Black Sheep and Walls of Silence

Gerd Muehlheusser  
University of Bern and IZA

Andreas Roider*  
University of Bonn and IZA

Revised version: February 2005  
First version: July 2004

Abstract

In this paper we analyze the frequently observed phenomenon that (i) some members of a team ("black sheep") exhibit behavior disliked by other (honest) team members, who (ii) nevertheless refrain from reporting such misbehavior to the authorities (they set up a "wall of silence"). Much cited examples include hospitals and police departments. In this paper, these features arise in equilibrium. An important ingredient of our model are benefits that agents receive when cooperating with each other in a team. Our results suggest that asymmetric teams where these benefits vary across team members are especially prone to the above mentioned phenomenon.

Keywords: teams, misbehavior, wall of silence, asymmetric information  
JEL-Classification: D82, C73.

*Corresponding author: Andreas Roider, University of Bonn, Department of Economics, Wirtschaftspolitische Abteilung, Adenauerallee 24-42, 53113 Bonn, Germany; roider@uni-bonn.de (address from August 2004 till July 2005: Andreas Roider, Stanford University, Stanford Graduate School of Business, 518 Memorial Way, Stanford, CA 94305-5015, USA). We thank Georg Nöldeke, Jörg Oechssler, and Urs Schweizer, Stefan Reichelstein, Ilga Vossen, and seminar participants at the Köln-Bonner Colloquium on Personnel Economics for suggestions, and to two referees and an Associate Editor for extensive and very helpful comments. Both authors gratefully acknowledge financial support from the Graduiertenkolleg "Quantitative Economics" at the University of Bonn. The second author gratefully acknowledges the hospitality of Stanford Graduate School of Business and a DFG research fellowship.
1 Introduction

Motivation In July 2003, a test driver of DaimlerChrysler drove a Mercedes prototype from corporate headquarters in Stuttgart (Germany) to the company’s test site in Papenburg, which is located about 300 miles to the north. On the highway he drove very fast (allegedly 150 m.p.h.) thereby massively tailgating a slower car. The driver of that car became so scared by the incident that she hit two trees on the roadside after losing control over her vehicle. Both, the driver and her two-year old daughter were killed. In the courtroom, the key question was whether it had really been the test driver who had tailgated the slower car. Hence, the timing of the test driver’s trip became an issue, and precise evidence on his departure time from headquarters and his arrival time at the test site was crucial. Yet such information was very hard to elicit as colleagues of the test driver were claiming they could not remember any details at all. In the end, the test driver was convicted by testimony of two other motorists whom he had passed shortly before the accident. After the trial, the judge complained about the test driver’s colleagues’ strong reluctance to cooperate with the authorities, presuming that none of them liked to be considered a denigrator (see Süddeutsche Zeitung, February 17, 2004, p. 17).

In this paper, we study two interrelated questions. First, we ask why individuals such as the test driver’s colleagues might implicitly tolerate certain actions by fellows even if they strongly dislike them? That is, why might they “set up a wall of silence”? Second, we simultaneously study how such potential walls of silence affect the incentives of would-be “black sheep” to misbehave.

Apart from the above example, there are many other settings where similar phenomena arise. The most prominent example are certainly police departments, where the so-called blue wall of silence refers to police officers’ reluctance to testify against their colleagues. For example, in an anonymous survey conducted among US police officers, 79% of 1,116 respondents confirmed that a code of silence existed (see Trautman (2000)). Furthermore, according to the Mollen Commission that investigated police violence in New York “the vast majority of honest police officers still protect the minority of corrupt officers”. ¹ In

---

¹See Commission to Investigate Allegations of Police Corruption and the Anti-Corruption Procedures of
a similar vain, Chevigny (1995, p. 92) reports that according to members of New York’s Civilian Complaint Review Board, “it had never had a case in which a police witness testified against another”. Moreover, this phenomenon does not seem to be confined to the US police departments, see Huberts, Lamboo, and Punch (2003) for evidence on the Netherlands and Ekenvall (2003) for evidence on Sweden and Croatia. Walls of silence also exist in other areas of law enforcement: According to the Hagar report which investigated California’s Department of Corrections, there is a "pervasive code of silence" among prison guards "that protects rogue guards...").

Further examples abound: According to the white wall of silence, doctors are reluctant to testify against colleagues in cases of malpractice. In labor arbitration, "arbitrators are aware that many employers refrain from calling co-workers as witnesses out of respect for ‘the code’ that prohibits employees from testifying against one another" (Gosline (1988, p.45)). In the education sector, high-school teachers remain silent about blatantly failing colleagues in the classroom, while students underreport misbehavior by their classmates: For example, in a survey among 3400 Toronto high school students, Tanner and Wortley (2002) found that more than half of them did not report to adult authority figures - parents, teachers or police - after being victimized, and they conclude that there exists a teen code of silence. Finally, in a community context, individuals are often reluctant to report crimes to the police when the criminal belongs to their community or even family. In summary, these examples suggest that walls of silence are empirically relevant, in fact, Neal Trautman, director of the US-based National Institute of Ethics, goes even so far as to suggest that "the code of silence exists within virtually all organizations" (see Trautman (2004)).

---

2 See Special Master’s Final Report Re Department of Corrections Post Powers Investigations and Employee Discipline, United States District Court, No.C90-3094-T.E.H.

3 For example, according to the Committee on Quality of Health Care in America, while the number of deaths in US hospitals due to malpractice is estimated to be up to 98,000 per year, "two thirds of the nation’s hospitals haven’t reported a single adverse incident involving a physician in the last eight years" (Kohn, Corrigan, and Donaldson (1999)). More examples in this context are provided in a recent book by Gibson and Singh (2003).


6 See e.g. Freeman (1999), Donohue and Levitt (2001), and Finkelhor and Ormrod (2001).
Given these examples, two principal (and interrelated) questions arise: First, how can one explain why misbehavior is often disliked, but nevertheless tolerated? The explanation we evoke is based on three observations. First, there is abundant evidence that reporting does not occur out of reputational concerns. For example, violators of the code of silence are often labeled "rats", "snitches" or "squealers" and are thus no longer respected by their colleagues and fellows. Second, individuals often derive substantial (cooperation) benefits from being an accepted group member instead of being ostracized. For example, police officers or prison guards need to be backed up in dangerous situations so that it is important for them to have attentive colleagues around. Whistle-blowers might not receive maximum backup in such situations. Also in less life-threatening situations, whistle-blowers suffer from the consequences of ostracism: Quoting R. Krupp: "...I would go for promotional interviews, and some of the defendants in my [whistle-blower] case were sitting on the panel - my interview panel. So, I ... stopped participating in the interviews, because ... it was a waste of time...". In the medical context, Gibson and Singh (2003, p.137) report the case of a physician who concedes, after having published a study about the unusually higher number of cardiac arrests in his hospital: "I can't prove it, but I suspect my appointment to full professor was delayed for several years as a result of this paper". Furthermore, as explained by an anonymous doctor in a CNN interview, a doctor might easily suffer through ostracism because "other doctors could put him out of business by refusing to refer him patients". Hence, potential whistle-blowers worry about their reputation because they do not want to forego future cooperation benefits and, as a consequence, refrain from reporting.

Second, how do potential walls of silence influence the behavior of potential “black sheep”, i.e., team members who may pursue activities that increase their own payoff but are disliked by their fellows? For example, doctors may save on effort costs when not taking appropriate care thereby causing harm to patients. Furthermore, police officers may handle suspects in a

---

7Thus, we are not enquiring why criminal teams, where all members are misbehaving, are stable in the sense that, if caught, none of them cooperates with the authorities.

8See e.g. Gosline (1988). Moreover, in the above-mentioned survey reported in Trautman (2000), the fear of being “ostracized” and “blackballed” were the most frequent answers to the question why police officers would not report misbehavior by colleagues. Very similar explanations are also given for the other above-mentioned examples. See the references given above.


manner that, while acceptable to themselves, may be considered unduly harsh or even brutal by others.\textsuperscript{11}

**Framework and results** We analyze a model that exhibits the basic features of the above mentioned examples as equilibrium phenomena. The aim is to provide conditions under which “black sheep” misbehave and such misbehavior is tolerated by honest team members. In our framework individuals decide whether or not to cooperate and form a team. When deciding upon team formation, each individual compares the expected gain from the cooperation to his outside option he earns absent cooperation. As for reputational concerns, honest team members differ with respect to their (privately known) willingness to report misbehavior. For example, they might simply face different opportunity costs from filing a report, or a fraction of them dislikes cooperation with the authorities. Alternatively, they might differ either with respect to their (intrinsic) willingness to treat team fellows favorably, or to obey a given norm (such as “Do not testify against your fellows”). Moreover, it seems plausible to assume that an agent’s willingness to report misbehavior is not common knowledge. Indeed, as Hoover Fellow and former San Jose police chief Joseph McNamara notes: “A corrupt, racist or brutal cop will abstain from misconduct only when he looks at the cop next to him and believes that the officer will blow the whistle if he hits the suspect”.\textsuperscript{12,13}

As a consequence, the reporting decision may convey information about an agent’s type and this, in turn, might affect their future payoffs. The basic mechanism at work is that, when reporting misbehavior, honest team members may forego future cooperation benefits (with ”black sheep” or with other team members who also observe the reporting). In turn, anticipating that reporting may not occur leads black sheep to misbehave in the first place.

Our aim is to provide conditions under which black sheep indulge in misbehavior and honest team members set up a wall of silence. In particular, we find that equilibria where

\textsuperscript{11}In the survey reported by Trautman (2000), excessive use of force was indeed the most frequent kind of misbehavior for which fellow officers confirmed to have refrained from reporting it.


\textsuperscript{13}Of course, this assumption seems the more plausible, the higher the degree of turnover. There is some evidence that turnover is considerable in the police context: According to a San Francisco police official “sixty percent of our patrol officers have been on the force less than five years” (See "Cracking the Code of Silence," *San Francisco Chronicle*, March 9, 2003). We thank an Associate Editor for pointing out this issue.
walls of silence occur jointly with potentially high levels of misbehavior exist in cases where teams are asymmetric in the sense that the benefit from cooperation is more important for honest team members than for black sheep compared to their respective outside option. On the other hand, when teams are sufficiently symmetric, equilibria with walls of silence cannot be sustained and, as a consequence, black sheep choose to not misbehave.

2 Relation to the Literature

While the phenomenon of “walls of silence” has received attention in the law and sociology literatures, to the best of our knowledge there exists only one other formal treatment of this phenomenon which is Benoit and Dubra (2004). They ask “Why Do Good Cops Defend Bad Cops?” and aim at explaining why honest team fellows protect dishonest ones. Also relying on a framework of asymmetric information, they enquire why a single (and potentially honest) agent would favor the representation of all agents (the union) to indiscriminately defend misbehaving colleagues over employing a candid strategy in which the union honestly reports all information it has. The basic idea is that in the first case a court will tend not to listen too much to the union’s statement because it contains no information. If the prior of the court is good enough, then even some honest agents prefer the indiscriminate strategy because this reduces the probability of being subject to a type II error. In contrast, the explanation we evoke for this phenomenon does not rely on enforcement errors, but is based on interaction benefits. While such benefits are also mentioned in Benoit and Dubra (2004, p. 787), they are not part of their model. Moreover, in contrast to Benoit and Dubra (2004, p. 787), in our model the level of misbehavior by an agent is not exogenously given, but arises endogenously from that agent’s optimization problem, taking into account the costs and benefits from misbehavior. Finally, a wall of silence is not the result of an action of a non-strategic player (the union) as in Benoit and Dubra (2004), but is derived from an honest agent’s optimal reporting decision.

Moreover, our paper contributes to the growing economic literature on social norms

14 With respect to the former, see e.g., the above cited articles by Gosline (1988) Epstein (2002), and the empirical work by Finkelhor and Ormrod (2001). In sociology there is a substantial literature on "police integrity": see e.g., Ekervall (2003), Huberts, Lamboo, and Punch (2003), and the references cited therein.
(see e.g. Akerlof (1980) and Huck, Kübler, and Weibull (2003)). In particular, since our analysis relies on a signaling game, it is related to a recent literature also based on signaling approaches that aims to explain related phenomena in the contexts of social interaction and law enforcement: Bernheim (1994) shows how heterogenous individuals are willing to conform to a single standard when popularity is deemed sufficiently important. Contrary to our paper, the issue of how this willingness can be exploited by group members is not addressed.\textsuperscript{15} Battaglini, Benabou, and Tirole (2003) analyze the incentive of individuals to interact with peers to learn more about their own characteristics.

The present paper could also be seen as a contribution to the literature analyzing factors that hinder the flow of information within organizations: while previous research (see e.g. Levitt and Snyder (1997)) has analyzed this issue in standard moral hazard frameworks such that exerting (unobservable) effort is costly, we provide a different rationale to explain why agents might have an incentive to hinder the flow of information within organizations.

The remainder of the paper is organized as follows: In Section 3 the model is set up, which is then analyzed in Section 4. Section 5 concludes. All proofs are relegated to an appendix.

3 The Model

We consider two risk-neutral individuals, $B$ and $G$ who may derive benefits $b^c > 0$ and $g^c > 0$, respectively, from cooperating with each other in a team. Individual $B$ is a (potential) ”black sheep” who might engage in activities disliked by the ”good guy” $G$. Throughout the paper, we will refer to such activities as ”misbehavior”. If no team is formed, $G$ and $B$ work on their own (thereby foregoing the benefits from cooperation), where the values of their outside options are denoted by $g^o > 0$ and $b^o > 0$, respectively.

\textbf{Stage game} To capture dynamic effects, we assume that $B$ and $G$ play a stage-game that is repeated twice, where the two periods are denoted by $t \in \{1, 2\}$. The stage-game itself

\textsuperscript{15}A framework of incomplete information is also used by Kim and Ryu (2003) and Kim and Lee (2001) to explain (non)conformity of agents.
consists of four dates (see Figure 1 below):

![Figure 1: Stage game](image)

**Dates 1 and 2 (team formation).** To model the issue of team formation and cooperation as simply as possible, we assume that at date 1, B decides whether or not to offer G to form a team \((T_B \in \{1, 0\})\). In case an offer has been made, G decides at date 2 whether or not to accept \((T_G \in \{1, 0\})\). When a team is formed (i.e., when \(T_B \cdot T_G = 1\)), B and G receive their respective cooperation benefits \(b^c\) and \(g^c\), while otherwise they receive their reservation payoffs, \(b^o\) and \(g^o\).

**Date 3 (misbehavior).** Given that a team has been formed, B might choose to "misbehave" by taking an action \(m \in [0, M]\) that generates a private gain \(b(m)\), where \(b\) is an increasing, concave function satisfying \(b(0) = 0\). As explained above, such behavior is disliked by G and reduces his payoff by \(m\).

**Date 4 (reporting).** At date 4 G decides whether or not to report the (potential) misbehavior by B to some authority \((R \in \{1, 0\})\) that may then investigate the case.

**Payoff consequences of reporting** We assume that G derives an expected (gross) benefit \(r(m)\) from reporting, which might for example reflect his satisfaction from seeing B penalized for his misbehavior.\(^{17}\) We impose the following three (as we think realistic) properties on \(r(m)\): i) there is no benefit from reporting unless there is misbehavior by B, i) the benefit

---

\(^{16}\) Alternatively, consider a setting where the parties first decide whether to form a team, and subsequently whether to cooperate (if a team has indeed been formed). Our assumptions below ensure that if the parties decide to forego the benefits from cooperating, they will prefer their outside options. As a consequence, we do not consider the team formation and cooperation decisions separately, and use these terms interchangeably. Note that our results below do not rely on the order of moves and would continue to hold if G moves first.

\(^{17}\) Benoit and Dubra (2004) report that moral considerations constitute a major reason for many (out of the few) police officers who testify against their colleagues.
is the higher, the higher the level of misbehavior, and iii) to rule out the unrealistic case that $G$ prefers high levels of misbehavior because the private benefit from conviction is so large, we assume that, although there is a benefit from conviction, $G$ prefers lower levels of misbehavior. Formally, these assumptions amount to $r(0) = 0$, $r' > 0$, and $0 < r' < 1$. However, as discussed in the introduction, $G$'s willingness to cooperate with the authorities is not common knowledge. To model such heterogeneity as simply as possible we assume that $G$ is one of two possible types $\theta \in \{H, D\}$. While a “conscientious” type $H$ has low reporting costs (which we set equal to zero), an “opportunistic” type $D$ faces fixed reporting costs $\tau > 0$. Thus while type $H$ is readily willing to report any level of misbehavior, type $D$ is more reluctant to do so: it follows that the net benefit from reporting differs across types and is $r(m)$ for type $H$ and $r(m) - \tau$ for type $D$. As a result, type $D$ will only cooperate with the authorities if the level of misbehavior is sufficiently large.\footnote{This is strongly confirmed by existing empirical research on reporting behavior, see e.g. Huberts, Lamboo, and Punch (2003) and Finkelhor and Ormrod (2001).} $G$’s type $\theta$ is private information, and $G$ learns it in the first period after a team has been formed.\footnote{This assumption allows us to focus on the potential signaling effect of the reporting decisions.}

With prior $h \equiv \text{Prob}(\theta = H) > 0$, he is “conscientious” and with probability $(1 - h)$ he is “opportunistic”, and the prior $h$ is common knowledge.

As for the payoff consequence of reporting for $B$, we take the authority’s enforcement technology as exogenously given, i.e., we assume that it is independent of the players’ actions and can be represented by a mapping from the level of misbehavior to expected penalties. These penalties consist of a fine and/or the monetary equivalent of imprisonment. In particular, given that $B$ has taken action $m$ and has been reported by $G$, the expected penalty that $B$ faces is given by a function $p(m)$ with the following properties: i) enforcement is "non-erroneous" in the sense so that there is no penalty unless there is misbehavior, ii) the penalty is the higher, the higher the level of misbehavior, and iii) penalties are sufficiently high to deter misbehavior if reporting occurs with certainty. Formally, this amounts to assuming $p(0) = 0$, $p' > b'$, and $p'' \geq 0$.\footnote{Of course, in reality the expected penalty $p(m)$ is the product of a certain probability with which $B$ is found guilty and the resulting penalty from conviction. However, for ease of notation we only use the expected penalty. Our assumption that the expected penalty is increasing in the level of misbehavior is consistent with the notion of "marginal deterrence", see e.g., Stigler (1970) and Mookherjee and Png (1992).}
Information structure and equilibrium concept Throughout, we assume that, while $G$ has private information concerning his type, the parties are symmetrically informed about all other variables. The above definitions and assumptions apply to both periods of the game, and the two periods differ in only two ways. First, $G$ knows his type at the beginning of the second period because he has learned it in the first period, and second, while the first-period belief equals $h$, based on the observed reporting behavior in the first period, $B$ might hold a belief $\beta \neq h$ at the beginning of the second period. To solve this game of incomplete information, we focus on pure-strategy Perfect Bayesian equilibria that are robust with respect to the Intuitive Criterion as proposed by Cho and Kreps (1987).

4 Equilibrium Analysis

4.1 Static Problem

In this section, we derive the properties of all potential period 2 equilibrium strategies. Below, we show that given our assumptions the last period of the game can be solved by backwards induction because the circularity between equilibrium strategies and equilibrium beliefs normally present in dynamic games of incomplete information is not an issue. As a consequence, the period 2 equilibrium outcome is identical to the outcome of a static version of the model, where the stage game is only played once. Note that in the following we omit the time subscript $t = 2$ for ease of notation.

Since at the end of the game, $G$ does no longer have to worry about his reputation, optimality of his strategy implies that he will report whenever he expects a positive net benefit from doing so. This implies that (with the exception of cases of indifference) the equilibrium reporting strategies $R^* (m; \theta)$ of both types $\theta$ of $G$ only depend on $m$ (and not on other parts of the history of the game).\textsuperscript{21} This leads to the following result:

Lemma 1 (reporting strategies in the static case) In period 2, type $H$ reports whenever misbehavior occurs, while type $D$ does so only if the level of misbehavior is sufficiently

\textsuperscript{21}In the following we proceed in a similar manner and include only those parts of the history as arguments in the equilibrium strategies that might have a non-trivial impact.
large, i.e., \( R^*(m; H) = 1 \ \forall m > 0 \) and \( R^*(m; D) = 1 \Leftrightarrow m > \tilde{m} \), where \( \tilde{m} \) is implicitly defined by \( r(\tilde{m}) \equiv \tau \), and \( \tilde{m} > 0 \).

To rule out uninteresting cases, in the following we assume that there exists sufficiently large levels of \( m \) for which type \( D \) reports (i.e., \( \tilde{m} < \overline{m} \)). When determining the optimal level of misbehavior, \( B \) takes \( G \)'s subsequent reporting strategy into account.\(^{22}\) This implies that in equilibrium, the period 2 level of misbehavior depends only on \( B \)'s belief \( \beta \) at this point in time to face type \( H \). It follows that the level of misbehavior optimally chosen is given by

\[
m^* \equiv \arg\max_m \{ b(m) - ER^*(m, \beta) \cdot p(m) \}, \tag{1}
\]

where \( ER^*(m, \beta) \equiv \beta \cdot R^*(m; H) + (1 - \beta) \cdot R^*(m; D) \) denotes the expected reporting decision given a belief \( \beta \) to face type \( H \). Denote the unique solution to (1) as a function of \( \beta \) by \( m^*(\beta) \). Figure 2 below illustrates the optimal level of misbehavior chosen by \( B \).

**Lemma 2 (misbehavior in the static case)** The optimal period 2 level of misbehavior does not exceed \( \tilde{m} \). In particular, \( m^*(0) = \tilde{m} \), \( m^*(1) = 0 \), and \( m^*(\beta) \) is weakly decreasing in \( \beta \).

If \( B \) is certain to face type \( D \) he chooses the maximal level of \( m \) for which no reporting occurs. If \( B \) is certain to face type \( H \), he chooses \( m = 0 \) because misbehavior does not pay in this case (recall that \( p' > b' \)). Finally, as type \( H \) always reports, a higher probability to face this type induces \( B \) to choose a lower level of misbehavior.

That only type \( H \) does report in equilibrium implies that there is a (partial) wall of silence in the static case because misbehavior is only reported with probability \( h \). While, given that type \( D \) faces reporting costs \( \tau \), this result is not surprising, we show in the next section that in a dynamic setup a wall of silence, where neither type reports, may emerge due to reputational concerns.

Finally, we turn to team formation. Whether the parties are indeed willing to form a team, (in principle) depends on the subsequent level of misbehavior by \( B \) and the resulting reporting behavior of \( G \). For the next section, where we consider a dynamic setup, it turns

\footnote{Note that if in cases of indifference \( G \) would report, equilibria might fail to exist.}
out to be instructive to distinguish two cases: a *symmetric team case*, where cooperation is sufficiently attractive for both parties (such that the team is always formed), and an *asymmetric team case*, where it depends on the anticipated behavior of the parties within a team whether they indeed decide to cooperate. In the dynamic setup, we are mainly interested in the reporting behavior of the honest $G$ (which is potentially driven by $B$’s subsequent willingness to cooperate with him). Consequently, we assume that $G$ always prefers to be part of the team, and vary $B$’s cooperation benefit to distinguish the two cases.

**Assumption 1 (G’s benefit from cooperation)** *Cooperation is sufficiently attractive for party $G$, i.e., $g^c > \overline{m} + g^o$.*

Assumption 1 implies that either type of $G$ prefers cooperation with $B$ over being on his own independent of the belief of $B$. Hence, whenever $B$ proposes to form a team, both types of $G$ accept, which implies that in equilibrium $G$’s team formation decision has no effect on the belief held by $B$.

Now consider $B$’s team formation decision. Note that $B$’s payoff inside the team is given by $b^c + [b(m^*(\beta)) - \beta \cdot p(m^*(\beta))]$, while his outside option is given by $b^o$. As $B$ always has the option not to misbehave, it follows that the term in square brackets is non-negative. Hence, if $b^c \geq b^o$ (the *symmetric case*), $B$ will always want to form a team. On the other hand, if $b^c < b^o$ (the *asymmetric case*), this does not necessarily hold true: party $B$ will only propose to form a team if his belief $\beta$ to face type $H$ is sufficiently low such that the term in square brackets is non-negative.
brackets above is sufficiently large. For this asymmetric case it is useful to implicitly define
a critical value $\tilde{\beta}$ by $b^o - b^c = b(m^*(\tilde{\beta})) - \tilde{\beta} \cdot p(m^*(\tilde{\beta}))$, where $\tilde{\beta} < 1$ holds.\(^{23}\) This leads to the following result:

**Lemma 3 (team formation in the static case)** *In equilibrium, each type of $G$ accepts the offer by $B$ to cooperate ($T^{G*}(H) = T^{G*}(D) = 1$), while $B$’s optimal team formation decision is given by $T^{B*}(\beta) = 0$ if $b^c - b^o < 0$ and $\beta > \tilde{\beta}$, and $T^{B*}(\beta) = 1$ otherwise.*

Lemma 3 implies that $B$ might choose not to offer to cooperate with $G$ when both his outside option and his belief that he faces type $H$ are sufficiently large (a larger probability to face type $H$ reduces $B$’s profit arising within a team). In order to avoid trivial outcomes, we assume that the critical value $\tilde{\beta}$ is sufficiently large such that $B$ offers to cooperate given the prior belief $h$. Formally, this amounts to $\tilde{\beta} > h$ if $b^c < b^o$.\(^{24}\)

For a given belief $\beta \in [0, 1]$ at the beginning of period 2, the period 2 equilibrium outcome is unique and described by Lemmata 1, 2 and 3. This equilibrium outcome would also obtain in a static, one-shot version of the present game, where the stage game is only played once. In particular, in this static case $\beta = h$ holds, and the static equilibrium outcome is given by (under slight abuse of notation)

$$\{ T^{B*} = T^{G*} = 1, \ m^*(h), \ R^*(m^*(h); D) = 0, \ R^*(m^*(h); H) = 1 \}. \quad (2)$$

In the next section we turn to a dynamic version of the game and show how in equilibrium, a wall of silence may be set up by both types.

### 4.2 Dynamic Problem

In the dynamic case, $G$ may potentially signal his type through his first period reporting decision, and hence reputational concerns might influence his willingness to cooperate with

\(^{23}\)The fact that the equilibrium payoff of $B$ given that a team is formed is decreasing in $\beta$ is obvious for all $\beta$ such that $m^*(\beta) = \tilde{m}$. For all other values of $\beta$ this relationship follows from the Envelope-Theorem. If a $\tilde{\beta}$ satisfying the above equality fails to exist, we have $T^{B*}(\beta) = 0$ for all $\beta$.

\(^{24}\)If this assumption is violated, then in the asymmetric team case there is a unique equilibrium outcome where the parties prefer to exercise their outside options in both periods.
the authorities.\textsuperscript{25} In the following, we speak of a \textit{separating equilibrium} if the parties cooperate in period 1 and \( R_1^*(m_1^*; H) \neq R_1^*(m_1^*; D) \), and of a \textit{pooling equilibrium} otherwise. In a separating equilibrium, at the beginning of period 2 \( B \) knows which type of \( G \) he faces, whereas in a pooling equilibrium \( B \) receives no additional information through the period 1 reporting decision. Therefore, in a pooling equilibrium \( B \)'s belief \( \beta \) at the beginning of period 2 has to equal the prior belief \( h \).

**Separating equilibria** In a first step, we show that separating equilibria fail to exist, and hence in any equilibrium, \( B \) cannot distinguish between the two types at the beginning of period 2.

**Proposition 1 (no separating equilibria)** In any equilibrium it holds that \( R_1^*(m_1; H) = R_1^*(m_1; D) \) \( \forall m_1 \), i.e., separating equilibria fail to exist.

To see the intuition behind this result note that in a separating equilibrium, both types of \( G \) would make different reporting decisions in the first period, implying that both possible actions \( R_1 \in \{0, 1\} \) would be on the equilibrium path. Suppose, for example, that for a given level \( m_1 \) of first-period misbehavior only type \( H \) (but not type \( D \)) is supposed to report. The consistency requirement for the beliefs of \( B \) at the beginning of period 2 implies \( \beta = 1 \) if \( G \) has reported, and \( \beta = 0 \) otherwise. There is no leeway in forming off-equilibrium beliefs because both possible actions by \( G \) are on the equilibrium path. Any other beliefs would conflict with the consistency requirement for the beliefs in a PBE. It then follows that in each candidate separating equilibrium, one type has an incentive to deviate. First, consider the case of a symmetric team where the team is always formed independent of \( B \)'s belief (see Lemma 3). If type \( H \) is supposed to report on the equilibrium path, type \( D \) has an incentive to report as well because the resulting reduction in the second period level of misbehavior would outweigh his first period reporting costs. Second, in the case of an asymmetric team \( B \) would not cooperate with type \( H \) which induces the latter to refrain from reporting.

\textsuperscript{25}In reality it may sometimes happen that upon finding (sufficiently large) misbehavior the authorities effectively rule out further (second period) interaction with a black sheep \( B \) (e.g., if as a consequence of a conviction \( B \) is fired or, in the case of a doctor, he loses his licensure). In this case, our model nevertheless applies if one assumes that there are other team members (such as other colleagues) who observe the first period interaction between \( G \) and \( B \) and with whom \( G \) may want to interact in the second period.
Pooling equilibria  In a next step, we consider pooling equilibria. In order to economize on notation, the period 1 reporting decision in a pooling equilibrium is denoted by \(R^*_1(m_1)\). Note that because \(\beta = h\) in any candidate pooling equilibrium, the unique period 2 equilibrium outcome is given by (2). An important preliminary step to identify pooling equilibria is to characterize under which circumstances \(R^*_1(m_1) = 0\) and \(R^*_1(m_1) = 1\), respectively, are consistent with equilibrium. As has been argued above, in the present framework the One-Deviation Principle applies, and hence only simple deviations from the candidate period 1 reporting strategies need to be considered. This observation allows to derive the following result.

**Lemma 4 (only one type is relevant)** Independent of off-equilibrium beliefs, \(R^*_1(m_1) = 0 (R^*_1(m_1) = 1)\) is consistent with equilibrium if and only if type \(H\) (type \(D\)) has no incentive to deviate.

To illustrate the intuition behind Lemma 4 suppose that the equilibrium strategies prescribe \(R^*_1(m_1) = 1\), and consider a deviation to non-reporting. In the first period, (relative to type \(H\)) type \(D\) saves reporting costs \(\tau\). In the second period (relative to type \(D\)) type \(H\) obtains a reporting benefit \([r(m^*(\beta)) - r(m^*(h))]\) that is smaller than \(\tau\).\(^{26}\) Hence, type \(D\) has a larger incentive to deviate. A similar logic applies to the case \(R^*_1(m_1) = 0\).

**Off-equilibrium beliefs** We now briefly turn to the issue of off-equilibrium beliefs in pooling equilibria. Given that for a certain \(m_1\) the equilibrium period 1 reporting strategy prescribes \(R^*_1(m_1) = 1\), denote the off-equilibrium belief following a deviation to non-reporting by \(\beta^1(m_1)\). Analogously, when the equilibrium strategy prescribes \(R^*_1(m_1) = 0\), denote the belief following a deviation to reporting by \(\beta^0(m_1)\). At the outset the concept of Perfect Bayesian equilibrium does not impose any restrictions on the off-equilibrium beliefs party \(B\) may hold in a pooling equilibrium. If the Intuitive Criterion has bite, it follows from Lemma 4 that \(\beta^0(m_1) = 1\) respectively \(\beta^1(m_1) = 0\) has to hold because in the former (latter) case a deviation is potentially more profitable for type \(H\) (\(D\)).\(^{27}\) In many cases, however, the Intu-

---

\(^{26}\)This follows from the fact that both \(m^*(\beta)\) and \(m^*(h)\) are not larger than \(\bar{m}\) (see Lemma 2).

\(^{27}\)For example, in the case \(R^*_1(m_1) = 0\) the Intuitive Criterion has bite if, given that \(B\) holds the most favorable beliefs, a deviation would be profitable for type \(H\) but not for type \(D\).
itive Criterion will have no bite, but nevertheless in light of Lemma 4 certain off-equilibrium beliefs seem to be implausible. In order to ensure that our results do not rely on potentially unrealistic off-equilibrium beliefs, we impose an additional requirement. Suppose that in a pooling equilibrium for a given $m_1$ both types are supposed to report ($R_1^*(m_1) = 1$) but no reporting occurs resulting in an off-equilibrium belief $\beta_1(m_1)$. If $R_1^*(m_1) = 1$ is indeed an equilibrium strategy, both types of $G$ would lose through such a deviation. However, Lemma 4 implies that this loss would always be larger for type $H$. Hence, it would seem plausible that in this case $B$ does not update in the direction of type $H$. An analogous argument applies to the case $R_1^*(m_1) = 0$. Consequently, it seems natural to impose the following restriction on off-equilibrium beliefs:\footnote{Assumption 2 is weaker than Cho and Kreps' (1987) D1 criterion, which in effect insists that a deviation has to be attributed to the type who has the highest incentive to deviate. However, note that even if we impose the stronger assumption $\beta_0(m_1) = 1$ and $\beta_1(m_1) = 0$, this would not alter our results in any way.}

**Assumption 2 (off-equilibrium beliefs)** $\beta_0(m_1) > h > \beta_1(m_1)$ for all $m_1$.

Our results for the asymmetric case (see Section 4.2.2) would not change at all if Assumption 2 were not imposed. In the symmetric case (see Section 4.2.1), however, in the absence of Assumption 2, there might exist additional "wall of silence"-equilibria where the first period level of misbehavior is strictly positive, but neither type reports to the authorities.\footnote{It can be shown that a necessary and sufficient condition for the existence of such equilibria (that survive the Intuitive Criterion) is given by $\bar{m} - \tau \geq \tau - r(m(h))$. This condition is, for example, satisfied if $\tau \leq \frac{1}{2}\bar{m}$ holds, i.e., if type $D$’s cost of reporting is not too high.} Although these equilibria survive the Intuitive Criterion, they involve off-equilibrium beliefs that, in light of Lemma 4, may be questionable. By imposing Assumption 2, we thus make it more difficult to support equilibria that exhibit a wall of silence.

In the following two subsections we describe the equilibrium outcomes in the symmetric case and the asymmetric case, respectively, where the main difference between these cases is that in the former the team will always be formed, while in the latter this is not necessarily the case.
4.2.1 The Symmetric Case ($b^c > b^o$)

When the cooperation benefits of both, $G$ and $B$ are sufficiently large, $R_1^*(m_1) = 0$ can never be consistent with equilibrium because $b^c > b^o$ implies that there is always cooperation in period 2. Hence, deviating from $R_1^*(m_1) = 0$ would always be profitable because it would lead to a smaller level of misbehavior in period 2. Second, $R_1^*(m_1) = 1$ is consistent with equilibrium as long as type $D$ wants to conceal his type and has no incentive to deviate: on the equilibrium path type $D$ would derive $r(m_1) - \tau$ from reporting in period 1, and face a level of misbehavior $m^*(h)$ in period 2. By deviating he would forego $r(m_1) - \tau$ and face a higher level of misbehavior $m^*(\beta^1(m_1))$ instead of $m^*(h)$. Hence, type $D$ has no incentive to deviate as long as

$$r(m_1) - \tau - m^*(h) \geq -m^*(\beta^1(m_1)) \iff r(m_1) + m^*(\beta^1(m_1)) - m^*(h) - \tau \geq 0, \quad (3)$$

i.e., as long as the utility loss due to the higher level of misbehavior is larger than the reporting cost. If (3) holds, reporting is a ”credible threat” for either type. In this case it is optimal for $B$ to choose $m_1 = 0$ because misbehavior would be reported with certainty. As $B$ still gains $b^c$ from cooperating with $G$, he nevertheless prefers this outcome to his (low) outside option $b^o$. Off-equilibrium beliefs $\beta^1(m_1)$, such that (3) is satisfied, exist if the prior belief to face type $H$ is sufficiently large.30

**Proposition 2 (symmetric case)** In the symmetric case,

(i) equilibria exist if the prior belief to face type $H$ is sufficiently large,

(ii) in all equilibria, any level of misbehavior will be reported by either type

(i.e., $R_1^*(m_1) = 1$ for all $m_1$), the team is formed (i.e., $T_1^{B*} = T_1^{G*} = 1$),

and $B$ chooses not to misbehave (i.e., $m_1^* = 0$).

Intuitively, when cooperation is sufficiently important for the potential ”black sheep” $B$, he is disciplined by the uncompromising reporting behavior of $G$, and chooses not to misbehave.

30To see this, note that $r(m_1) + m^*(\beta^1(m_1)) - m^*(h) - \tau \geq m^*(\beta^1(m_1)) - m^*(h) - \tau$. Now, suppose that $h$ is sufficiently large and that $\beta^1(m_1) = 0$ holds. In this case $m^*(\beta^1(m_1)) - m^*(h) - \tau \geq 0$ simplifies to $\tilde{m} \geq m^*(h) + \tau$ which is satisfied for sufficiently large $h$, because we have $m^*(1) = 0$ and $\tilde{m} \geq \tau$ (due to $r(\tilde{m}) = \tau$ and $r' < 1$).
4.2.2 The Asymmetric Case \((b^c < b^o)\)

We now turn to the case of asymmetric teams where \(B\)’s outside option \(b^o\) is assumed to be relatively attractive, such that \(b^o > b^c\) holds. For example consider the case where \(G\) is junior who is eager to gain experience while \(B\) is more senior, so that the \(G\)’s benefit from working with \(B\) is much larger than vice versa.

Obviously, if \(b^o\) is too large relative to \(b^c\), in equilibrium \(B\) will always decline to cooperate in period 1. In order to rule out this uninteresting case, in the following we focus on settings where \(b^o\) is not too large, i.e., where \(b^o < b^c + b(m)\) holds.

We now show that in the asymmetric team case two types of period 1 equilibrium outcomes are possible. In a first class of equilibria (i) the parties cooperate, (ii) the level of misbehavior is strictly positive, but (iii) reporting does not occur in equilibrium, i.e., there is a wall of silence. It is shown that such equilibria always exist. In a second class of equilibria (that may also exist) \(B\) and \(G\) fail to cooperate. However, if both types of equilibria exist simultaneously, equilibria of the first type payoff-dominate the equilibria of the second type. Hence, even if other equilibria exist it seems plausible that the parties will coordinate on a (payoff-dominant) “wall of silence” outcome.

To prove these claims, in a first step, we will now characterize which period 1 reporting behavior is consistent with equilibrium.

**Proposition 3 (reporting in asymmetric teams)** In the asymmetric case,

(i) for all \(m_1\) there exist off-equilibrium beliefs \(\beta^0(m_1)\) such that neither type of \(G\) has an incentive to deviate from \(R^*_1(m_1) = 0\),

(ii) for a given \(m_1\) there exist off-equilibrium beliefs \(\beta^1(m_1)\) such that neither type of \(G\) has an incentive to deviate from \(R^*_1(m_1) = 1\) if \(r(m_1) - \tau + \tilde{m} - m^*(h) \geq 0\).

Intuitively, for a given \(m_1\) where the equilibrium strategies prescribe \(R^*_1(m_1) = 0\), type \(H\) can only be prevented from deviating if the future loss is sufficiently high. Only for off-equilibrium beliefs above the threshold \(\tilde{\beta}\) this is the case because such beliefs induce \(B\) to reject cooperation in period 2. For a given prior \(h\), \(R^*_1(m_1) = 1\) is only consistent with equilibrium for sufficiently high levels of \(m_1\) (see the discussion above Proposition 2).

Proposition 3 implies that while for certain \(m_1\) the equilibrium strategies might require both
types to report, for other levels of misbehavior the equilibrium strategies might prescribe non-reporting.

Now consider $B$’s optimal choice of $m_1$. For given equilibrium reporting strategies $R^*_1(m_1) \in \{0, 1\}$, $B$ optimally chooses the largest level of misbehavior for which reporting does not occur. Hence, in any equilibrium the maximizer $m^*_1 = \max\{m_1 \mid R^*_1(m_1) = 0\}$ must be well defined, which implies $R^*_1(m^*_1) = 0$ for any $m^*_1 > 0$. Moreover, given $b^o > b^c$, party $B$ will only propose to form a team if $m^*_1 > 0$. That is, if a team is indeed formed, there is both, misbehavior and a complete wall of silence in period 1, where neither type reports in equilibrium. In contrast to the symmetric case, these wall of silence equilibria survive a strong requirement on the off-equilibrium beliefs. In particular, it follows from Proposition 3(i) that there always exists an equilibrium where the parties cooperate, but $m^*_1 = \overline{m}$ and $R^*_1(\overline{m}) = 0$. That is, the maximum level of misbehavior $\overline{m}$ is chosen, but reporting does not occur. If the equilibrium reporting strategies are such that the resulting $m^*_1$ is relatively low, the parties will not cooperate in equilibrium. Such non-cooperation equilibria are, however, necessarily payoff-inferior because in the cooperation equilibria, non-cooperation would have been an option for both parties. The discussion above is summarized in the following proposition.

**Proposition 4 (asymmetric case)** In the asymmetric case,

(i) in any equilibrium where a team is formed there is a strictly positive first period level of misbehavior $m^*_1 > 0$ accompanied by a wall of silence (i.e., $R^*_1(m^*_1) = 0$). Equilibria of this kind always exist. In particular, there always exists an equilibrium where $m^*_1 = \overline{m}$ and $R^*_1(\overline{m}) = 0$, and

(ii) there might exist additional equilibria where the parties choose not to cooperate in period 1, but such equilibria are payoff-dominated.

The discussion above (and in particular Proposition 3) implies that the wall of silence equilibria identified in Proposition 4 differ only with respect to the level of period 1 misbehavior $m^*_1$.

\[^{31}\text{To see this, recall that all pooling equilibria result in the same period 2 equilibrium outcome.}\]
least arise in any wall of silence equilibrium. In particular, it is of interest how this lower bound \( \tilde{m}_1^* \) varies with parameters of the model. This threshold value is also interesting for a second reason. Observe that as \( G \)'s equilibrium payoff is decreasing in \( m_1^* \), it might seem plausible that \( G \) selects reporting strategies that yield him a maximal payoff, in which case the equilibrium level of misbehavior would indeed be given by \( \tilde{m}_1^* \).

The exact value of the lower bound on the equilibrium level of misbehavior is determined by two forces. First, Proposition 3(ii) implies that \( R^*(m_1) = 1 \) is consistent with equilibrium only for values of \( m_1 \) equal to or above some threshold level \( \tilde{m}_1^R \) that is implicitly defined by

\[
 r(\tilde{m}_1^R) - \tau + \tilde{m} - m^*(h) = 0. \tag{4}
\]

Hence, in any equilibrium it must be the case that any level of misbehavior below \( \tilde{m}_1^R \) would not be reported by either type. In turn, this implies that it can never be optimal for \( B \) to choose some \( m_1 < \tilde{m}_1^R \). To see this, note that if instead \( B \) were to choose some \( m_1 + \epsilon \) for some small \( \epsilon \), he would still not be reported and earn a higher payoff contradicting the presumption that \( m_1 \) was optimal. Second, it is clear from Proposition 4 that if he is to cooperate, \( B \)'s equilibrium payoff must be larger than his outside option, which is only the case if he gets away with at least some misbehavior. A threshold value \( \tilde{m}_1^T \) for the level of misbehavior such that team formation occurs is implicitly defined by \( b^c + b(\tilde{m}_1^T) = b^o \). To summarize, the above two observations imply \( \tilde{m}_1^* = \max\{\tilde{m}_1^R, \tilde{m}_1^T\} \), where \( \tilde{m}_1^* \) is strictly interior.\(^{32}\)

It is now possible to make a number of interesting observations and to conduct some instructive comparative static exercises. For the moment, consider the case that \( B \)'s outside option and cooperation benefit are relatively similar (i.e., the difference \( b^o - b^c \) is relatively small). First, if there is a sufficient amount of uncertainty in the sense that \( h \) is not too large, non-cooperation equilibria (as described in Proposition 4(ii)) fail to exist because \( \tilde{m}_1^T \leq \tilde{m}_1^R \) holds, and as a consequence there is always misbehavior combined with a wall of silence.\(^{32}\)

\(^{32}\)The fact that we have \( 0 < \tilde{m}_1^* < m \) follows from two observations. First, we have \( 0 < \tilde{m}_1^T < m \), where the first inequality follows from \( b^c < b^o \) (the asymmetric case) and where the second inequality follows from \( b^c + b(m) > b^o \). Second, we have \( 0 < \tilde{m}_1^R < m \), i.e., condition (4) is satisfied for \( m_1 = m \) (because \( r(m) > \tau \) and \( \tilde{m} \geq m^*(h) \)) and it is satisfied for \( m_1 = 0 \) iff \( h \) is sufficiently large.
silence in this case. Second, it is interesting to note that in the present case Proposition 4 even applies when $h$ is relatively large.\footnote{Note that $b^o \approx b^c$ implies that the critical belief level $\bar{\beta}$ above which cooperation does not occur is close to one, hence our assumption $h < \bar{\beta}$ is satisfied even for large values of $h$.} That is, even if $B$ is relatively certain to face the conscientious type $H$, a wall of silence-equilibrium with $m^*_1 = \overline{m}$ might emerge. Third, if $(b^o - b^c)$ is relatively small, $\tilde{m}^T_1$ is close to zero, and hence it is likely that we have $\tilde{m}^*_1 = \tilde{m}^R_1$, i.e., that the lower bound to the equilibrium level of misbehavior is determined by the reporting incentive constraint of the opportunistic type $D$. For the moment, assume that this is the case. As $\tilde{m}^R_1$ is decreasing in $h$, it follows that $\tilde{m}^*_1$ is decreasing in $h$ up to the point where the participation constraint of $B$ becomes binding, after which $\tilde{m}^*_1$ is independent of $h$. Furthermore, the implicit function theorem implies that $\tilde{m}^R_1$ is decreasing in reporting costs $\tau$ (see condition (4) above). Intuitively, a larger $\tau$ has both a direct and an indirect effect. Directly, it makes it more costly for type $D$ to report. Indirectly, it makes it also more costly for him to deviate to $R = 0$ because this would lead to a higher level of misbehavior $\tilde{m}$ in period 2. However, it is straightforward to show that the second effect dominates so that the incentive constraint (4) is relaxed and $\tilde{m}^R_1$ decreases when $\tau$ increases.\footnote{In order to see this, recall that $\tilde{m} \equiv r^{-1}(\tau)$ and note that $0 < r' < 1$ implies $r^{-1} r' > 1$.} Interestingly, interpreting $\tau$ as a measure of heterogeneity between types, this result suggests that the level of misbehavior in period 1 tends to be the higher, the lower the degree of heterogeneity. Fourth, as $\tilde{m}^T_1$ is increasing in $(b^o - b^c)$, the larger $B$’s outside option $b^o$ is relative to his cooperation benefit $b^c$, the more likely is it that $B$’s team formation decision defines the lower bound on the equilibrium level of misbehavior. In this case, the lower bound on equilibrium misbehavior is weakly increasing in $(b^o - b^c)$. Finally, it is also interesting to consider parameter changes that lead to a switch from the symmetric to the asymmetric case. As for $B$’s outside option, as long as $b^o$ is low enough such that one is in the symmetric case, there is cooperation and no misbehavior in period 1 (see Proposition 2). But as $b^o$ becomes larger such that one moves to the asymmetric case, in equilibrium there is either a discrete jump to a strictly positive level of misbehavior (and possibly even to $\overline{m}$) and a wall of silence is erected (or cooperation breaks down altogether).
5 Conclusion

In this paper we aim at exploring the interplay between the behavior of black sheep (i.e., members of a team engaging in activities disliked by their (honest) fellows) and the behavior of honest team members who often fail to report such activities. In our model, such behavior arises as an equilibrium phenomenon: black sheep choose to misbehave, and honest team members set up a wall of silence.

The reason why honest team members set up a wall of silence is that they do not want to forego future benefits from cooperation. The basic mechanism at work is that the reporting decision may convey information about the type of a honest team member. Depending on his own benefit from cooperation, this influences the decision of a potential black sheep to cooperate in the first place. Our analysis suggests that the joint occurrence of misbehavior by black sheep and a wall of silence set up by its team mates seems to be most likely in asymmetric teams where the cooperation benefit is relatively large for the honest team members and relatively small for the potential black sheep.
Appendix

6.1 Proof of Lemma 2

We prove Lemma 2 by proving the following claim:

\[
m^*(\beta) = \begin{cases} 
\tilde{m} & \text{if } b'(\tilde{m}) - \beta \cdot p'(\tilde{m}) \geq 0, \\
\hat{m}(\beta) & \text{if } b'(\hat{m}) - \beta \cdot p'(\hat{m}) < 0 < b'(0) - \beta \cdot p'(0), \text{ and} \\
0 & \text{otherwise},
\end{cases}
\]

(5)

where \(\hat{m}(\beta)\) is implicitly defined by \(b'(\hat{m}) - \beta \cdot p'(\hat{m}) = 0\), and where \(\hat{m}(\beta) \in [0, \bar{m}]\) holds for all \(\beta\). Proof of the claim: Note that \(b''(m) - \beta \cdot p''(m) < 0 \forall m, \beta\) by assumption. For a given \(\beta\), \(B\) may choose some \(m \leq \tilde{m}\) or some \(m > \tilde{m}\). From Lemma 1 it follows that only type \(H\) reports for \(m \leq \tilde{m}\), and that both types report for all \(m > \tilde{m}\). Note that \(m > \tilde{m}\) can never be optimal because \(b(\tilde{m}) - \beta \cdot p(\tilde{m}) > b(\tilde{m}) - p(\tilde{m})\) for all \(\beta < 1\), and \(b'(m) - p'(m) < 0\) by assumption. This observation also implies that \(m = 0\) is optimal for \(\beta = 1\). Therefore, we only need to consider \(m \leq \tilde{m}\) (i.e., values of \(m\) where only type \(H\) reports). If \(b'(\tilde{m}) - \beta \cdot p'(\tilde{m}) \geq 0\), concavity implies that \(b(m) - \beta \cdot p(m)\) is increasing for all \(m \leq \tilde{m}\), and hence \(\tilde{m}\) is optimal. If \(b'(0) - \beta \cdot p'(0) \leq 0\), concavity implies that \(b(m) - \beta \cdot p(m)\) is decreasing for all \(m \leq \tilde{m}\), and hence \(m = 0\) is optimal. If \(b'(0) - \beta \cdot p'(0) > 0 > b'(\tilde{m}) - \beta \cdot p'(\tilde{m})\), concavity and the Intermediate Value Theorem imply that there exist some \(\hat{m}(\beta) \in (0, \tilde{m})\) that solves \(b(\tilde{m}) - \beta \cdot p(\tilde{m}) = 0\). Finally, define a critical value \(\beta^{\text{crit}}\) implicitly by \(b'(\tilde{m}) - \beta^{\text{crit}} \cdot p'(\tilde{m}) = 0\), and note that \(b'(\tilde{m}) - \beta \cdot p'(\tilde{m}) < 0\) is equivalent to \(\beta > \beta^{\text{crit}}\). Hence, for all \(\beta > \beta^{\text{crit}}\) the optimal \(m\) is strictly below \(\tilde{m}\) and decreasing in \(\beta\).

6.2 Proof of Proposition 1

Suppose that in a candidate equilibrium \(R_1^*(m_1; H) \neq R_1^*(m_1; D)\) for some \(m_1 \in [0, \bar{m}]\). In order to prove that such behavior is not consistent with equilibrium it has to be shown that at least one type of \(G\) can gain from deviating.

Case 1 \((b^c - b^o \geq 0)\). First, suppose that \(R_1^*(m_1; H) = 1\) and \(R_1^*(m_1; D) = 0\). In this case the incentive compatibility condition for type \(D\) is given by \(-m_1 + g^c - \tilde{m} \geq -m_1 + b^c - b^o \geq 0\)
\( r(m_1) - \tau + g^e \Leftrightarrow -\tilde{m} \geq r(m_1) - \tau \). If \( m_1 > \tilde{m} \), then \( r(m_1) - \tau > 0 \), and \( D \)'s incentive compatibility condition cannot be satisfied. If \( m_1 \leq \tilde{m} \), then \( r(m_1) - \tau \leq 0 \). Note that \(-\tilde{m} \geq \\) \( r(m_1) - \tau \Leftrightarrow -r(m_1) \geq \tilde{m} - r(\tilde{m}) \). Moreover, \( \tilde{m} - r(\tilde{m}) > 0 \) because \( r(0) = 0 \) and \( r' < 1 \), which again yields a contradiction because \( -r(m_1) \leq 0 \) for all \( m_1 \). Second, suppose that \( R^*_{1}(m_1; H) = 0 \) and \( R^*_{1}(m_1; D) = 1 \). The incentive compatibility condition of type \( H \) is given by \(-m_1 + g^c \geq -m_1 + r(m_1) + g^c - \tilde{m} + r(\tilde{m}) \Leftrightarrow \tilde{m} \geq r(m_1) + \tau \). The incentive compatibility condition of type \( D \) is given by \(-m_1 + r(m_1) - \tau + g^c - \tilde{m} \geq -m_1 + g^c \Leftrightarrow r(m_1) - \tau \geq \tilde{m} \). Hence, if both incentive compatibility conditions were satisfied simultaneously this would imply that \(-\tau \geq \tau \) which is not possible.

**Case 2** \((b^e - b^o < 0)\). First, suppose that \( R^*_{1}(m_1; H) = 1 \) and \( R^*_{1}(m_1; D) = 0 \). In this case the incentive compatibility condition of type \( H \) is given by \(-m_1 + r(m_1) + b^o \geq -m_1 + g^c - \tilde{m} + r(\tilde{m}) \Leftrightarrow 0 \geq \left[ g^c - g^o \right] + \left[ r(\tilde{m}) - \tilde{m} \right] - r(m_1) \), which is violated for all levels of \( m_1 \) if it is violated for \( m_1 = \overline{m} \). This is the case because \( 0 \geq \left[ g^c - g^o \right] + \left[ r(\tilde{m}) - \tilde{m} \right] - r(\overline{m}) \Leftrightarrow 0 \geq \left[ g^c - \overline{m} - g^o \right] + \left[ r(\tilde{m}) - \tilde{m} \right] - \left[ r(\overline{m}) - \overline{m} \right] \) cannot be satisfied due to Assumption 1, \( r' < 1 \) and \( \tilde{m} < \overline{m} \). Second, suppose that \( R^*_{1}(m_1; H) = 0 \) and \( R^*_{1}(m_1; D) = 1 \). In this case the incentive compatibility condition of type \( H \) is given by \(-m_1 + g^o \geq -m_1 + r(m_1) + g^c - \tilde{m} + r(\tilde{m}) \Leftrightarrow 0 \geq \left[ g^c - \tilde{m} - g^o \right] + r(m_1) + r(\tilde{m}) \) which cannot be satisfied due to Assumption 1.

### 6.3 Proof of Lemma 4

As discussed in Section 4.1, despite the fact that we study a framework of incomplete information, in our setup the One-Deviation Principle (see e.g., Fudenberg and Tirole, 1991, p.109) applies. Consequently, in order to verify which period 1 reporting strategies are consistent with equilibrium one only needs to consider deviations from the candidate reporting strategies while the equilibrium continuation in period 2 may be taken as given. In the following, we prove the lemma for the case that both types of \( G \) are supposed not to report (i.e., \( R^*_{1}(m_1) = 0 \)) and \( b^e - b^o \geq 0 \) holds. The proof for the remaining cases is analogous, and therefore omitted. The claim holds if type \( H \) has a larger incentive to deviate than type \( D \). This is the case if the difference between type \( H \)'s candidate equilibrium payoff and his payoff following a deviation, which is given by \( [g^c - m^*(h) + r(m^*(h))] - [r(m_1) + g^c - m^*(\beta) + r(m^*(\beta))] \) is smaller than the difference...
between type D’s candidate equilibrium payoff and his payoff following a deviation, which is
given by \([g^c - m^*(h)] - [r(m_1) - \tau + g^c - m^*(\beta)]\), where \(\beta \in [0, 1]\) denotes the off-equilibrium belief. As \(r(m^*(h)) - \tau - r(m^*(\beta)) \leq r(\tilde{m}) - \tau - r(m^*(\beta)) = -r(m^*(\beta)) \leq 0 \forall \beta\) this is indeed the case.

### 6.4 Proof of Proposition 2

Recall that if the Intuitive Criterion has bite it implies \(\beta^0(m_1) = 1\) respectively \(\beta^1(m_1) = 0\), which will imply that the results derived below are robust to the Intuitive Criterion.

First, Lemma 4 implies that \(R^*_1(m_1) = 0\) is consistent with equilibrium if and only if \(-r(m_1) + [r(m^*(h)) - m^*(h)] - [r(m^*(\beta^0(m_1))) - m^*(\beta^0(m_1))]) \geq 0\), which, however, is violated due to Assumption 2 and \(r' < 1\). Second, Lemma 4 implies that \(R^*_1(m_1) = 1\) is consistent with equilibrium if and only if \(r(m_1) - \tau + m^*(\beta^1(m_1)) - m^*(h) \geq 0\). Note that for all \(m_1\) there exist off-equilibrium beliefs such that this inequality is satisfied, if it can be satisfied for \(m_1 = 0\). It immediately follows from the discussion above Proposition 2 that this is indeed the case if \(h\) is sufficiently large. The period 1 level of misbehavior \(m^*_1\) has to be optimal given the equilibrium reporting strategies and given the equilibrium continuation in period 2. It follows from the reasoning above that in equilibrium the period 1 choice of the level of misbehavior has no impact on B’s period 2 belief, which just equals \(h\). Hence, B chooses the level of misbehavior that maximizes his period 1 payoff, and given that any misbehavior is reported it follows that \(m^*_1 = 0\) is optimal. Finally, given Assumption 1 and \(b^c \geq b^o\) both parties choose to cooperate.

### 6.5 Proof of Proposition 3

Recall that if the Intuitive Criterion has bite it implies \(\beta^0(m_1) = 1\) respectively \(\beta^1(m_1) = 0\), which will imply that the results derived below are robust to the Intuitive Criterion.

First, consider \(R^*_1(m_1) = 0\). Lemma 4 implies that the incentive compatibility condition of type \(H\) is decisive. It follows from the proof of Proposition 2 in Appendix 6.4 that \(R^*_1(m_1) = 0\) is not consistent with equilibrium if \(\beta^0(m_1) \leq \tilde{\beta}\). However, if \(\beta^0(m_1) > \tilde{\beta}\), the parties do not cooperate in period 2, and hence the incentive compatibility condition of type \(H\) is given by
$g - m^*(h) + r(m^*(h)) \geq r(m_1) + g^o \iff [g - g^o] + [r(m^*(h)) - m^*(h)] - r(m_1) \geq 0$. The above inequality is satisfied if it is satisfied for $m_1 = \overline{m}$: $[g - g^o] + [r(m^*(h)) - m^*(h)] - r(\overline{m}) \geq 0 \iff [g - g^o - \overline{m}] + [r(m^*(h)) - m^*(h)] - [r(\overline{m}) - \overline{m}] \geq 0$, which holds due to Assumption 1 and $r' < 1$.

Second, consider $R_1^*(m_1) = 1$. Lemma 4 implies that the incentive compatibility condition of type D is decisive. For a given $m_1$ the proof of Proposition 2 in Appendix 6.4 implies that $R_1^*(m_1) = 1$ is consistent with equilibrium if $r(m_1) - \tau - m^*(h) + m^*(\beta^1(m_1)) \geq 0$. Off-equilibrium beliefs $\beta^1(m_1)$ such that this inequality is satisfied exist if and only if $r(m_1) - \tau - m^*(h) + m^*(0) \geq 0$.

6.6 Proof of Proposition 4

Note that in any equilibrium no additional information regarding the type of G is revealed. Hence, the period 2 equilibrium outcome is independent of the choice of the period 1 equilibrium strategies. In particular, this implies that both the period 1 level of misbehavior and the period 1 cooperation decisions have to maximize period 1 payoffs.

Ad (i): Suppose the candidate equilibrium strategies are such that at date 1 $B$ anticipates that $m_1^* = 0$. In this case his period 1 payoff would be given by $b^c$ which is smaller than $b^o$. Hence, $B$ will choose $T_1^{B^*} = 0$. This proves that in any equilibrium where the parties cooperate $m_1^* > 0$ has to hold. As Proposition 1 shows that only pooled reporting decisions are consistent with equilibrium, if a certain level of $m_1$ is reported it is reported with certainty. Moreover, as $b(m_1) > b(m_1) - p(m_1)$ for all $m_1 > 0$, $B$ will choose the highest level of $m_1$ such that $R_1^*(m_1) = 0$. If such a maximizer fails to exist, the respective candidate reporting strategies cannot be part of an equilibrium. Finally, it immediately follows from Proposition 3(i) that there exists an equilibrium where $m_1^* = \overline{m}$ and $R_1^*(\overline{m}) = 0$. In such an equilibrium both parties want to cooperate. G wants to cooperate due to Assumption 1. B wants to cooperate due to the fact that in equilibrium he is not reported in period 1, and hence gets away with a level of misbehavior $\overline{m}$ resulting in a period 1 payoff of $b^c + b(\overline{m}) > b^o$.
References


Huck, S., D. Kübler, and J. Weibull (2003): “Social Norms and Optimal Incentives in Firms,” *mimeo, University College London*.


