t_h = sup(t : x(t) = x_p)$.*

PROOF: This proof is a straightforward replication of Bernheim's (1994) proof of Theorem 3, pages 869-870.

Let $U^*(t)$ denote the equilibrium payoff received by a type t agent, and let $H(t)$ denote the equilibrium esteem of a type t agent, where b denotes the inference from an action x over the type space T:

$H(t) \equiv \lambda \sum_T h(b)\phi(b, x(t))$

Finally, let $I(x,t) \equiv U^*(t) - g(t-x)$ denote the esteem that would make type t indifferent between choosing x and his or her equilibrium choice. We now establish two claims.

Claim 1. Consider any t', t'' with $t' < t''$ and any x, b such that $U(x, t'', b) \geq U^*(t'')$ and $U(x, t', b) \leq U^*(t')$. Then for any action $z > x$, the esteem that would make t' indifferent is larger then what would make t'' indifferent, $I(z, t') > I(z, t'')$.

To prove this claim, note that

$I(z,t') - I(z,t'') = [U^*(t') - g(t'-z)] - [U^*(t'') - g(t''-z)]$

$\geq [U(x,t',b) - g(t'-z)] - [U(x,t'',b) - g(t''-z)]$

$= g(t'-x) + \lambda s(x) - g(t'-z) - g(t''-x) - \lambda s(x) + g(t''-z)$

$= \int_{t'}^{t''} g'(w-z) - g'(w-x)dw$

$= -\int_x^z \int_{t'}^{t''} g''(w-q)dqdw > 0.$

Claim 2. Consider any t', t'' with $t' < t''$ and any x, b such that $U(x, t', b) \geq U^*(t')$ and $U(x, t'', b) \leq U^*(t'')$. Then for any $z < x$, $I(z, t'') > I(z, t')$.

The proof is completely symmetric to the proof of Claim 1.

We use Claim 1 and Claim 2 to prove Lemma 10. First, we note that as there are as many types as possible actions; whenever there exists an off-equilibrium action, there must exists some pool. Suppose that there exists (in equilibrium) some pool at x_p and let H_p be the status conferred on members of this pool. In equilibrium, type t_h recieves utility $U^*(t_h) = g(t_h - x_p) + H_p$. Let $\hat{x}$ be $x(t_h + \frac{1}{k})$. By Lemma 7, $\hat{x} \geq x_p$. As t_h is the largest type in the pool, $\hat{x} > x_p$. Now we consider off-equilibrium actions $x \in (x_p, \hat{x})$.

Consider any $t < t_h$. Take $x = x_p$ and b s.t. $\lambda h(b) = H_p$, and apply claim 1. It follows that, for $x \in (x_p, \hat{x})$, $I(x,t) > I(x,t_h)$. Under the D1 criterion, this implies that $\phi(t,x) = 0$ for $x \in (x_p, \hat{x})$ and $t < t_h$. Now consider any $t > t_h$. Choose some t' such that $t > t' > t_h$. Note that $x(t') \geq \hat{x}$. Take $x = x(t')$ and b such that $\lambda h(b) = H(t')$, and apply claim 2. It follows that, for $x \in (x_p, \hat{x})$, $I(x,t) > I(x,t')$. Under the D1 criterion, this implies that $\phi(t,x) = 0$ for $x \in (x_p, \hat{x})$ and $t > t_h$. For completeness, one must rule out the possibility that, for any $x \in (x_p, \hat{x})$, there is some t for which $I(x, t_h) > I(x,t)$. The preceding argument rules out $t < t_h$. It also demonstrates that $I(x,t)$ is strictly decreasing in t for $t > t_h$. Thus, if there exists such a t, it must be the case that $\lim_{t \downarrow t_h} I(x,t) > I(x,t_h)$. But this implies that $\lim_{t \downarrow t_h} U^*(t) > U^*(t_h)$, which in turn implies that type $t_h$would imitate some type t slightly greater than t_h. Thus the claim is established. *Q.E.D.*

LEMMA 11 *In any signaling equilibrium where $g(.)$ is strictly concave, $s(x)$ is decreasing.*

PROOF: Follows directly from Lemma 7 and 10. *Q.E.D.*

Denote by x_p a pool, by t_l the smallest type $t_l : x(t_l) = x_p$ and t_h the largest type $t_h : x(t_h) = x_p$. Note that by the definition of a pool and Lemma 7, $t_l < t_h$. Then we state the following lemma. Let $\lambda^* \equiv -kg(\frac{1}{k})$.

LEMMA 12 *If $\lambda > \lambda^*$, then for any signaling equilibrium that satisfies the D1 criterion, there exists at most one pool $x_p \in X$ s.t. $\frac{1}{k} < t_h - t_l$ and it satisfies $0 = x_p$.*

PROOF: This is theorem 3, Bernheim 1994. The proof uses only Lemma 7 and Lemma 10 to show that whenever there exists a pool $x_p > 0$ s.t. $t_l < t_h - \frac{1}{k}$, then either t_l or t_h can pay a small intrinsic cost $g(\frac{1}{k})$ for a large gain in esteem $s(x) - s(x_p)$, where $x \in \{x_p - \frac{1}{k}, x_p + \frac{1}{k}\}$.

To see this, suppose that there exist some second pool not at 0. Let $x_p > 0$ be s.t. $t_l = x_p$ and $t_h = x_p + \frac{2}{k}$, and $x(t_h) = x_p$ and $x \notin x(t_h) \forall x > x_p$. For this to be an equilibrium two things must be true: $u(t_h, t_h) < u(x_p, t_h) \Leftrightarrow \lambda > \frac{-g(t_h - x_p)}{s(x_p) - s(t_h)}$ and $u(x_p, t_l) > u(x_p - \frac{1}{k}, t_l) \Leftrightarrow \lambda < \frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{s(x_p - \frac{1}{k}) - s(x_p)}$. There only exists such a λ that fulfills both the upper and the lower bound if $\frac{-g(t_h - x_p)}{s(x_p) - s(t_h)} < \lambda < \frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{s(x_p - \frac{1}{k}) - s(x_p)}$. A necessary condition is $\frac{\frac{-g(t_h - x_p)}{t_h - x_p}}{\frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{\frac{1}{k}}} < \frac{\frac{s(x_p) - s(t_h)}{t_h - x_p}}{\frac{s(x_p - \frac{1}{k}) - s(x_p)}{\frac{1}{k}}}$. Now, we will show that the lower bound on λ cannot be lower than the upper bound on λ.

Note that $x_p \leq t_l$. This follows directly from that by Lemma 11, any action $x(t) > t$ is strictly dominated.

As all t_h choose the action x_p, no other type chooses the action t_h (there are no better types to pool with for types $t > t_h$), hence by Lemma 10, $s(t_h) = -t_h$. Further, by Lemma 7 and Lemma 10 $s(x_p - \frac{1}{k}) \geq -t_l$. Hence $\frac{s(x_p) - s(t_h)}{s(x_p - \frac{1}{k}) - s(x_p)} \frac{\frac{1}{k}}{t_h - x_p} \leq \frac{s(x_p) + t_h}{-t_l - s(x_p)} \frac{\frac{1}{k}}{t_h - x_p}$. As all t_h and at least some t_l choose x_p, $s(x_p) \leq -\frac{t_h + t_l}{2}$, which is equivalent to $2s(x_p) \leq -(t_h + t_l) \Leftrightarrow s(x_p) + t_h \leq -t_l - s(x_p) \Leftrightarrow \frac{s(x_p) + t_h}{-t_l - s(x_p)} \leq 1$. Further, as $\frac{1}{k} < t_h - t_l$ and $x_p \leq t_l$ then $\frac{\frac{1}{k}}{t_h - x_p} < 1$ and hence $\frac{\frac{s(x_p) - s(t_h)}{t_h - x_p}}{\frac{s(x_p - \frac{1}{k}) - s(x_p)}{\frac{1}{k}}} < 1$.

As g is strictly concave, $g(a) + g(b) > g(a + b)$ hence $g(-\frac{1}{k}) + g(t_l - x_p + \frac{1}{k}) > g(t_l - x_p)$ and hence $\frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{\frac{1}{k}} < \frac{g(-\frac{1}{k})}{1/k}$. Hence $\frac{\frac{-g(t_h - x_p)}{t_h - x_p}}{\frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{\frac{1}{k}}} > \frac{\frac{-g(t_h - x_p)}{t_h - x_p}}{\frac{g(1/k)}{\frac{1}{k}}}$ and as $t_h - x_p > \frac{1}{k}$, $\frac{\frac{-g(t_h - x_p)}{t_h - x_p}}{\frac{g(1/k)}{\frac{1}{k}}} > 1$ and we have reached a contradiction; it cannot be true that $\frac{\frac{-g(t_h - x_p)}{t_h - x_p}}{\frac{g(t_l - x_p) - g(t_l - x_p + \frac{1}{k})}{\frac{1}{k}}} < \frac{\frac{s(x_p) - s(t_h)}{t_h - x_p}}{\frac{s(x_p - \frac{1}{k}) - s(x_p)}{\frac{1}{k}}}$ as the LHS >1 and the RHS <1.

We finally note that there can exist a pool at 0. At $x_p = 0$ there exists no action $x = x_p - \frac{1}{k}$ for t_l to consider, hence the condition for the existence of a pool reduces to $u(t_h, t_h) < u(0, t_h) \Leftrightarrow \lambda > \frac{-g(t_h)}{s(0) - s(t_h)}$. *Q.E.D.*

A.1.1 Existence of a fully separating equilibrium where $x(t) = t \forall t$ when $\lambda \leq \lambda^*$.

Let $\lambda^* \equiv -kg(\frac{1}{k})$. Suppose $\lambda \leq \lambda^*$ and that $x(t) = t, \forall t$. Then $s(x) = -x$ for all x. Then, as$u(t - \frac{1}{k}, t) - u(t, t) = g(\frac{1}{k}) + \lambda(\frac{1}{k} - t) - \lambda(-t) = g(\frac{1}{k}) + \lambda\frac{1}{k} \leq g(\frac{1}{k}) + \lambda^*\frac{1}{k} = g(\frac{1}{k}) - g(\frac{1}{k}) = 0$, hence t does not profit from choosing action $x = t - \frac{1}{k}$. By Lemma 7, if $u(x, t) < u(t, t)$, for $x < t$, then $u(x, t') < u(t, t')$ for all $t' > t$. Hence, as t does not deviate to $x = t - \frac{1}{k}$, no $t' > t$ deviates to this action either. This is true for each type $t \in [\frac{1}{k}, 1]$ and action $x \in [0, \frac{k-1}{k}]$, and we conclude that no type has a profitable deviation to some $x < t$. Note finally that, as $s(x)$ is decreasing, any action $x > t$ is strictly dominated.

A.1.2 Existence of equilibrium with properties 1-4. when $\lambda > \lambda^*$.

Suppose $\lambda > \lambda^*$, that actions are according to 1-3 and beliefs are according to 4.

If $\lambda > \lambda^*$ then no $t \in [0, \bar{t}]$ wants to deviate:

The esteem over this interval is given by $s(0) \geq -\frac{\bar{t}}{2}$ and $s(x) = -\bar{t}$ for all $x \in (0, \bar{t}]$. In this proof of existence, we look at the boundary condition for when type $\bar{t} = \frac{1}{k}$ does not want to deviate from a zero-mass of types $\bar{t}$ choosing action $x = 0$ and the rest choose action $x = \bar{t}$. Type $\bar{t}$ do not want to deviate if $u(0, \bar{t}) \geq u(\bar{t}, \bar{t}) \Leftrightarrow \lambda \geq \underline{\lambda} \equiv \frac{-g(\bar{t})}{s(0)-s(\bar{t})} = \frac{-2kg(\bar{t})}{k\bar{t}+1}$. Evaluate at $s(0) = 0$ and $\bar{t} = \frac{1}{k}$: $\underline{\lambda} = -kg(\frac{1}{k}) = \lambda^*$. Note that as $s(0)$ is decreasing in the share of types $t = 1$ who choose action $x(1) = 0$, $\underline{\lambda}$ is increasing, confirmingthat it is a lower bound. By Lemma 7, if $x(\bar{t}) = 0$ then $x(t) = 0 \ \forall t < \bar{t}$, and hence no $t < \bar{t}$ deviate either.

No $t \in (\bar{t}, 1]$ wants to deviate:

First we show that in equilibrium some share of each type will move one step towards the norm. Suppose not, that if $0 \notin x(t)$ then $x(t) = t$. Then, $s(x) = -x$ for each $x > \bar{x}$. Types $t > \bar{x}$ do not deviate if $u(t - \frac{1}{k}, t) < u(t, t) \Leftrightarrow \lambda < \frac{-g(\frac{1}{k})}{s(t-\frac{1}{k})-s(t)} = -kg(\frac{1}{k}) = \lambda^*$ and hence we have reached a contradiction. It directly follows that it must be true that everywhere in this interval $\lambda(s(t - \frac{1}{k}) - s(t)) = -g(\frac{1}{k})$, hence $s(x)$ is linear over $[\bar{x}, 1]$, i.e. $\lambda(s(t - \frac{m}{k}) - s(t)) = -mg(\frac{1}{k})$. We also note that as $s(x)$ is decreasing, any action $x > t$ is strictly dominated. There are then two possible deviations from equilibrium strategies, $x(t) \in (\bar{x}, t - \frac{2}{k}]$ and $x(t) \in [0, \bar{x}]$ and we disprove each in turn.

t does not deviate to $t - \frac{m}{k} \in [\bar{x}, 1)$,where $m > 1$, since $u(t - \frac{m}{k}, t) - u(t, t) = g(\frac{m}{k}) + \lambda(s(t - \frac{m}{k}) - s(t)) = g(\frac{m}{k}) - mg(\frac{1}{k}) < 0$. The last inequality follows from $g(.) < 0$ and concave.

Suppose $t > \bar{t}$ wants to deviate to 0. By Lemma 8, $t > \bar{t}$ require higher esteem to choose 0 than does $\bar{t}$, i.e. that $u(x, t) = g(t - x) + \lambda' s(x)$ where $\lambda' > \lambda$ and we have reached a contradiction.

Finally, suppose $t > \bar{t}$ wants to deviate to $x \in (0, \bar{x})$, where we note that x are off-equilibrium actions and by property 4, $s(x) = -\bar{t}$. It directly follows that for any $t \geq \bar{x}, \frac{\partial u(x,t)}{\partial x} > 0$ over $(0, \bar{x})$. By definition, $\bar{x}$ maximizes $u(x, \bar{t} + \frac{1}{k})$, inluding if there is a zero-mass of types $t = \bar{t} + \frac{1}{k}$: $x(t) = \bar{x}$. Let $\bar{x} = inf(x(\bar{t} + \frac{1}{k}))$. If there is a zero-mass choosing $\bar{x}$, then $s(\bar{x}) = -\bar{t} = s(x)$, $\forall x \in (0, \bar{x})$, hence $u(\bar{x}, \bar{t} + \frac{1}{k}) > u(x, \bar{t} + \frac{1}{k})$ for all $x \in (0, \bar{x})$. If there is a positive mass choosing $\bar{x}$, $s(\bar{x}) = -\bar{t} + \frac{1}{k}$ and some would want to

choose $x(t) = \bar{x} - \frac{1}{k}$, then $inf(x(t)) = \bar{x} - \frac{1}{k} < \bar{x}$ and we have reached a contradiction. If not all $\bar{t}$ choose $x(\bar{t}) = 0$ then $\bar{x} \in x(\bar{t})$ and $\bar{x} > 0$, hence by Lemma 7, $x(t) \geq sup(x(\bar{t}))$ for all $t \geq \bar{t}$ and we have reached a contradiction.

D1 beliefs for an off-equilibrium action $x \in (0, \bar{t})$ assign probability=1 to $t = \bar{t}$:

This is Lemma 10.

A.1.3 There exists precisely one equilibrium fulfilling properties 1-4 for any given $\lambda \geq \lambda^*$

By Lemma 8, $\bar{t}$ is strictly increasing in λ. It directly follows that $\bar{t}$ is uniquely determined by λ. Hence, as λ increases, a larger mass of types $\bar{t}$ chooses action 0, and if λ increases further, $\bar{t}$ increases.

Let λ^{**} solve $u(0,1) = u(1,1)$ for $s(0) = -\mathbb{E}[t]$, i.e. $\lambda^{**} = -2g(1)$. Then if $\lambda > \lambda^{**}$, $\bar{t} = 1$, i.e. the equilibrium is fully pooling.

A.1.4 Uniqueness of a separating equilibrium when $\lambda \leq \lambda^*$.

As in Bernheim (1994) λ is bounded from below in all partially pooling equilibria. To see this in our setting, note that a zero-mass of types $t = \frac{1}{k}$ want to pool at 0 if and only if $u(0, \frac{1}{k}) \geq u(\frac{1}{k}, \frac{1}{k}) \Leftrightarrow \lambda(s(0) - s(\frac{1}{k})) \geq -g(\frac{1}{k}) \Leftrightarrow \lambda \geq \lambda^*$. Hence if $\lambda < \lambda^*$, the equilibrium cannot be partially or fully pooling. As there can be no other equilibrium than fully pooling, partially pooling or separating, equilibria must be separating if $\lambda < \lambda^*$. For uniqueness it then only remains to show that there exists only one separating equilibrium when $\lambda < \lambda^*$ which follows from Lemma 9.

A.1.5 When $\lambda > \lambda^*$ no equilibrium exists where 1-4 do not hold.

Let $\lambda > \lambda^*$.

Property 1 ($x(t)$ is weakly increasing in t)

If property 1 does not hold, this contradicts Lemma 7.

Property 2 (there exists a $\bar{t} \leq 1$ such that $x(t) = 0$ for all $t \leq \bar{t}$)

From Lemma 9 follows that the only separating equilibrium is one where $x(t) = t \forall t$. Since $u(0, \frac{1}{k}) \geq u(\frac{1}{k}, \frac{1}{k}) \longleftrightarrow \lambda > \lambda^*$ a separating equilibrium cannot exist when $\lambda > \lambda^*$. Further, by Lemma 7, if $t' < t$, then $x(t') \leq x(t)$, hence if there exists some type t' such that $x(t') = 0$ then $x(t) = 0$ for all $t < t'$.

Next, suppose that there exists some other pooling equilibrium, not fulfilling property 2, i.e., one where there is partial or full pooling at a $x > 0$. Then some type $t > \frac{1}{k}$ moves to some $x(t) < t$ but $t = \frac{1}{k}$ does not pool at 0. By monotonicity in actions (Lemma 7) there is then only type 0 at 0, hence $s(0) = 0$. Further, as only types $t \geq \frac{1}{k}$ choose $x = \frac{1}{k}$, it must be that $s(\frac{1}{k}) \leq -\frac{1}{k}$. But then, the marginal gain for type $\frac{1}{k}$ of deviating to $x(\frac{1}{k}) = 0$ is at least as large as for any other type from choosing an action $x(t) < t$.

Property 3 (all types $t > \bar{t}$ take actions spanning a range $x(t) \in [\bar{x}, 1]$ for some $\bar{x} > 0$)

We have shown that property 2 must hold. Hence, suppose property 2 holds but not property 3. Whenever there exists some $t : x(t) > 0$ then $\bar{x} > 0$ and by Lemma 7, $x(t) \geq \bar{x}\ \forall t > \bar{t}$. For property 3 not to hold then, there must exist some $x' \in (0, 1)$ s.t. $\max x(1) = x'$.

Suppose $x' = \frac{k-1}{k}$. But, then $s(x' - \frac{1}{k}) - s(x') \geq s(x') - s(1)$ and hence $x(x') < x'$ why $\phi(x', 1) = 1$ and $s(x') = -1$ and $t = 1$ has a profitable deviation to $x = 1$, which is a contradiction.

Suppose $x' \in [\bar{x}, \frac{k-1}{k})$. By Lemma 12, there can only be one pool at 0, hence x' cannot be a pool. But then, $s(x') = -1 = s(1)$ why $u(1, 1) - u(x', 1) = \lambda s(1) - g(1 - x') - \lambda s(x') = -g(1 - x') > 0$ and we have a contradiction.

Finally, we remark that by Lemma 8, $\bar{t}$ is uniquely determined by λ.

Property 4 (beliefs for an off-equilibrium action x' assign a probability 1 to $t(x') = \bar{t}$)

This is Lemma 10.

A.2 Proof of Lemma 2

Part 1. Note that whenever a type $t = x'$ misrepresents herself, $x(x') \neq x'$, then no other type will choose the action x'. To see this, note that for any strictly convex function, $f(a) + f(b) < f(a + b)$ hence we can write $g(t' - t) + g(t - x) < g(t' - x) \Leftrightarrow -g(t - x) > g(t' - t) - g(t' - x)$. Hence if $\lambda s(x) - \lambda s(t) \geq -g(t - x)$ then $\lambda s(x) - \lambda s(t) > g(t' - t) - g(t' - x)$ which is equivalent to that if $u(x, t) \geq u(t, t) \Rightarrow u(x, t') > u(t, t')$.

Part 2. The divinity criterion selects the type that has the most to gain from an off-equilibrium action. If t has more to gain than any other t' by deviating to the off-equilibrium path action t, then Divinity selects t with probability 1. Let $U^*(t)$ be the equilibrium utility of type t. NTS (1) : $u(t, t) - U^*(t) > u(t, t') - U^*(t')$, $\forall t' \neq t$. The action t is empty, hence $\exists x : u(t, t) \leq u(x, t) = U^*(t)$. Hence, we write (1) : $u(t, t) - u(x, t) > u(t, t') - U^*(t')$. Further, it must be true that $U^*(t') \geq u(x, t')$, and note that a sufficient condition for (1) is (1)$' : u(t, t) - u(x, t) > u(t, t') - u(x, t')$ as $u(t, t') - u(x, t') \geq u(t, t') - U^*(t')$. Now: (1)$' : g(0) + \lambda s(t) - g(t - x) - \lambda s(x) > g(t' - t) + \lambda s(t) - g(t' - x) - \lambda s(x) \Leftrightarrow g(t' - x) > g(t' - t) + g(t - x)$ where we note that $t' - x = t' - t + t - x$, and that g is strictly convex, $f(a + b) > f(a) + f(b)$, why this is true for all $t' \neq t$.

A.3 Proof of Lemma 3

Suppose not, suppose there exists two points, on or off the equilibrium set, $v \in int(T)$ s.t. $s(v) > s(v - \epsilon)$ where $\epsilon \in \{\frac{1}{k}, \frac{2}{k}, ..., v\}$. There are two ways in which this could happen. Either, (i) some $t > v$, choose $v - \epsilon$, or (ii) some $t < v - \epsilon$, choose v. Or both. We will disprove each alternative in turn. First, suppose both v and $v - \epsilon$ are on the equilibrium path. Then,

(i): Suppose that there exists some $t > v$ s.t. $u(v - \epsilon, t) \geq u(v, t)$. But, since $s(v) > s(v - \epsilon)$, for

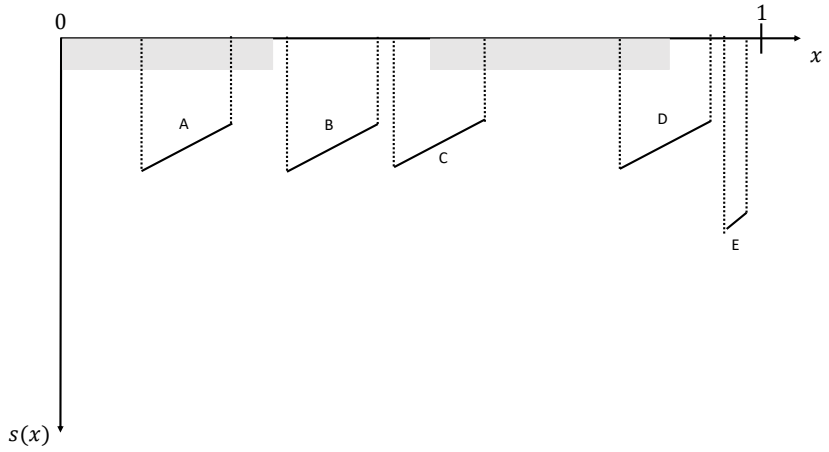
$t > v$, then $u(v,t) - u(v-\epsilon,t) = g(t-v) - g(t-v+\epsilon) + \lambda\big(s(v) - s(v-\epsilon)\big) > 0$, since $g(t-v) > g(t-v+\epsilon)$. Therefore, this is a contradiction.

(ii): Suppose that there exists some $t < v-\epsilon$ s.t. $u(v,t) \geq u(t,t) \Leftrightarrow g(t-v)+\lambda s(v) \geq \lambda s(t)$,condition (*). But, by Lemma 2, no other type will choose action t and hence $s(t) = -t$. Further, by the same Lemma 2, if $v \in x(t)$ then $x(v) = v$. But then, as $t < v$, $s(v) < s(t)$ which with $g(.)$ being negative contradicts condition (*).

If one or both points v and $v-\epsilon$ are off the equilibrium path, we apply the divinity criterion, then there are five ways in which this could happen. Call points within the equilibrium action set x and off the equilibrium action set $\tilde{x}$. Suppose that $\tilde{x}$ is within some interior interval off the equilibrium set, where actions both larger and smaller than $\tilde{x}$ are taken. We must then exclude (A) a point $\tilde{x}$ yields higher esteem than a point $x < \tilde{x}$, (B) esteem is increasing over two off equilbrium actions $\tilde{x_1}$ and $\tilde{x}_2$, where $\tilde{x_1} < \tilde{x_2}$, and (C) a point $\tilde{x}$ yields lower esteem than a point $x > \tilde{x}$. Then, suppose that there is an upper interval off the equilibrium set. We must then exclude (D) a point $\tilde{x}$ yields higher esteem than a point $x < \tilde{x}$, and (E) esteem is increasing over two off equilbrium actions $\tilde{x_1}$ and $\tilde{x}_2$, where $\tilde{x_1} < \tilde{x_2}$. Alternatives (A)-(E) are illustrated graphically in Figure 3.

By Lemma 2, the esteem off the equilibrium set is $s(x) = -x$ and hence we have contradicted claims (B) and (E). For claim (A) then, the esteem at x must be lower than at $\tilde{x}$, hence there must exist some type $t > \tilde{x}$ that chooses x. This is contradicted by argument (i). For claim (D), there must exist some type $t < x$ that chooses $\tilde{x}$. This is contradicted by argument (ii). Finally, claim (C) require that $s(x) > -\tilde{x}$, hence there must exist some type $t < \tilde{x}$ that chooses x. This is contradicted for the same reason as in (ii).

FIGURE 3.— Cases of increasing esteem



Notes: Sketch of possible places where the esteem function may be increasing that are refuted by the proof in Section A.3. The grey zones are those where actions are taken and the white zones are those where no action is taken so off-equilibrium beliefs are needed. For visibility the illustration looks as if the action space is continuous while in the model it is discrete.

A.4 Proof of Lemma 5

A. If $u(x',t') \geq u(x,t')$ for $x \in (x',t')$ then $u_x(x,t') \leq 0$ over (x',t'). Because $u_{xt}(.,.) < 0$, $u_x(x,t) < u_x(x,t') \leq 0$ for all $t > t'$, and it thus follows directly that $u(x',t) \geq u(x,t)$.

B. That $t' = \max_x u(x,t')$ is equivalent to that $u_x(x,t') \geq 0$ for all $x < t'$, with strict inequality for some x. Because $u_{xt}(x,t) < 0$, it is always true that $u_x(x,t) > u_x(x,t')$ for $t < t'$. Then, $u_x(x,t) > u_x(x,t') \geq 0$ and it must be true that $t = \max_x u(x,t)$ for all $t < t'$.

C. $u(x,t') = u(x',t') \Leftrightarrow \lambda\big(s(x)-s(x')\big) = g(t'-x')-g(t'-x)$. Then, $u(x',t)-u(x,t) = g(t-x')-g(t-x)-\lambda\big(s(x)-s(x')\big) = g(t-x')-g(t'-x')-\big(g(t-x)-g(t'-x)\big)$. Because for any $t < t' \Rightarrow g(t-y) > g(t'-y)$, for each $y \in \{x,x'\}$ and $x < x' \Rightarrow g(y-x') > g(y-x)$, for each $y \in \{t,t'\}$, it directly follows that $g(t-x') - g(t'-x') - \big(g(t-x) - g(t'-x)\big) > 0$.

B Proof of Proposition 1

B.1 Proof of Proposition 1, part a:

We here prove that some anormative take the norm action in public, $0 \in x(1)$. Equivalently, if $\lambda > 1$ then $u(0,1) \geq u(1,1)$.

We show this in two steps. First, claim 1: whenever some $x(t) < t$ then there exists some $t > 0$ s.t. $x(t) = 0$. Second, claim 2: if there exists some $t > 0$ s.t. $x(t) = 0$ then $x(1) = 0$.

B.1.1 Claim 1: whenever some $x(t) < t$ then there exists some $t > 0$ s.t. $x(t) = 0$.

We prove this by contradiction. Suppose that there exists some equilibrium strategy profile $\min\{x(t')\} = b \in (0, t')$, while there exists no $t > 0$ s.t. $x(t) = 0$. Then, $s(0) = 0$ and $s(b) = -k$, where by Lemma 5.1, $k > b$ because of the pooling of worse types at $x = b$. Define $\Delta_x u(x,t)$ over the domain $[x_1, x_2]$ as the gain in utility for t from choosing action x_2 rather than x_1. For this to be part of an equilibrium, it must be true that $\Delta_x u(x,t') > 0$ over $[0,b]$ and that $\Delta_x u(x,t') \leq 0$ over $[b,1]$. Consider each interval in turn.[14] Denote by l_1 the number of points in $[0,b]$ and by l_2 the number of points in $[b,1]$. We can rewrite the conditions as $\Delta_x g(t'-x) + \lambda \Delta_X s(x) > 0$ over $[0,b]$ and $\Delta_x g(t'-x) + \lambda \Delta_X s(x) \leq 0$ over $[b,1]$. As $s(x)$ is decreasing in any equilibrium, Lemma 3, $\Delta_x s(x) < 0$ and we write $\lambda < \frac{\frac{\Delta_x g(t'-x)}{k_1}}{\frac{-\Delta_x s(x)}{k_1}}$ and $\lambda \geq \frac{\frac{\Delta_x g(t'-x)}{k_2}}{\frac{-\Delta_x s(x)}{k_2}}$.Denote by x_1 any $x \in [0,b]$ and by x_2 any $x \in [b,1]$. $\frac{\frac{\Delta_{x2} g(t'-x_2)}{k_2}}{\frac{\Delta_{x1} g(t'-x_1)}{k_1}} < \frac{\frac{s(b)-s(1)}{k_2}}{\frac{s(0)-s(b)}{k_1}}$, A necessary condition is then that the difference in the average relative intrinsic utility increase is smaller than the average relative increase in social esteem. As $x_1 < x_2$ and $g''(.) > 0$, $\frac{\frac{\Delta_{x2} g(t'-x_2)}{k_2}}{\frac{\Delta_{x1} g(t'-x_1)}{k_1}} > 1$. As $s(b) < -1$, $s(0) = 0$ and $s(1) = -1$, $-\frac{s(b)-s(1)}{k_2} < 1$ and $-\frac{s(0)-s(b)}{k_1} > 1$, why $\frac{\frac{s(b)-s(1)}{k_2}}{\frac{s(0)-s(b)}{k_1}} < 1$ and we have arrived at a contradiction.

B.1.2 Proof of Claim 2: if there exists some $t > 0$ s.t. $x(t) = 0$ then $x(1) = 0$.

If $u(0,t) \geq u(t,t)$ Lemma 5 applies and the first part tells us that all $t' > t$ prefers 0 to $x \in (0,t)$. Note that $t = 1$ has more to gain from going to 0 than does t. By Lemma 2, $s(t) = -t$ and $s(1) = -1$. The net utility gain for t is hence $u(0,t) - u(t,t) = g(t) + \lambda s(0) - \lambda s(t)$ while for 1 it is $u(0,1) - u(1,1) = g(1) + \lambda s(0) - \lambda s(1)$. Note that $u(0,t) - u(t,t) < u(0,1) - u(1,1) \Leftrightarrow g(t) - g(1) < \lambda s(t) - \lambda s(1) \Leftrightarrow \frac{g(t)-g(1)}{1-t} < \lambda$ and note that $\frac{g(t)-g(1)}{1-t} < 1 < \lambda$. Finally, note that there is no $x \in (t,1)$ s.t. $x \in x(1)$ while $0 \notin x(1)$.

Suppose there is some $x \in (t,1)$ s.t. $x \in x(1)$ while $0 \notin x(1)$. Then, the following two things must be true. (1) : $u(x,1) \geq u(1,1)$ and (2) : $u(0,1) < u(1,1)$. Further, as $x > t$ is dominated we have (3) : $u(0,t) > u(x,t)$. By simple algebra, (1) and (3) gives that $g(t) + \lambda s(0) - g(t-x) > \lambda s(x) \geq \lambda s(1) - g(1-x)$ and hence it must be true that $\lambda s(0) - \lambda s(1) > -g(t) + g(t-x) - g(1-x)$. Rewrite (2) to $\lambda s(0) - \lambda s(1) < -g(1)$. Hence all three only hold if $g(1) - g(1-x) < g(t) - g(t-x)$. But as g is strictly convex, its derivative is increasing and this is a contradiction.

Then, $s(x) < -x$. But then, the marginal gain in esteem, $\frac{s(t)-s(x)}{x-t} > 1$ while the marginal cost

[14] The continuous argument is equivalent as when the function increases then so must the sequence defined by the same function. Over $x \in [0,b]$, denote by x_a some particular x in this domain and $s_a(.)$ the social esteem function. We have that $U_x(x,t') > 0 \Leftrightarrow \lambda < \frac{g_x(t'-x_a)}{s'_a(x)}$. Similarly, over $[b,1]$, denote by x_b some particular x in this domain and $s_b(.)$ the social esteem function. Then we have that $U_x(x,t') < 0 \Leftrightarrow \lambda > \frac{g_x(t'-x_b)}{s'_b(x)}$. For there to exist a λ that supports this equilibrium it must be true that $\frac{g_x(t'-x_b)}{s'_b(x)} < \lambda < \frac{g_x(t'-x_a)}{s'_a(x)}$. A necessary condition is then that $(-g_x(t-x) > 0,\ -s'(x) > 0)$ $\frac{g_x(t'-x_b)}{g_x(t'-x_a)} < \frac{s'_b(x)}{s'_a(x)}$, where $x_a < x_b$ and $g''(.) > 0$ why $\frac{g_x(t'-x_b)}{g_x(t'-x_a)} > 1$. Now, because some types $t > b$ pool at b, $s(b) < -b$. Because no t go to 0, $s(0) = 0$, why $-s'(x_a) > 1$ and $-s'(x_b) < 1$. Thus, $\frac{-s'_b(x)}{-s'_a(x)} < 1$ and we have arrived at a contradiction.

of misrepresentation $\frac{g(1-t)-g(1-x)}{x-t} < \frac{-g(x-t)}{x-t} < 1$, where the first inequality follows from $g(.)$ symmetric around 0 and strictly convex, hence $g(a)+g(b) < g(a+b)$, and the second inequality by normalization of g, $-g(1) < 1$, holds for all $x - t \leq 1$.

B.2 Proof of Prop 1, part b:

We here prove that no other type takes the norm action in public, $0 \notin x(t)$, $t \in (0,1)$. In other words, $0 \in x(t)$ *only if* $t \in \{0,1\}$.

By Proposition 1.a, $x(1) = 0$. It remains to show that $0 \notin x(t)$ $\forall t \neq \{0,1\}$. We show this in two steps. Claim 1: There exists no equilibrium where all type 1 choose action 0, $x(1) = 0$. Claim 2: If $x(t) \neq t$ for some $t < 1$, then by part 2 of Lemma 5, $1 \notin x(1)$. We show that then $\frac{1}{k} \in x(1)$ and by part 3 of Lemma 5 $u(0,t) < u(\frac{1}{k},t)$ for all $t < 1$.

B.2.1 Claim 1: There is no equilibrium where all type 1 pool at 0:

A necessary (not sufficient) condition for all type 1 to go to 0 is that s is monotonically decreasing over X. The only way to increase $s(0)$ is for some better-than-average type to go to 0, because $s(0) = -\frac{1}{2} = -\mathbb{E}[t]$. But by monotonicity in s, we have monotonicity in incentives of t, and this cannot be part of an equilibrium. As we cannot increase $s(0)$ we must decrease $s(\frac{j}{k})$ for all $j \in \{1,2,..\frac{k+1}{2})$, as by monotonicity in t, $s(\frac{\frac{k+1}{2}}{k}) \leq -\frac{1}{2} = s(0)$. In particular, we must decrease s at more points than there are worse-than-average types. But, for any j, we need more than one type to pool for s to be strictly decreasing; If $\frac{j}{k} \in x(\frac{k-j}{k})$, then $s(\frac{j}{k}) \geq -\frac{1}{2} = s(0)$, with equality if all type $\frac{k-j}{k}$ take action $\frac{j}{k}$; $x(\frac{k-j}{k}) = \frac{j}{k}$ (with the logic of Gauss's arithmetic sequence).

B.2.2 Claim 2: If $1 \notin x(1)$ then $\frac{1}{k} \in x(1)$.

Suppose not, that $1 \notin x(1)$ and that $\frac{1}{k} \notin x(1)$. Then, there exists some $x > \frac{1}{k}$ s.t. $u(x,1) > u(\frac{1}{k},1) \Leftrightarrow \lambda\frac{s(\frac{1}{k})-s(x)}{x-\frac{1}{k}} < \frac{g(1-x)-g(1-\frac{1}{k})}{x-\frac{1}{k}}$. By convexity of g, $\frac{g(1-x)-g(1-\frac{1}{k})}{x-\frac{1}{k}} = \frac{g(1-x)-g(1-\frac{1}{k})}{(-1)(1-x-(1-\frac{1}{k}))} \leq -g'(1-\frac{1}{k})$. By Lemma 5.3, it must be true that $u(x,t) > u(\frac{1}{k})\forall t < 1$, why $s(\frac{1}{k}) = -\frac{1}{k}$. Further, by $x \in x(1)$, $-s(x) > x$, thus $\lambda\frac{s(\frac{1}{k})-s(x)}{x-\frac{1}{k}} > \lambda\frac{-\frac{1}{k}+x}{x-\frac{1}{k}} = \lambda > -g(1)$, and we have arrived at a contradiction because $-g(1) > -g'(1-\frac{1}{k})$. Now, as both $\{0,\frac{1}{k}\} \in x(1)$ it must be true that $u(0,1) = u(\frac{1}{k},1)$ and Lemma 5.3 applies and action $x = 0$ is dominated for all $t \neq \{0,1\}$.

B.3 Proof of Proposition 1, part c:

We here prove that the normative take the norm action in public, $x(0) = 0$.

PROOF: Follows directly from Lemma 2, where whenever there is a pool at some point x_p then $x(x_p) = x_p$. *Q.E.D.*

B.4 Proof of Claim: For any finite number of types and actions, $k \in (1, \infty)$, there can always exist some mass of hypocrites $p > 0$ s.t. $s(x)$ is strictly decreasing.

Mass $\frac{1}{k+1}$ of each type, of which share p of type 1, meaning mass $\frac{p}{k+1}$, are hypocrites and pool at 0. Then $s(0) = \Pr(t=0|x=0)h(0) + \Pr(t=1|x=0)h(1) = -\frac{\Pr(x=0|t=1)\Pr(t=1)}{\Pr(x=0|t=0)\Pr(t=0)+\Pr(x=0|t=1)\Pr(t=1))} = -\frac{p\frac{1}{1+k}}{\frac{1}{k+1}+p\frac{1}{k+1}} = -\frac{p}{1+p}$.

Proof: $s(x)$ is then strictly decreasing if $s(0) > \max\{s(\frac{1}{k})\} \Leftrightarrow -\frac{p}{1+p} > -\frac{1}{k} \Leftrightarrow p < \frac{1}{k-1}$. Note that $\lim_{k\to\infty} \frac{1}{k-1} = 0$. *Q.E.D.*

B.5 Proof of uniqueness

Finally, we prove that this equilibrium is unique. We have shown that when $\lambda > 1$ each claim (a), (b) and (c) in Proposition 1 is true (Appendix B1, B2 and B3). In this section, we show that when $\lambda > 1$, there is no other equilibrium.

First, note that (by proof of (c), section B3) (c) follows necessarily from (a), hence is always true whenever (a) is true. Second, note that (by proof of (b), section B2) (b) follows necessarily from (a). Hence, we show uniqueness by posing the contradictory claim that there exists some other equilibrium where (a) is not true. By Lemma **3** s is decreasing in every equilibrium, and if s is decreasing, actions $x > t$ are strictly dominated. In the proof of (a), section B.1.1., we show that whenever there exists some type t s.t. $x(t) < t$ then there exists some t' s.t. $x(t') = 0$. In section B.1.2., on the other hand, we show that whenever there exists some t' s.t. $x(t') = 0$ then it must be true that $x(1) = 0$, and we have arrived at a contradiction.

Proof: Finally, we note that the share of extremists who go to 0 is uniquely determined by λ. $u(0,1) \geq u(1,1) \Leftrightarrow \lambda \geq \lambda_1 \equiv \frac{-g(1)}{s(0)-s(1)}$ where we note that $g(1), s(1)$ are constant and that $s(0)$ is smoothly decreasing in the share of types 1 who choose 0, hence this lower bound λ_1 is increasing in the share of types 1 who go to 0. Note that this is true for any $\lambda \geq \frac{-g(t-x)}{s(x)-s(t)}$ and that by Lemma 5 part 2, whenever $t \in x(t)$, $t' \in x(t')$ $\forall t' < t$, hence λ must increase for further types to move. *Q.E.D.*

B.6 Corollary: $s(x)$ is concave in any equilibrium.

In any equilibrium, $\frac{s(\frac{1}{k})-s(0)}{\frac{1}{k}} \in (-1, 0)$. To see this, note that for small λ, $s(0) = -\frac{p}{1+p}$, where p small, increasing in λ. For small λ, $s(\frac{1}{k}) = -\frac{1}{k}$, increase λ, some type 1 go to $\frac{1}{k}$. In any equilibrium where $\frac{1}{k} \in x(1)$, it must be true that $u(0,1) = u(\frac{1}{k}) \Leftrightarrow s(\frac{1}{k}) - s(0) = \frac{g(1)-g(\frac{k-1}{k})}{\lambda} \Rightarrow \frac{s(\frac{1}{k})-s(0)}{\frac{1}{k}} = \frac{k}{\lambda}\big(g(1) - g(\frac{k-1}{k})\big) \in (-1, 0)$. This is easy to iterate over all x. For any $\frac{j}{k} \in X$, it must either be true that $\frac{s(\frac{j}{k})-s(\frac{j-1}{k})}{\frac{1}{k}} = -1$ (no pooling at

j nor $j-1$), or some $t > \frac{j}{k}$ is indifferent: $\frac{s(\frac{j}{k})-s(\frac{j-1}{k})}{\frac{1}{k}} = \frac{k}{\lambda}\big(g(t-\frac{j-1}{k}) - g(t-\frac{j}{k})\big) \in (-1,0)$, or very similarly, t prefers $\frac{j-1}{k}$ and $t-\frac{1}{k}$ is indifferent between truth and $\frac{j-1}{k}$.