ELSEVIER

Journal of Economic Behavior & Organization
Vol. 65 (2008) 387–408

JOURNAL OF
Economic Behavior
& Organization

www.elsevier.com/locate/econbase

Black sheep and walls of silence

Gerd Muehlheusser [a], Andreas Roider [b,*]

[a] *Department of Sports Science, University of Bielefeld and IZA, P.O. Box 100131, D-33501 Bielefeld, Germany*
[b] *Wirtschaftspolitische Abteilung, Department of Economics, University of Bonn, IZA and CEPR, Adenauerallee 24-42, 53113 Bonn, Germany*

Received 29 July 2004; accepted 8 November 2005
Available online 12 December 2006

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 teams in which the importance of these benefits varies across team members are especially prone to the above-mentioned phenomenon.

JEL classification: D82; C73

Keywords: Teams; Misbehavior; Wall of silence; Whistle-blowing; Asymmetric information

1. Introduction

1.1. 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 mph), thereby aggressively tailgating a slower car. The driver of that car became so scared by the incident that

* Corresponding author. Tel.: +49 228 739246; fax: +49 228 739221.
E-mail addresses: gerd.muehlheusser@uni-bielefeld.de (G. Muehlheusser), roider@uni-bonn.de (A. Roider).

0167-2681/$ – see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.jebo.2005.11.004

she hit two trees on the roadside after losing control of her vehicle. Both the driver and her 2-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 the 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, 2004).

In this paper, we study two interrelated questions. First, we ask why (otherwise honest) individuals such as the test driver's colleagues might implicitly tolerate certain actions by fellows even if they dislike such behavior? 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.[1]

Apart from the above example, there exist a number of other settings where similar phenomena arise. Most prominently, in police departments, 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 respondents confirmed that they were aware that a code of silence existed, and 46% stated that they had witnessed misconduct but concealed what they knew (see Trautman, 2000). Furthermore, according to the Mollen Commission (1994, p. 51) that investigated police violence in New York "the vast majority of honest police officers still protect the minority of corrupt officers". In a similar vein, 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".[2] Walls of silence also seem to exist in other areas of law enforcement: according to the recent Hagar Report (2004) that 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.[3] Furthermore, in the law literature on labor arbitration it has been noted that "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 [emphasis added]" (Gosline, 1988, p. 45). In addition, it has been suggested that, in the education sector, high school teachers remain silent about blatantly failing colleagues in the classroom (see Los Angeles Daily News, 2001). Moreover, there is systematic evidence that teenagers are reluctant to report misbehavior by their peers; for example, in a survey among 3400 Toronto high school students, Tanner and Wortley (2002) found that more than half of those surveyed did not report to adult authority figures (parents, teachers, or police) after being

[1] It might be less puzzling to observe walls of silence in criminal teams, where all members are misbehaving. Such settings have, for example, been analyzed in the recent literature on leniency programs in antitrust, see for example, Motta and Polo (2003), Feess and Walzl (2004), and also the seminal paper on self-reporting by Kaplow and Shavell (1994).

[2] Moreover, this phenomenon does not seem to be confined to US police departments; see for example, Huberts et al. (2003) and Ekenvall (2003) for evidence on the Netherlands, and on Sweden and Croatia, respectively.

[3] The Committee on Quality of Health Care in America reports that, while the number of deaths in US hospitals due to malpractice is estimated to be up to 98,000 per year (see Kohn et al., 1999), "two thirds of the nation's hospitals haven't reported a single adverse incident involving a physician in the last eight years" (CNN and TIME, 2000). See also the recent book by Gibson and Singh (2003).

victimized, and they conclude that there exists a *teen code of silence*.[4] Finally, in community contexts, individuals are often reluctant to report crimes to the police when the criminal belongs to their community or even family (see e.g., Freeman, 1999; Donohue and Levitt, 2001; Finkelhor and Ormrod, 2001). To summarize, the above discussion suggests that walls of silence are relevant in a variety of contexts and, together with the related misbehavior, are perceived as serious problems.[5]

In this paper, we propose an explanation for walls of silence, where agents do not report out of reputational concerns because they worry about benefits from future cooperation. In the following, this idea is spelled out in more detail. First, there is considerable evidence that potential whistle-blowers worry about the reputational repercussions from cooperating with authorities. For example, violators of the code of silence are often labeled "rats", "snitches", or "squealers" and are no longer respected by their colleagues or fellows (see e.g., Gosline, 1988; Washington Post, 1999). Likewise, in Tanner and Wortley's survey among teenagers, 55% of all respondents identified the potential stigma of being labeled an "informer" as a reason for not reporting. In a similar spirit, in the above-mentioned survey among police officers by Trautman, the fear of being "ostracized" and "blackballed" were the most frequent answers of why police officers would not report misbehavior by colleagues.

Second, individuals often derive substantial (cooperation) benefits from being an *accepted* group member rather than being ostracized. For example, police officers or prison guards need to be backed up in dangerous situations. Hence, it is important for them to have attentive colleagues around, and there is indeed evidence suggesting that whistle-blowers do not always receive maximum backup in such situations. Also, being ostracized might be harmful in terms of career prospects. This has been witnessed by Richard Krupp, a whistle-blower in the California Department of Corrections, who reports: "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 . . ." (Sacramento News & Review, 2004).[6] Finally, ostracism might also be harmful in terms of business prospects. To give just one example, in CNN & TIME (2000), a physician argues that a doctor might easily suffer through ostracism because "other doctors could put him out of business by refusing to refer him patients". Note that cooperation benefits can be expected to differ across individuals. For example, depending on the size of their business or on how many patients they acquire on their own, different physicians might be harmed to a different degree in the case that colleagues refuse to refer them patients. With respect to promotional concerns, being an accepted group member might be considerably more important for juniors than for their more senior colleagues.

The second question we address in this paper is how the prospect of a wall of silence influences the behavior of potential "black sheep" (i.e., team members who may pursue activities that increase

[4] According to Washington Post (1999), "they hate that [misbehavior] but they won't rat on the culprits". For survey evidence on the US, see, e.g., Finkelhor and Ormrod (2001). For a discussion from a law perspective on campus safety, see for example, Epstein (2002).

[5] For example, in the medical sector, the costs from preventable medical error (which are argued to arise partly as a consequence of walls of silence) are estimated to be substantial (see, e.g., Kohn et al., 1999; Gibson and Singh, 2003). Apart from direct negative effects of misbehavior, it is frequently pointed out that walls of silence might undermine public trust in institutions such as the police. Also, the availability of (costly) educational instruments, such as on-site training courses to reduce walls of silence, hints at the relevance of the problem.

[6] In a similar vein, Gibson and Singh (p. 137) report the case of a physician who concedes (after having published a study about an unusually high 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".

their own payoff but are disliked by their fellows). For example, anticipating a wall of silence, doctors (teachers) may save on effort costs by not taking appropriate care (effort), thereby causing harm to patients (students). Furthermore, police officers may handle suspects in a manner that, while acceptable to themselves, may be considered unduly harsh or even brutal by others.[7] That is, in the analysis below we endogenize the level of misbehavior, giving rise to some interesting predictions.

1.2. Framework and results

We analyze a model that exhibits the basic features of the above-mentioned examples as equilibrium phenomena. Our aim is to provide conditions under which "black sheep" misbehave and such misbehavior is tolerated by honest team members who set up a wall of silence. In the model, individuals decide whether or not to cooperate and form a team. In order for reputational concerns to have the potential to matter, we assume that honest team members differ with respect to their privately known willingness to report misbehavior. For example, Huberts et al. (2003) and Ekenvall (2003) have investigated police officers' perceptions about their fellows' willingness to report. Interestingly, they find that while most officers disapprove of certain forms of misbehavior and think it should be reported, there is considerable *uncertainty* about whether fellows would do the same. In a similar spirit, Joseph McNamara, Hoover Fellow and former San Jose police chief, notes that "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 [emphasis added]" (San Francisco Chronicle, 2003). As a consequence, the reporting decision may convey information about an agent's type, and this might affect his future payoffs. The basic mechanism at work is that, by 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, the anticipation of a wall of silence leads black sheep to misbehave in the first place.

Our results on the existence of walls of silence (combined with potentially high levels of misbehavior) are most robust in cases where teams are *asymmetric* in the sense that, compared to their respective outside options, the benefit from cooperation is relatively high for honest team members and relatively low for black sheep. Intuitively, in this case the potential whistle-blower must provide sufficient incentives for the black sheep to engage in the team: an opportunity to misbehave (at least to some degree) without being reported provides this incentive. On the other hand, if teams are sufficiently *symmetric* in the sense that cooperation is relatively beneficial for both parties, equilibria with walls of silence may arise, but they rely on potentially questionable off-equilibrium beliefs. When ruling out such beliefs, walls of silence cannot be sustained in the symmetric case and, as a consequence, black sheep choose not to misbehave.

The remainder of the paper is structured as follows. In Section 2, we discuss the related literature. The model is introduced in Sections 3 and 4 contains our main results, while Section 5 concludes. All proofs are relegated to Appendix A.

[7] However, it is important to note that while cases of police violence often receive the highest public attention, walls of silence are likely to extend to many other forms of misbehavior. For example, the survey studies on police integrity by Ekenvall (2003) and Huberts et al. (2003) document the phenomenon (to a varying degree) for 11 different forms of misbehavior, of which only one relates to excessive use of force.

2. Relation to the literature

The present paper contributes to previous research on (i) walls of silence, (ii) social norms, and (iii) information transmission in organizations.

First, walls of silence have so far primarily been addressed in the law literature (see e.g., Gosline, 1988; Epstein, 2002) and in some survey studies in criminology and sociology (see e.g., Finkelhor and Ormrod, 2001; Ekenvall, 2003; Huberts et al., 2003). These studies have mainly focused on documenting walls of silence for either various types of misbehavior or across countries. In general, they have not aimed at disentangling potential reasons for this phenomenon, but have more generally alluded to loyalty towards the group.

Such willingness to treat group members favorably is well-known in social psychology, where experimental subjects often exhibit *ingroup bias* even in setups based on the *minimal group paradigm* (Tajfel et al., 1971) where group membership is based on arbitrary characteristics. One prominent explanation of this effect is based on *social identity theory* (Tajfel and Turner, 1986) according to which the favoring of group fellows helps individuals to maintain self-esteem and a positive identity (see e.g., Hewstone et al. (2002) and Mullen et al. (1992) for surveys of this literature).[8] So far, this literature has not focussed on walls of silence, and typically it has considered experiments involving one-sided, one-shot transactions in anonymous environments, the only information being to which group other subjects belong. In contrast, we enquire under which conditions own group members are treated favorably when interaction is non-anonymous, two-sided, and when it is not one-shot so that an individual's behavior may have an impact on her future payoff. While ingroup bias might play a role for the emergence of walls of silence, it is interesting to note that those empirical studies that have explicitly looked into various reasons for the emergence of walls of silence reveal that reputational concerns loom large in the decision not to report misbehavior (see e.g., Trautman, 2000; Tanner and Wortley, 2002).[9]

To the best of our knowledge, there exists only one other formal analysis of walls of silence, namely Benoit and Dubra (2004). They ask "Why Do Good Cops Defend Bad Cops?" and enquire why a majority of agents (including some good ones) would favor the representation of all agents ("the union") to defend misbehaving colleagues *indiscriminately* over employing a *candid* strategy in which the union honestly reports all information it has. In their model, there is a continuum of agents who differ in their (exogenously given) action-type, ranging from "bad" agents who display very inappropriate actions to better agents whose misbehavior is only minor. Their basic idea is that in the case of an indiscriminate strategy a court will tend not to listen to the union's statement because it contains no information. If the court's prior belief of facing a sufficiently good type is high enough, then even some of the relatively good agents prefer the indiscriminate strategy because this reduces their probability of being subject to a type II error (i.e., of being erroneously found to be bad).

Our approach differs in several important respects from Benoit and Dubra. First, in our model the level of misbehavior by an agent is not exogenously given by his type, but arises endogenously from that agent's optimization problem. It seems likely that in a number of settings the level of misbehavior is indeed a choice variable, and consequently, treating the level of misbehavior as

[8] Recently, Hertel and Kerr (2001) have pointed out that the degree of the ingroup bias might depend on the context in which a group operates. In particular, they show that appropriate *priming* of subjects can reduce this bias.

[9] Interestingly, this is the case even in the context of law enforcement or among teenagers, where the threat of physical retaliation might also be of relevance. For example, in Tanner and Wortley's study among teenagers, reputational concerns were cited more often as a reason for non-reporting than the threat of retaliation across all types of misbehavior studied.

endogenous allows us to explore how the prospect of a wall of silence might affect it.[10] Second, while in Benoit and Dubra walls of silence emerge out of the fear of being subject to type II error in the enforcement technology, we show that even in the absence of such errors, walls of silence might arise as a result of the concern for (endogenously arising) cooperation benefits within teams. Benoit and Dubra (p. 787f) do discuss such benefits, but they are not part of their model. Finally, while Benoit and Dubra assume that a majority of agents preferring an indiscriminate union suffices for a (complete) wall of silence to emerge, we focus on an individual agent's incentive for (not) reporting.

While not considering walls of silence, other economics papers have formally analyzed various forms of misbehavior in the context of law enforcement. For example, Dharmapala and Miceli (2003) analyze a related signaling model, where a court has to decide whether to trust evidence that may have been planted by the police. They investigate how warrant requirements and tort liability of officers, respectively, affect officers' behavior and the truth-finding of courts. Similar to our model, separating equilibria fail to exist so that there is no revelation of information in equilibrium.

Second, as a wall of silence might be viewed as a social norm, our paper contributes to a growing economic literature on this topic. While one strand of this literature has explored how the existence of social norms affects economic behavior (for a recent contribution, see e.g., Huck et al., 2003), a second strand has studied how such norms might emerge, seeking formal explanations as to why agents obey certain behavioral rules for which they have no direct preference.[11] A number of these papers also employ signaling approaches, but with different focuses. For example, Bernheim (1994) shows how heterogenous individuals are willing to conform to a single standard when popularity in itself is deemed sufficiently important. Contrary to our paper, the issue of how this willingness can be *exploited* by group members is not addressed. Furthermore, Kim and Ryu (2003) study factors that lead to a culture of deviance. In a behavioral framework, Battaglini et al. (2005) analyze the incentives of individuals to interact with peers in order to learn more about their *own* characteristics, such as their willpower.

Finally, our paper could also be seen as a contribution to the literature on factors that hinder the flow of information within organizations. While these papers typically rely on a standard principal-agent framework (see e.g., Levitt and Snyder, 1997), we provide a different rationale based on interaction *between* agents to explain such phenomena.

3. The model

We consider two risk-neutral individuals, B and G who 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.

[10] For example, police officers would presumably be more reluctant to treat subjects unduly harsh if they knew that their colleagues would not tolerate such behavior.

[11] Postulating that there exists an (ad hoc) social norm not to report on one's fellows would seem to leave open the question why there does not exist an alternative (and probably socially more desirable) norm under which the black sheep are the ones being ostracized rather than the (honest) whistle-blowers (for a similar argument, see Benoit and Dubra).

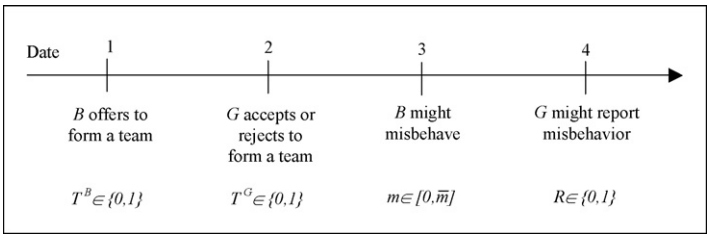

Fig. 1. The stage game.

3.1. 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 consists of four dates (see Fig. 1).

- *Dates 1 and 2 (team formation).* To model the issues of team formation and cooperation as simply as possible, we assume that at date 1, B decides whether to invite G to form a team ($T^B \in \{1, 0\}$). In case an offer has been made, G decides at date 2 whether to accept ($T^G \in \{1, 0\}$).[12] 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, respectively.
- *Date 3 (misbehavior).* Given that a team has been formed, B might choose to "misbehave" by taking an action $m \in [0, \bar{m}]$ that generates a private gain $b(m)$, where b is a strictly 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 to report misbehavior $m > 0$ by B to some authority ($R \in \{1, 0\}$), which may then investigate the case.

3.2. 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. Benoit and Dubra report that such moral considerations indeed constitute a major reason for many (out of the few) police officers who testify against their colleagues. We impose the following three natural properties on $r(m)$: (i) there is no benefit from reporting unless there is misbehavior by B, (ii) this benefit is the higher, the higher the level of misbehavior, and (iii) although there is a benefit from conviction, G prefers lower levels of misbehavior. Formally, this amounts to assuming that $r(0) = 0$, and $0 < r' < 1$. As discussed in Section 1, G's willingness to report is his private information. To model this 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

[12] 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.

"opportunistic" type D faces fixed reporting costs $\tau > 0$. For example, the types might simply face different opportunity costs from filing a report (e.g., testifying in court can be time-consuming). Alternatively, they might differ with respect to their (intrinsic) willingness to cooperate with authorities or to treat team fellows favorably.[13] Thus, in the absence of reputational concerns, type H is readily willing to report any level of misbehavior, while type D is more reluctant to do so: the *net* benefit from reporting differs across types and is given by $r(m)$ for type H and by $r(m) - \tau$ for type D. As a result, type D will only report if the level of misbehavior is sufficiently large. With prior $h \equiv \text{Prob}(\theta = H) > 0$, G is "conscientious" and with probability $(1 - h)$ he is "opportunistic", where h is common knowledge.[14] We assume that G learns his type in the first period after a team has been formed.[15]

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 into expected penalties).[16] 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) the enforcement technology is not subject to type II error (i.e., 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 $p(0) = 0$ and $p' > b'$, and, for technical convenience, we assume $p'' \geq 0$.

3.3. Information structure and equilibrium concept

Throughout, we assume that G has private information concerning his type, but that 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 Perfect Bayesian Equilibria (PBE) that are robust with

[13] For example, in the literature on ingroup bias discussed above, individuals typically differ with respect to the strength of this bias.

[14] One might wonder whether, in practice, team members might not become aware of their fellows' types over time. In this respect, it is interesting to note that even in the police context (where there is the most evidence on walls of silence) turnover can be considerable, probably limiting the above effect. For example, in the San Francisco police department, which has been plagued by a number of adverse incidents, "sixty percent of patrol officers have been on the force less than five years" (San Francisco Chronicle, 2003). Moreover, below we show that a "wall of silence"-outcome might emerge even if the degree of uncertainty is low.

[15] This assumption only serves to simplify the exposition. If G would learn his type already at the beginning of the game, then (in addition to the later reporting decision) the team formation decision might also be used as a signaling device, which makes the analysis much more tedious. However, we have analyzed this alternative timing, and it can be shown that (i) in any equilibrium, the team formation decisions of both types of G must be identical, and (ii) the same team formation decisions as in the case where G learns his type only after date 2 obtain in equilibrium (details are available from the authors upon request). Consequently, while adding considerable formal complexity, this alternative timing would not alter our results in any way.

[16] That is, the authority does not necessarily discover the exact level of misbehavior, and as a consequence there is a possibly stochastic connection between the level of misbehavior and the result of all legal enforcement activities. Given that the penalty function $p(m)$ is exogenous, our results continue to hold if the authority can uncover m.

respect to the Intuitive Criterion as proposed by Cho and Kreps (1987), and restrict attention to pure strategies.

4. Analysis of the model

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 played only 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 no longer has 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 θ of G only depend on the level of misbehavior m (and not on other parts of the history of the game).[17] This leads to the following result.

Lemma 1 (reporting strategies in static case). *In period 2, type H reports whenever misbehavior occurs, while type D does so only if the level of misbehavior is sufficiently large. Formally,* $R^*(m;H)=1\ \forall\ m>0$ *and* $R^*(m;D)=1 \Leftrightarrow m>\tilde{m}$, *where* $\tilde{m}$ *is implicitly defined by* $r(\tilde{m})\equiv\tau$, *and we have* $\tilde{m}>0$.

To rule out uninteresting cases, in the following we assume that there exist sufficiently large levels of m for which type D reports (i.e., $\tilde{m}<\bar{m}$).

When determining the optimal level of misbehavior, B takes G's subsequent reporting strategy into account. This implies that, in equilibrium, the period 2 level of misbehavior depends only on B's belief β of facing type H at this point in time. It follows that B's optimal level of misbehavior as a function of β is given by

$$m^*(\beta)\equiv \underset{m}{\operatorname{argmax}}\{b(m)-[\beta\cdot R^*(m;H)+(1-\beta)\cdot R^*(m;D)]\cdot p(m)\}, \tag{1}$$

where the term in square brackets denotes the expected reporting decision given a belief β of facing type H. We obtain the following result.

Lemma 2 (misbehavior in 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* β.

Fig. 2 serves to illustrate Lemma 2: note that for all levels of misbehavior above $\tilde{m}$, B would be reported by both types (i.e., with certainty), and hence our assumption that $p'>b'$ implies that such high levels of misbehavior cannot be optimal. On the one hand, if B is relatively certain of facing type D (see the left panel), he chooses the maximal level of m for which no reporting occurs. On the other hand, if B's belief of facing type H is sufficiently high (see the right panel), he chooses $m=0$ because misbehavior does not pay in this case. Finally, for intermediate values

[17] In the following we proceed in a similar manner, and only those parts of the history that might have a non-trivial impact are included as arguments in the equilibrium strategies.

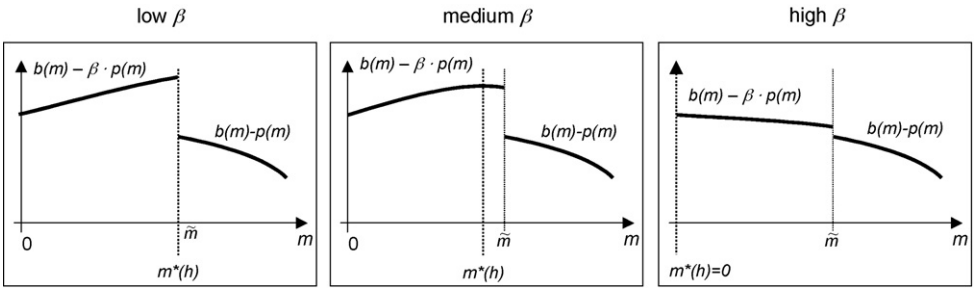

Fig. 2. The optimal level of misbehavior in the static problem.

of β, the optimal level of misbehavior lies between 0 and $\tilde{m}$ and is weakly decreasing in the probability of facing type H. As Lemma 2 implies $m^*(\beta) \leq \tilde{m}$, it follows that only type H reports in equilibrium: there is a (partial) wall of silence in the static case because misbehavior is only reported with probability h. Given that type D faces reporting costs τ, this result is not surprising. In the next section, however, we show how a wall of silence, where neither type reports, may emerge due to reputational concerns in a dynamic setup.

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. It will be useful to distinguish the following 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 cooperation depends on the anticipated behavior of the parties within a team. 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. Formally, $g^c > \bar{m} + g^o$.

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: agent B will only propose to form a team if he expects to derive a strictly positive expected payoff from misbehaving, which can only be the case for m strictly positive. To put it differently, B's belief β of facing type H must be sufficiently low such that the term in square brackets above is sufficiently large. This leads to the following result.

Lemma 3 (team formation in static case). *In equilibrium, each type of G accepts the offer by B to cooperate. In the symmetric case, B offers to cooperate. In the asymmetric case, B offers to cooperate if and only if his belief of facing type H is not too large. Formally,* $T^{G^*}(H) = T^{G^*}(D) = 1$, $T^{B^*}(\beta) = 0$ *if* $b^c - b^o < 0$ *and* $\beta > \tilde{\beta}$, *and* $T^{B^*}(\beta) = 1$ *otherwise, where* $\tilde{\beta}$ *is implicitly defined by* $b^o - b^c = b(m^*(\tilde{\beta})) - \tilde{\beta} \cdot p(m^*(\tilde{\beta}))$, *and where* $\tilde{\beta} < 1$ *holds.*

Lemma 3 implies that B might choose not to offer to cooperate with G when both his outside option and his belief of facing type H are sufficiently large. In order to avoid trivial outcomes, we assume that the critical belief level $\tilde{\beta}$ is sufficiently large such that, given the prior belief h, B offers to cooperate. Formally, this amounts to $\tilde{\beta} > h$ if $b^c < b^o$. If this assumption is violated, then in the asymmetric team case there is a unique equilibrium outcome where the parties exercise their outside options in both periods.

For a given belief $\beta \in [0, 1]$ at the beginning of period 2, the period 2 equilibrium outcome is unique and described by Lemmas 1–3. 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 in the static equilibrium outcome the team is formed, there is misbehavior $m^*(h)$, and only type H reports. Formally (under slight abuse of notation):

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

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 the authorities.[18] In the following, we speak of a *separating equilibrium* if the parties cooperate in period 1 and make different period 1 reporting decisions (i.e., if $R_1^*(m_1^*; H) \neq R_1^*(m_1^*; D)$), and of a *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 β at the beginning of period 2 has to equal the prior belief h.

Separating equilibria. In a first step, we show that 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 for any level of misbehavior both types of G make identical reporting decisions* (*i.e.*, *separating equilibria fail to exist*). *Formally*, $R_1^*(m_1; H) = R_1^*(m_1; D)\, \forall m_1$.

To see the intuition behind this result 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 (i.e., there is no leeway in forming off-equilibrium beliefs). It then follows that there is always one type who 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). In this case, 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. In the remaining case

[18] In reality, upon finding (sufficiently large) misbehavior, authorities might 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.

of a candidate equilibrium where only type D (but not type H) is supposed to report, a similar intuition applies.

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 identifying 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 (Fudenberg and Tirole, 1991) applies, and hence only simple deviations from the candidate period 1 reporting strategies need to be considered. This observation allows us 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 τ. In the second period (relative to type D) type H obtains a reporting benefit $[r(m^*(\beta)) - r(m^*(h))]$ that is smaller than τ. This follows from the fact that both $m^*(\beta)$ and $m^*(h)$ are not larger than $\tilde{m}$ (see Lemma 2). 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 $H(D)$. 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.

In many cases, however, the Intuitive Criterion will have no bite, but nevertheless in light of Lemma 4 certain off-equilibrium beliefs might seem to be less plausible. In order to ensure that our results on the emergence of walls of silence do not depend on potentially questionable 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. 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, one could argue that in this case, B should not update in the direction of type H. An analogous argument applies to the case $R_1^*(m_1) = 0$. Consequently, in order to ensure the robustness of our results, we impose the following additional restriction on off-equilibrium beliefs.[19]

Assumption 2 *(off-equilibrium beliefs).* $\beta^0(m_1) > h > \beta^1(m_1)$ for all m_1.

[19] Assumption 2 is related to Cho and Kreps' D1 criterion, which would in effect require attributing any deviation to the type who has the larger incentive to deviate. However, even if we were to impose the stronger assumption $\beta^0(m_1) = 1$ and $\beta^1(m_1) = 0$, our results would continue to hold. Intuitively, this is because under any of the equilibrium reporting strategies identified below, the respective critical type will prefer the prior belief to having his type revealed.

As will become clear below, in the asymmetric case (Section 4.2.2) our results do not depend on whether or not Assumption 2 is imposed. In the symmetric case (Section 4.2.1), however, "wall of silence"-equilibria (which survive the Intuitive Criterion) are eliminated by Assumption 2. Hence, invoking Assumption 2 makes it somewhat more difficult to sustain equilibria that exhibit a wall of silence.

4.2.1. The symmetric case

When the cooperation benefits of both G and B are sufficiently large (i.e., if $b^c > b^o$), there is always cooperation in period 2. First, Lemma 4 implies that $R_1^*(m_1) = 0$ is consistent with equilibrium if type H has no incentive to deviate. On the one hand, if H sticks to $R_1^*(m_1) = 0$, he faces a level of misbehavior $m^*(h)$ in the second period. On the other hand, if he deviates, he obtains a reporting benefit of $r(m_1)$, but is subject to misbehavior $m^*(\beta^0(m_1))$ in the next period. Hence, taking the reporting benefits in period 2 into account, H's incentive condition is given by

$$-m^*(h) + r(m^*(h)) \geq r(m_1) - m^*(\beta^0(m_1)) + r(m^*(\beta^0(m_1))). \tag{3}$$

It is clear from (3) that only if deviating would lead to a higher level of misbehavior in period 2 might it be the case that H has an incentive to refrain from reporting some positive level of misbehavior in period 1 (recall that $r' < 1$). However, Assumption 2 precludes this case (because it implies $m^*(\beta^0(m_1)) \leq m^*(h)$), and therefore, deviating is profitable for type H.

Second, $R_1^*(m_1) = 1$ is consistent with equilibrium as long as type D 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 (given Assumption 2) 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

$$r(m_1) - \tau - m^*(h) \geq -m^*(\beta^1(m_1)) \Leftrightarrow r(m_1) + m^*(\beta^1(m_1)) - m^*(h) - \tau \geq 0, \tag{4}$$

that is, as long as the utility loss due to the higher level of misbehavior is sufficiently large. Inspecting inequality (4) reveals that this condition is easier to satisfy if Assumption 2 is imposed. If (4) 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. It can be shown that off-equilibrium beliefs $\beta^1(m_1)$, such that (4) is satisfied, exist if the prior belief of facing type H is sufficiently large.

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

(i) *equilibria exist if the prior belief of facing type H is sufficiently large, and*
(ii) *in all equilibria, any level of misbehavior will be reported by either type, the team is formed, and B chooses not to misbehave. Formally,* $R_1^*(m_1) = 1$ *for all* $m_1 > 0$, $T_1^{B*} = T_1^{G*} = 1$, *and* $m_1^* = 0$.

Intuitively, when cooperation is sufficiently important for B, he is disciplined by the uncompromising reporting behavior of G and chooses not to misbehave.[20]

[20] If Assumption 2 is not imposed, G's reporting behavior might not be as uncompromising. It follows from the discussion above that in this case, condition (3) can be satisfied for m_1 not too large, and hence for such levels of period 1

4.2.2. The asymmetric case

We now turn to the case where B's outside option is relatively attractive ($b^o > b^c$) such that he is willing to cooperate with G only if he expects to get away with some strictly positive level of misbehavior. Since, by Assumption 1, G is always willing to cooperate, this case is referred to as *asymmetric*.

For example, one could think of G being either a young or new team member for whom being accepted is very important so that his benefit from cooperation would be large. On the other hand, B is more senior, and working with the unexperienced G may impose some cost on him, thereby making cooperation with G considerably less attractive. As the subsequent analysis suggest, underreporting might be particularly severe for such young (or new) group members.[21] Furthermore, in the decision to enter a certain profession or team, some agents might take for granted certain "fringe benefits" that arise from the possibility to indulge in some form of misbehavior. Indeed, Huberts et al. (p. 226) provide evidence that in a number of countries, it is not uncommon for police officers to consider certain forms of misbehavior (such as accepting gifts and free services) as being part of the standard, informal benefits of being in the force.[22]

We now show that in the asymmetric 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: there is a wall of silence. It is shown that such equilibria always exist. There may exists a second class of equilibria in which 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* $R_1^*(m_1) = 0$, *and*
(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

misbehavior non-reporting by both types would be consistent with equilibrium. Then, in addition to the equilibria identified in Proposition 2, "wall of silence"-equilibria may emerge, in which the level of misbehavior in period 1 is strictly positive and neither type reports. It can be shown that such equilibria exist *and* survive the Intuitive Criterion if and only if $\tilde{m} - \tau \geq \tau - r(m(h))$ holds. This condition is, for example, satisfied if the reporting benefit is sufficiently small (i.e., if $r' \leq 1/2$).

[21] While systematic evidence is still scarce, this observation is consistent with available empirical studies on police officers and teenagers that document a positive correlation between age and reporting rates; see, for example, Ekenvall (2003, p. 226) and Finkelhor and Ormrod (2001, p. 222), respectively. In this respect, it is also interesting to note that in San Francisco's problem-ridden police department (see Footnote 14), the fraction of officers with short tenure has indeed been substantial. Assuming that these officers are often paired with more senior partners, this hints at the existence of asymmetric teams in this example.

[22] Also, consider the following (admittedly stark) statement by a LA police officer, recorded in the course of investigating the Rodney King case in 1991 reported in Chevigny (p. 35): "They give me a stick, they give me a gun, they pay me 50*G*s to have some fun".

beliefs above the threshold $\tilde{\beta}$ is this 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 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 period 1 misbehavior 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. To see this, suppose to the contrary that there exists some $\hat{m}_1 > m_1^*$ that would remain unreported by both types: in this case, by choosing $\hat{m}_1$, B would still not be reported and earn a higher payoff, contradicting the presumption that m_1^* was optimal. Hence, in any equilibrium the maximizer $m_1^* = \max\{m_1 | R_1^*(m_1) = 0\}$ 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 forming 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 are consistent with Assumption 2.[23] In particular, it follows from Proposition 3(i) that there always exists an equilibrium where the parties cooperate, the maximum level of misbehavior $\bar{m}$ is chosen, but reporting does not occur. If the equilibrium reporting strategies are such that the resulting m_1^* would be 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. We summarize this discussion 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* $m_1^* = \bar{m}$ *and* $R_1^*(\bar{m}) = 0$, *and*
(ii) *there might exist additional equilibria where the parties choose not to cooperate in period 1, but such equilibria are payoff-dominated.*

4.3. Implications

In the following, we discuss some implications arising from the above results. First, in the presence of asymmetric information, walls of silence can emerge even in short-lived relationships of known length.[24] Furthermore, since such "wall of silence"-equilibria are particularly robust in the asymmetric case (if possible) such team configurations should be avoided.

Second, it is interesting to note that, even if B is relatively certain to face the conscientious type H, a "wall of silence"-equilibrium exhibiting a potentially high level of misbehavior might arise

[23] Proposition 3 also holds when Assumption 2 is not imposed. This is obvious for part (i). As for part (ii), even without imposing the restriction $\beta^1(m_1) < h$, there exist off-equilibrium beliefs such that the crucial incentive condition of type D is satisfied if and only if it is satisfied for $\beta^1(m_1) = 0$.

[24] Note that while we consider only two periods, we conjecture that the intuition gained in the present paper extends to a multi-period version of the model where equilibria exhibiting misbehavior and walls of silence can be supported in all but the last period; the availability of multiple future opportunities of interaction should make the threat of possible non-cooperation even stronger.

(see Propositions 3 and 4). Intuitively, even given a relatively high prior belief h, a team is formed as long as B's outside option is not too attractive. At the same time, the threat of non-cooperation, which would result if B were sure to face H, is still sufficient to deter both types from reporting. Thus, the existence of walls of silence is not necessarily sensitive to the fraction of agents who would in principle be inclined to report any misbehavior. For example, in the police context this suggests that, even if the majority of police officers are conscientious, high levels of misbehavior and walls of silence might still emerge.

Third, while there exists a continuum of equilibria exhibiting a wall of silence, Proposition 4(i) implies that the respective equilibrium outcomes differ only with respect to the level of misbehavior in period 1. It is therefore instructive to derive a *lower bound* on m_1^* (denoted by $\tilde{m}_1^*$) and to use it as a measure for the *minimum* level of misbehavior to emerge in *any* equilibrium.[25] The lower bound is determined by two forces. First, it follows from the above discussion that m_1^* cannot be smaller than the minimum level for which type D would be willing to report (see Proposition 3(ii)). Second, at the same time, as $b^c < b^o$, B prefers cooperation to his outside option only if he expects to get away with at least some minimum level of misbehavior. In the following, we study in more detail how the lower bound on m_1^* varies with parameters of the model. One interesting aspect will be to consider variations with respect to how much G dislikes misbehavior by B. The simplest way of capturing this is to assume that a level m of misbehavior reduces G's payoff by $s \cdot m$, where $s \in (0, 1]$ is commonly known (i.e., in all of the previous analysis we have assumed $s = 1$). We obtain the following comparative statics result.

Proposition 5 (comparative statics). *The minimum period* 1 *level of misbehavior in any* "*wall of silence*"*-equilibrium is weakly decreasing in h and s, and weakly increasing in the attractiveness of B's outside option relative to cooperation.*

Intuitively, while the latter part of Proposition 5 is driven by B's participation constraint, the results on h and s follow from the critical type D's incentive constraint.

First, consider an increase in h: since only pooling equilibria exist, a higher h leads to a lower level of misbehavior in period 2 (see Lemma 2). In turn, this increases the difference between the level of misbehavior off and on the equilibrium path ($m(\beta^1(m_1)) - m(h) > 0$), and as a consequence, a deviation from reporting becomes less attractive for type D. Consequently, he is willing to report lower levels of misbehavior. Taken together with the discussion above, we conclude that, while a high probability of facing a conscientious type does not preclude the emergence of walls of silence, it is nevertheless helpful in the sense that it makes lower levels of misbehavior sustainable in equilibrium.[26] From a practical point of view, it might, for example, be possible to increase the number of conscientious agents by generating some turnover in combination with a recruiting

[25] Note that the outcome $m_1^* = \tilde{m}_1^*$ and $R_1^*(\tilde{m}_1^*) = 0$ would emerge as the *unique* equilibrium prediction in the asymmetric case if the following two additional assumptions are imposed: (i) in the spirit of Cho and Kreps' D1 criterion, Assumption 2 is strengthened to $\beta^0(m_1) = 1$ and $\beta^1(m_1) = 0$ for all m_1, and (ii) G selects the reporting strategies that maximize his equilibrium payoff. Intuitively, part (i) ensures that $R_1^*(m_1) = 1$ can be supported for levels of m_1 as low as possible. Part (ii) implies that whenever both $R_1^*(m_1) = 1$ and $R_1^*(m_1) = 0$ can be supported, $R_1^*(m_1) = 1$ is selected unless this would lead B not to form the team.

[26] Of course, the equilibrium level of misbehavior does not necessarily decrease when the lower bound does. On the other hand, any factor that leads to an increase in the lower bound eliminates previously feasible equilibria with relatively low levels of misbehavior. In this sense, our results are stronger in one direction (namely, which changes in parameters should be avoided), and it applies that "You can pull a string, but you cannot [necessarily] push it".

policy that aims at hiring agents who tend to be less opportunistic than insiders (and to make this commonly known).[27]

Furthermore, when s increases, B's misbehavior is more strongly disliked by both types of G. Consequently, type D's willingness to report is enhanced, which again relaxes his incentive constraint and makes lower equilibrium levels of misbehavior sustainable. This finding is consistent with results from a comparative survey study by Huberts et al. (p. 225ff) on police integrity in the Netherlands (NL) and in the US. They find that on average, officers in NL consider the *same infractions* to be more serious than their US colleagues (which could be interpreted as s being higher among NL officers). Also, the stated willingness to report is more uncompromising in NL, and problems with police misbehavior and walls of silence are indeed less severe in NL compared to the US. Huberts et al. also point out that this stronger *awareness* of Dutch officers is not exogenously given, but the result of various measures to improve police integrity.[28]

Finally, Proposition 5 suggests that in countries where police officers earn relatively little in comparison to alternative occupations (i.e., where b^o is large), a large fraction of police officers may enter this job only because it enables them to earn certain fringe benefits through misbehavior. Everything else equal, the model suggests that misbehavior and walls of silence should be more prevalent in such countries.

5. Conclusion

This paper explores the interplay between the behavior of *black sheep* (i.e., members of a team engaging in activities disliked by their 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 decide to set up a wall of silence. The reason for doing so is that they worry about their reputation because they do not want to forego future benefits from cooperation. The basic mechanism at work is that the reporting decision may convey information about an honest team member's type. 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 setup by its team mates is most likely in asymmetric teams, where the cooperation benefit is relatively large for honest team members and relatively small for potential black sheep.[29]

At a more general level, our model of walls of silence driven by reputational concerns points to the importance of providing anonymity to potential whistle-blowers. Indeed, ensuring anonymity is considered crucial by many practitioners (see e.g., Trautman, 2000; Spitalli, 2004).[30] In this respect, it is interesting to note that the importance of the provision of anonymity would seem to be less obvious if walls of silence were mainly driven by a direct preference to treat team members favorably. On the other hand, in some settings secrecy will not be feasible or hard to achieve, in

[27] To achieve this aim, practitioners often stress the importance of conducting background investigations on potential candidates, which are considered a good predictor of future employee behavior.

[28] In this spirit, Kübler (2001) analyzes how socially undesirable norms may be successfully changed through instruments such as symbolic acts or educational programs.

[29] Our analysis thus hints at the potential importance of such asymmetries and suggests that special attention should be paid to this issue in future empirical research on walls of silence.

[30] See also the recent Sarbanes-Oxley Act of 2002 (in particular, Section 301), which includes anonymity provisions for employee complaints about questionable accounting practices in publicly traded corporations. Various US state laws (such as the California Whistle-Blower Protection Act) also contain anonymity clauses.

which case it will be difficult to eliminate the reputational considerations that are the focus of the present paper. For example, labor arbitrators will in general not admit anonymous (hearsay) evidence (see e.g., Gosline (1988)). In addition, even if an anonymity policy is nominally in place, information leakage might be hard to prevent in practice if administrative leaders do not pay special attention to the enforcement of such provisions (see e.g., Sacramento News & Review).

Finally, we would like to point out that a full-fledged (normative) analysis of anonymity provisions would have to take into account a further, potentially important dimension, namely the role of trust among individuals in an organization. Active encouragement of whistle-blowing might create an atmosphere of distrust that might adversely affect performance. While such issues are beyond the scope of the present paper, they seem to be an interesting topic for future research.

Acknowledgements

We are grateful to Björn Bartling, Eberhard Feess, Georg Nöldeke, Jörg Oechssler, Stefan Reichelstein, Urs Schweizer, and Ilga Vossen for suggestions, and to two Referees and an Associate Editor for extensive and very helpful comments. We also thank seminar participants at the University of Bonn, Simon Fraser University, the Köln-Bonner Colloquium on Personnel Economics 2004 in Bonn, the Canadian Economic Theory Conference 2005 in Vancouver, the SOLE/EALE World Congress 2005 in San Francisco, the European Economic Association meetings 2005 in Amsterdam, the Verein für Socialpolitik meetings 2005 in Bonn, and the SFB/TR15 Conference 2005 in Tutzing for helpful discussions and suggestions. Both authors gratefully acknowledge financial support from the Graduiertenkolleg "Quantitative Economics" at the University of Bonn. The second author also gratefully acknowledges financial support from the SFB/TR15 at the University of Bonn, a DFG research fellowship, as well as the hospitality of Stanford Graduate School of Business where part of this research was conducted.

Appendix A

A.1. Proof of Lemma 2

We prove Lemma 2 by proving the following claim:

$$m^*(\beta) = \begin{cases} \tilde{m} & \text{if } b'(\tilde{m}) - \beta \cdot p'(\tilde{m}) \geq 0, \\ \hat{m}(\beta) & \text{if } b'(\tilde{m}) - \beta \cdot p'(\tilde{m}) < 0 < b'(0) - \beta \cdot p'(0), \text{ and} \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $\hat{m}(\beta)$ is implicitly defined by $b'(\hat{m}) - \beta \cdot p'(\hat{m}) = 0$, and where $\hat{m}(\beta) \in [0, \tilde{m}]$ holds for all β. *Proof of the claim:* Note that $b''(m) - \beta \cdot p''(m) < 0 \forall m, \beta$ by assumption. For a given β, 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 $b(\hat{m}) - \beta \cdot p(\hat{m}) = 0$. Finally, define a critical value

β^{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 β.

A.2. Proof of Lemma 3

If B chooses $T^B = 0$, the game ends and both parties receive their reservation utilities. Hence, G decides about T^G if and only if $T^B = 1$. It immediately follows from Assumption 1 that both types of G strictly prefer to form the team independent of the belief subsequently held by B. Given this equilibrium continuation, $T^{B*}(\beta)$ immediately follows from the discussion above the lemma. Given that a team is formed, note that the equilibrium payoff of B is decreasing in β is obvious for all β such that $m^*(\beta) = \tilde{m}$. For all other values of β this relationship follows from the Envelope-Theorem. If a $\tilde{\beta}$ satisfying the definition in the Lemma fails to exist, we have $T^{B*}(\beta) = 0$ for all β.

A.3. 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 $-m_1 + g^c - \tilde{m} \geq -m_1 + r(m_1) - \tau + g^c \Leftrightarrow -\tilde{m} \geq r(m_1) - \tau$. 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 *($b^c - b^o < 0$).* 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 $-m_1 + r(m_1) + b^o \geq -m_1 + g^c - \tilde{m} + r(\tilde{m}) \Leftrightarrow 0 \geq [g^c - g^o] + [r(\tilde{m}) - \tilde{m}] - r(m_1)$, which is violated for all levels of m_1 if it is violated for $m_1 = \bar{m}$. This is the case because $0 \geq [g^c - g^o] + [r(\tilde{m}) - \tilde{m}] - r(\bar{m}) \Leftrightarrow 0 \geq [g^c - \bar{m} - g^o] + [r(\tilde{m}) - \tilde{m}] - [r(\bar{m}) - \bar{m}]$ cannot be satisfied due to Assumption 1, $r' < 1$ and $\tilde{m} < \bar{m}$. 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 $-m_1 + g^o \geq -m_1 + r(m_1) + g^c - \tilde{m} + r(\tilde{m}) \Leftrightarrow 0 \geq [g^c - \tilde{m} - g^o] + r(m_1) + r(\tilde{m})$ which cannot be satisfied due to Assumption 1.

A.4. 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

$b^c - b^o \geq 0$ 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 $[g^c - m^*(h) + r(m^*(h))] - [r(m_1) + g^c - m^*(\beta) + r(m^*(\beta))]$ and 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.

A.5. *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 $m_1 = 0$. This is the case if h is sufficiently large. To see this, note that for $m_1 = 0$ and $\beta^1(m_1) = 0$ type D's incentive condition simplifies to $[\tilde{m} - \tau] - m^*(h) \geq 0$, which is satisfied for large h because the term in square brackets is strictly positive. 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.

A.6. *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 A.5 that, given Assumption 2, $R_1^*(m_1) = 0$ 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 $g^c - m^*(h) + r(m^*(h)) \geq r(m_1) + g^o \Leftrightarrow [g^c - g^o] + [r(m^*(h)) - m^*(h)] - r(m_1) \geq 0$. The above inequality is satisfied if it is satisfied for $m_1 = \bar{m} : [g^c - g^o] + [r(m^*(h)) - m^*(h)] - r(\bar{m}) \geq 0 \Leftrightarrow [g^c - g^o - \bar{m}] + [r(m^*(h)) - m^*(h)] - [r(\bar{m}) - \bar{m}] \geq 0$, 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 A.5 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 $r(m_1) - \tau - m^*(h) + m^*(0) \geq 0$.

A.7. *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. 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 $m_1^* = \bar{m}$ and $R_1^*(\bar{m}) = 0$. 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 $\bar{m}$ resulting in a period 1 payoff of $b^c + b(\bar{m}) > b^o$.

A.8. Proof of Proposition 5

To ensure that G still prefers lower levels of misbehavior, one has to assume $r' < s$. In this case, Propositions 3 and 4 continue to hold with the only modification that Proposition 3(ii) now reads: 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 $r(m_1) - \tau + s \cdot \tilde{m} - s \cdot m^*(h) \geq 0$. As discussed above, the exact value of the lower bound $\tilde{m}_1^*$ is determined by two forces. On the one hand, it follows from Proposition 3(ii) that $R^*(m_1) = 1$ cannot be consistent with equilibrium for values of m_1 below some threshold level $\tilde{m}_1^R$ that is implicitly defined by $r(\tilde{m}_1^R) - \tau + s \cdot [\tilde{m} - m^*(h)] = 0$. On the other hand, it is clear from Proposition 4 that B is only willing to form the team if $m_1^* \geq \tilde{m}_1^T$, where $\tilde{m}_1^T$ is implicitly defined by $b(\tilde{m}_1^T) = b^o - b^c$. Together, these observations imply $\tilde{m}_1^* = \max\{\tilde{m}_1^R, \tilde{m}_1^T\}$, and Proposition 5 immediately follows from Lemma 2 and the Implicit Function Theorem.

References

Battaglini, M., Benabou, R., Tirole, J., 2005. Self-control in peer groups. Journal of Economic Theory 123, 105–134.

Benoit, J.-P., Dubra, J., 2004. Why do good cops defend bad cops? International Economic Review 45, 787–809.

Bernheim, B.D., 1994. A theory of conformity. Journal of Political Economy 102, 841–877.

Chevigny, P.B., 1995. Edge of the Knife: Police Violence in the Americas. New Press, New York.

Cho, I.-K., Kreps, D., 1987. Signaling games and stable equilibria. Quarterly Journal of Economics 102, 179–221.

CNN & TIME, 2000. Thin White Line. CNN & TIME, available at http://www.transcripts.cnn.com/transcripts/0009/24/impc.00.html.

Dharmapala, D., Miceli, T.J., 2003. Search, seizure and (false?) arrest: an analysis of Fourth Amendment remedies when police can plant evidence. Working paper, University of Connecticut.

Donohue, J.J., Levitt, S.D., 2001. The impact of race on policing, arrest patterns, and crime. Journal of Law and Economics 44, 367–394.

Ekenvall, B., 2003. Police attitudes towards fellow officers' misconduct: the Swedish case and a comparison with the USA and Croatia. Journal of Scandinavian Studies in Criminology and Crime Prevention 3, 210–232.

Epstein, J., 2002. Breaking the code of silence: bystanders to campus violence and the law of college and university safety. Stetson Law Review 32, 91–124.

Feess, E., Walzl, M., 2004. Self-reporting in optimal law enforcement when there are criminal teams. Economica 71, 333–348.

Finkelhor, D., Ormrod, R.K., 2001. Factors in the underreporting of crimes against juveniles. Child Maltreatment 6, 219–229.

Freeman, R.B., 1999. The economics of crime. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3. North-Holland, Amsterdam, pp. 3529–3571.

Fudenberg, D., Tirole, J., 1991. Game Theory. MIT Press, Cambridge.

Gibson, R., Singh, J., 2003. Wall of Silence: The Untold Stories of the Medical Mistakes that Kill and Injure Millions of Americans. Lifeline Press, Washington.

Gosline, A., 1988. Witnesses in labor arbitration: spotters, informers, and the code of silence. The Arbitration Journal 43, 44–54.

Hagar Report, 2004. Special master's final report re Department of Corrections post powers investigations and employee discipline. United States District Court for the Northern District of California, No. C90-3094-TEH.

Hertel, G., Kerr, N.L., 2001. Priming in-group favoritism: the impact of normative scripts in the minimal group paradigm. Journal of Experimental Social Psychology 37, 316–324.

Hewstone, M., Rubin, M., Willis, H., 2002. Intergroup bias. Annual Review of Psychology 53, 575–604.

Huberts, L.W., Lamboo, T., Punch, M., 2003. Police integrity in the Netherlands and the United States: awareness and alertness. Police Practice and Research 4, 217–232.

Huck, S., Kübler, D., Weibull, J., 2003. Social norms and optimal incentives in firms. Working paper, University College London.

Kaplow, L., Shavell, S., 1994. Optimal law enforcement with self-reporting behavior. Journal of Political Economy 102, 583–606.

Kim, J.-Y., Ryu, K., 2003. Yes men and no men: does defiance signal talent? Journal of Institutional and Theoretical Economics 159, 468–490.

Kohn, L., Corrigan, J., Donaldson, M., 1999. To Err is Human: Building a Safer Health System. National Academy Press, Washington.

Kübler, D., 2001. On the regulation of social norms. Journal of Law, Economics & Organization 17, 449–476.

Levitt, S.D., Snyder, C.M., 1997. Is no news bad news? Information transmission and the role of "early warning" in the principal-agent-model. RAND Journal of Economics 28, 641–661.

Los Angeles Daily News, 2001. See, hear, speak no evil; educators' code of silence keeps failing teachers in the classroom—to the detriment of hundreds or thousands of students. December 2, p. V1.

Mollen Commission, 1994. Commission to Investigate Allegations of Police Corruption and the Anti-corruption Procedures of the Police Department Report. The City of New York, New York.

Motta, M., Polo, M., 2003. Leniency programs and cartel prosecution. International Journal of Industrial Organization 21, 347–379.

Mullen, B., Brown, R., Smith, C., 1992. In-group bias as a function of salience, relevance, and status. European Journal of Social Psychology 22, 103–122.

Sacramento News & Review, 2004. The code of silence. Online edition May 13, available at http://www.newsreview.com/sacramento/Content?Foid=oid%3A29043.

San Francisco Chronicle, 2003. Cracking the code of silence. March 9, p. D1.

Spitalli, S., 2004. Breaking the code of silence: how students can keep schools – and each other – safe. National Association of Pupil Services Administrators Briefs, available at http://www.napsa.com/briefs/briefs.htm.

Süddeutsche Zeitung, 2004. Autobahnraser zu 18 Monaten Haft verurteilt. Online edition February 17, available at http://www.sueddeutsche.de/panorama/artikel/936/26910/.

Tajfel, H., Billig, M.G., Bundy, R.P., Flament, C., 1971. Social categorization and intergroup behaviours. European Journal of Social Psychology 1, 149–178.

Tajfel, H., Turner, J., 1986. The social identity theory of intergroup behavior. In: Worchel, S., Austin, W.G. (Eds.), Psychology of Intergroup Relations. Nelson-Hall, Chicago, pp. 7–24.

Tanner, J., Wortley, S., 2002. The Toronto Youth Crime and Victimization Survey: Overview Report. Centre of Criminology, University of Toronto, Toronto.

Trautman, N., 2000. Police code of silence facts revealed. In: Paper Presented at Annual Conference of the International Association of Chiefs of Police, available at http://www.aele.org/loscode2000.html.

Washington Post, 1999. No telling anymore; at today's schools, fearful kids keep a code of silence. April 27, p. C01.